2 ej

UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

INTRODUCCIÓN A LA TÉCNICA DE SEGMENTACIÓN CHAID

T E S I S

QUE PARA OBTENER EL TÍTULO DE
A C T U A R I A

P R E S E N T A:

GABRIELA MERAZ RIOS

DIRECTOR DE TESIS:
Dr. JOSÉ RODOLFO MENDOZA BLANCO

UNIVERSIDAD NACIONAL
AVENIDA DE
MEXICO

MAT. MARGARITA ELVIRA CHÁVEZ CANO
Jefa de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo de Tesis

INTRODUCCIÓN A LA TÉCNICA DE SEGMENTACIÓN CHAID.

realizado por GABRIELA MERAZ RIOS

con número de cuenta 9036375-9 , pasante de la carrera de ACTUARÍA

Dicho trabajo cuenta con nuestro voto aprobatorio

Atentamente

Director de Tesis
Propietario
DR. JOSÉ RODOLFO MENDOZA BLANCO

Propietario
M. EN A. P. MA. DEL PILAR ALONSO REYES

Propietario
M. EN C. JOSÉ ANTONIO FLORES DÍAZ

Suplente
ACT. YAZMÍN ILIANA BÁRCENAS OROZCO

Suplente
ACT. JAIME VÁZQUEZ ALAMILLA

Consejo Departamental de Matemáticas
M. EN A.P. MA. DEL PILAR ALONSO REYES
FACULTAD DE CIENCIAS

A mis padres

A mis hermanos

A Denis

AGRADECIMIENTOS

No hay palabras para agradecer al Dr. José Rodolfo Mendoza, el apoyo y dedicación que me brindó. Como mi asesor rebasó todas mis expectativas. Fue un honor conocer a una persona que ejerce su labor con respeto, responsabilidad, profesionalismo y amor.

Mi gratitud y reconocimiento a mis sinodales:

Mtra. Pilar Alonso.
Mtro. José Antonio Flores.
Act. Yazmín Bárcenas.
Act. Jaime Vázquez.

Su comprensión, disposición e interés permitieron culminar este trabajo. Sus aportaciones enriquecieron no sólo a la presente tesis, sino también a mi desarrollo profesional.

A la Universidad Nacional Autónoma de México: mi *alma mater*.

Índice

Introducción

La técnica "Detector Automático de Interacción Ji-cuadrada", CHAID por sus siglas en inglés (Chi-squared Automatic Interaction Detector), es una técnica de segmentación que tiene sus orígenes a principios de los años ochenta. A pesar de su reciente aparición, esta técnica goza actualmente de gran popularidad gracias a que se ha revelado como una herramienta sencilla, la cual ofrece resultados sustanciosos y fácilmente interpretables, requiriendo un tiempo corto de proceso. Estas características le han permitido ser utilizada en una amplia y diversa gama de aplicaciones. Sin embargo, a pesar de su creciente popularidad, existen pocos documentos que expongan el algoritmo central de CHAID.

El objetivo de esta tesis consiste en describir, analizar y ejemplificar el algoritmo CHAID, a fin de servir como una fuente explicativa y de consulta, que facilite al lector la comprensión del algoritmo y le permita aplicarlo adecuadamente, visualizando los beneficios que pueda obtener de él.

El primer capítulo presenta conceptos que sirven como antecedentes para comprender el procedimiento que sigue CHAID. En este capítulo se mencionan conceptos básicos de las pruebas estadísticas y la medición de variables; continuando con la teoría de tablas de contingencia y pruebas de significancia χ^2 para independencia e igualdad de distribuciones. Asimismo, se presentan los conceptos básicos referentes a árboles, particularmente de los de clasificación. Por último, este capítulo aborda el tema de segmentación, en especial la de mercados.

El segundo capítulo expone propiamente la teoría de la técnica CHAID. Comienza con una explicación descriptiva de ésta técnica y de su procedimiento, dando paso a sus antecedentes históricos y a sus propósitos iniciales; así como las extensiones que ha experimentado a través de los años y los variados usos que han encontrado cabida en su expansión a otras áreas, apreciándose su creciente potencial.

Una explicación del algoritmo CHAID sirve como preámbulo para su exposición en términos matemáticos, acotando las debidas consideraciones para el algoritmo nominal u

original, y para el correspondiente al de las variables dependientes ordinales. Asimismo, se explica como CHAID calcula la significancia de las predictoras que se utilizan en el análisis. A continuación se exponen las extensiones que se han aportado a la técnica (tablas de ganancias y matriz de clasificación), las cuales mejoran el aprovechamiento y evaluación de los resultados. Finalmente, se destina una sección a los puntajes de categoría, necesarios para el algoritmo ordinal, así como su uso en el algoritmo nominal.

El tercer capítulo presenta dos ejemplos explicativos de los conceptos manejados por CHAID. En el primero se desea determinar el perfil de grupos de alto y bajo riesgo de crédito con relación a ciertas variables predictoras. El grupo de alto riesgo está formado por individuos que solicitaron un crédito sin pagarlo, es decir, se considera un mal crédito. En este ejemplo se sigue paso a paso el algoritmo de CHAID, generando el árbol de segmentación del análisis. Además, se construye la matriz de clasificación para evaluar cada nivel del árbol y se analizan los resultados con las tablas de ganancias para obtener más información de la segmentación final. Particularmente se destaca el uso de las tablas resumen como una herramienta de evaluación del árbol.

El segundo ejemplo se basa en el estudio de la promoción por correo para la suscripción a una revista. En éste, se desea identificar los segmentos de la población que mejor respondan a la estrategia para considerarlos en acciones futuras. Este caso se analiza de tres formas. En la primera, la variable dependiente es dicotómica, es decir, sólo toma en cuenta que las personas respondan o no a la promoción. Se evalúa el árbol con la matriz de clasificación y se describen los cálculos para la construcción de las tablas de ganancias detalladas y de resumen. En la segunda, la variable dependiente es politómica nominal de tres categorías (respondió y pagó, respondió y no pagó, no respondió). A cada una de ellas se les asigna un puntaje para valorar la relación del costo-beneficio asociado. En la tercera forma, la variable dependiente es tratada como ordinal, utilizando los mismos puntajes de categoría para identificar las diferencias de segmentación del algoritmo ordinal con relación al algoritmo nominal.

Capítulo 1

ANTECEDENTES

En este capítulo se presenta información necesaria para aplicar el algoritmo de CHAID adecuadamente y para facilitar su interpretación estadística. En la Sección 1.1, se presenta una explicación de algunas pruebas estadísticas paramétricas y no paramétricas, así como conceptos importantes y útiles en el desarrollo de las pruebas estadísticas. La Sección 1.2 trata el tema de escalas de **medición de variables**, que como se verá en el siguiente capítulo, es de gran importancia para el algoritmo CHAID. Este tema abarca definiciones y ejemplos de escalas de medición. En la Sección 1.3, se muestra un panorama de las **tablas de contingencia**, y de algunas de sus pruebas estadísticas que son útiles para la aplicación y evaluación de la técnica CHAID. A partir de tablas de 2×2 se generaliza a tablas de $r \times c$. La Sección 1.4 aborda el tema de árboles, especialmente los de clasificación. Finalmente, la Sección 1.5 hace referencia a la segmentación, la cual es notable por sus usos en mercadotecnia y muy adecuada para los propósitos de esta tesis.

1.1 PRUEBAS ESTADÍSTICAS

Las pruebas de hipótesis conforman una de las mayores áreas de la inferencia estadística. En esta sección se revisarán los métodos generales para probar hipótesis y su aplicación a problemas comunes.

Primero se deben establecer algunas definiciones:

Definición 1.1 *Una **hipótesis estadística** es una proposición o conjetura acerca de la distribución o características de una o más variables aleatorias. Para denotar una hipótesis estadística se usará la letra H seguida por dos puntos y la proposición que especifica la hipótesis.*

Definición 1.2 *Una **prueba de la hipótesis estadística** H es una regla o procedimiento para decidir rechazar H en favor de una hipótesis alternativa H_1.*

Para probar una teoría a través de las pruebas estadísticas, se establece una hipótesis con respecto a esta teoría y se recaban datos que conduzcan a tomar la decisión de sostener, revisar o desechar la hipótesis como la teoría de la cual se originó.

Pruebas paramétricas y no paramétricas

Las pruebas estadísticas se dividen en dos importantes ramas: las paramétricas y las no paramétricas. En su elección se debe considerar la manera en que se obtuvo la muestra, las hipótesis particulares que se desean probar y el tipo de medición o escala que se empleó en las definiciones operacionales de las variables implicadas. Todas estas cuestiones determinan qué prueba estadística es óptima o más apropiada para analizar un conjunto particular de datos de investigación.

Definición 1.3 *Una **prueba estadística paramétrica** es una afirmación acerca de la distribución de respuestas en la población de la cual se ha obtenido la muestra investigada, la cual se establece en términos de los parámetros de la distribución de los datos. La significancia de los resultados de la prueba paramétrica depende de la validez de estas suposiciones.*

Definición 1.4 *Una **prueba no paramétrica** está basada en un modelo que especifica sólo condiciones generales sobre la distribución de la información y ninguna acerca de los parámetros de la distribución de la cual fue obtenida la muestra.*

1.1.1 Conceptos básicos

Las hipótesis nula y alternativa

- La *hipótesis nula* (H_0) es el supuesto a probar y por lo general se formula con el propósito expreso de controlar la probabilidad de rechazarla cuando es falsa.

- La *hipótesis alternativa* (H_1) es la aseveración o hipótesis que se acepta si se rechaza H_0.

La hipótesis estadística que especifica completamente la distribución, se le conoce como *simple*, si no, es llamada *compuesta*.

Niveles de significancia

Se rechaza H_0 en favor de H_1 cuando una prueba estadística proporciona un valor cuya probabilidad p de una ocurrencia tanto o más extrema que el observado (bajo H_0), es más pequeño que alguna probabilidad α especificada de antemano. A esa probabilidad α se le conoce como *nivel de significancia* y a p como *nivel de significancia descriptivo o valor p*. En otras palabras, si p$\leq \alpha$, se rechaza H_0 en favor de H_1. Los valores comunes de α son 0.05 y 0.01.

El propósito de asignar un nivel de significancia es definir un evento raro de acuerdo con H_0 cuando la hipótesis nula sea verdadera. Así, si H_0 fuera cierta y si el resultado de una prueba estadística en un conjunto de datos observados (tanto o más extremos) tuviera una probabilidad menor o igual a α, se tendría la ocurrencia de un evento raro lo que conduciría, sobre una base probabilista, a rechazar H_0.

Entonces, se puede ver que α proporciona la probabilidad de rechazar equivocada o falsamente H_0. Al error de rechazar H_0 equivocadamente se le conoce como error tipo I. El error tipo II se comete cuando no se rechaza la hipótesis nula, cuando en realidad es falsa.

La región de rechazo

La *región de rechazo* es una región construida con base en la distribución muestral nula. En esta distribución se incluyen todos los valores posibles que un estadístico de prueba puede adoptar. La región de rechazo consiste en un subconjunto de estos valores posibles, y se elige de forma tal que, la probabilidad de ocurrencia de que un estadístico de prueba según H_0, tenga valores en ese subconjunto es de α. Es decir, la región de rechazo consiste en un conjunto de valores posibles que son tan extremos cuando H_0 es verdadera, que la probabilidad es muy pequeña (es decir, igual a α), de manera que la muestra que se observó realmente proporcione un estadístico de prueba que esté entre esos valores. La probabilidad asociada con cualquier valor individual en la región de rechazo es, por supuesto, igual o menor que α.

En ocasiones resulta útil contar con un procedimiento objetivo para rechazar o bien acceptar una hipótesis particular, por lo que se sugiere tomar en cuenta las siguientes observaciones:

- Establecer la hipótesis nula (H_0) y la alternativa (H_1); así como, decidir qué datos se van a recabar y en qué condiciones.
- Seleccionar una prueba estadística para probar H_0. Elegir el modelo de prueba que se aproxime lo más cercanamente posible a las condiciones de la investigación en términos de las suposiciones en las cuales está basada la prueba.
- Especificar un nivel de significancia α y un tamaño de la muestra n.
- Especificar la región de rechazo para la prueba estadística.
- Recabar los datos y calcular el valor de la prueba. Si ese valor está en la región de rechazo, la decisión es rechazar H_0 con el nivel de significancia elegido. Por el contrario, si ese valor cae fuera de esa área, la hipótesis no puede ser rechazada.

1.2 MEDICIÓN DE VARIABLES

La *medición* consiste en establecer reglas para asignar o asociar números o valores a los objetos, con el propósito de representar cantidades o atributos. Los atributos indican que la medida tiene relación con alguna característica particular de los objetos y se describen frecuentemente como variables. Por ejemplo, si se estudia el precio de cierta acción en el mercado, la variable será el precio. Si desea estudiar la altura promedio de una población, la variable será la altura de los individuos de esa población.

Por los valores que pueden tomar las variables se clasifican en dos grupos: categóricas y cuantitativas. La observación de variables lleva a diferentes usos que se hacen de los números o "valores" de las variables, lo que se traduce en diferentes escalas de medición. Las escalas de medición para las variables categóricas son: nominal y ordinal; y para las cuantitativas: de intervalo y de razón o proporción.

1.2.1 Variables categóricas

Las *variables categóricas* difieren de las *cuantitativas* en que ellas no son medidas en forma continua sino que toman valores discretos, los cuales son clasificados en grupos. Se puede convertir variables cuantitativas en categóricas agrupando valores adyacentes. Por ejemplo, si se estudia la variable continua *edad,* se pueden agrupar sus valores en las siguientes categorías: 0 a 14,15 a 20. 21 a 34, 35 a 46, 47 a 58, 59 a 66, y todas las edades más grandes que 66.

Las *variables categóricas* pueden ser nominales u ordinales.

- **Nominales.** La medición en una *escala nominal* consiste simplemente en asignar los valores de la variable a categorías cualitativamente distintas. Las categorías de una variable *nominal* no tienen un orden natural, difieren en tipo más que en grado. Aquí los "valores" u otros símbolos se usan para clasificar un objeto, una persona o una característica en distintos grupos. La operación fundamental es determinar

la pertenencia a una misma categoría o clase; es decir, si se poseen características comunes. Por ejemplo, los estudiantes de una universidad se podrían clasificar por las carreras que están estudiando. Es claro que no existe una relación de orden o jerarquía entre las carreras, por lo que es posible clasificar a cada estudiante de acuerdo a la carrera que estudia y etiquetar las categorías resultantes como 1, 2, 3, etc en cualquier orden. Otro ejemplo se da con los números de teléfono, la asignación de los primeros dígitos que conforman el número telefónico está determinada por la zona del domicilio al que se le va a instalar la línea telefónica. Así, se pueden formar categorías por zona.

En una escala nominal, las operaciones de la escala dividen a una clase dada, en un conjunto de subclases mutuamente excluyentes. La única relación implicada es la equivalencia; esto es, los miembros de cualquier subclase deben ser equivalentes en la propiedad que está siendo analizada.

- **Ordinales.** Si los valores de una variable cumplen los requerimientos de una escala nominal, y existe además, algún tipo de relación entre ellos, se está hablando entonces de una *escala ordinal.* Es decir, las categorías representan series ordenadas de acuerdo a sus relaciones (por ejemplo, más alto, más preferido. más difícil, etc). *Tales relaciones se denotan por medio del símbolo* ">", *el cual en general significa* "mayor que" dependiendo de la naturaleza de la relación que define la escala. En la escala ordinal se tiene clasificación y orden, pero no se tiene ningún conocimiento sobre el tamaño de las unidades de la escala de medición, no hay indicación en sentido absoluto de cuánto es la diferencia entre posiciones sucesivas en la escala. El sistema de grados en el ejército es un ejemplo de escala ordinal: general > teniente > sargento. El nivel socioeconómico de una familia también constituye una escala ordinal, así como las calificaciones del desempeño de los alumnos en una asignatura.

1.2.2 Variables cuantitativas

Estas variables pueden tomar un número finito o infinito de valores y pueden ser clasificadas de acuerdo a escalas de intervalo o de razón.

- **De intervalo.** Cuando una escala tiene todas las características de una medición ordinal y cuando además tienen sentido las distancias o diferencias entre cualesquiera de dos valores de la escala, pero el "cero" se asigna de manera arbitraria, se tiene una *escala de intervalo*. Este tipo de medición se caracteriza por una unidad común y constante de medida, donde una diferencia de cierta magnitud significa lo mismo en todos los puntos de la escala. Por ejemplo, la medición de temperatura conserva un orden: dos escalas de medición para la temperatura, Celcius y Farenheit, observan la misma relación de orden, indicando que a mayor cantidad medida, mayor temperatura. Las escalas, aunque diferentes, guardan la propiedad de que la diferencia de una unidad es igual en toda la escala, por ejemplo, la diferencia entre 35 y 36 grados es igual a la diferencia entre 4 y 5 grados en cualquiera de las dos escalas. Sin embargo, estas escalas no concuerdan en su punto cero. En resumen, al medir la temperatura, la escala y el punto cero son diferentes, pero la información es equivalente Otros ejemplos son la hora del día, el año, etc ..

- **De razón.** Cuando una escala tiene todas las características de la de intervalo y además tiene un punto cero absoluto en su origen (que significa ausencia de la característica), se le llama *de razón*. La razón de cualesquiera dos puntos es independiente de la unidad de medida. Por ejemplo, la masa y el peso son medidos en una escala de razón ya que, si se considera el peso de un objeto en 2 escalas (*libras y gramos*), al reducir o aumentar el peso del objeto, las escalas guardarán la misma proporción, es decir, la razón es la misma para las dos medidas de gramos y en libras.

1.3 TABLAS DE CONTINGENCIA

Definición 1.5 *Una tabla de contingencia, es un arreglo de números naturales en una forma matricial que representan conteos o frecuencias asociados con clases de variables en una o más poblaciones.*

Por ejemplo, un vendedor tiene a la venta un modelo de vestido en tres colores· azul, rojo y verde. Se observa que en un mes vendió 2 vestidos azules, 15 rojos y 7 verdes, por lo que, en total vendió 24 vestidos de este modelo. Estos datos se pueden mostrar usando una tabla de contingencia de 1×3. :

	vestido			
	azul	rojo	verde	total
ventas	2	15	7	24

1.3.1 La tabla de contingencia de 2×2

Una *tabla de contingencia de* 2×2, es un arreglo de las frecuencias de clasificación de una población con dos variables de dos categorías o dos poblaciones en una variable con dos categorías.

Prueba χ^2 para proporciones en tablas de contingencia de 2×2

Se obtienen dos muestras aleatorias independientes de dos poblaciones con n_1 y n_2 observaciones, respectivamente. Cada observación se clasifica en una de dos categorías excluyentes (I o II). Se calculan las frecuencias de cada categoría por población y los datos se acomodan en una tabla de contingencia de 2×2:

Poblaciones	Categorías		Total
	I	II	
Población 1	O_{11}	O_{12}	$n_1 = O_{11} + O_{12}$
Población 2	O_{21}	O_{22}	$n_2 = O_{21} + O_{22}$
Total	$C_1 = O_{11} + O_{21}$	$C_2 = O_{12} + O_{22}$	$n = n_1 + n_2$

donde

O_{ij} = número de observaciones de la población i que se clasificaron en la categoría j.

n_i = número total de observaciones de la muestra de la población i.

C_j = número total de observaciones de las dos muestras que pertenecen a la categoría j.

n = número total de observaciones.

SUPUESTOS

A continuación se resumen los supuestos para esta prueba:

- Cada muestra es considerada aleatoria.
- Las dos muestras son mutuamente independientes.
- Cada observación puede ser clasificada en sólo una clase.

HIPÓTESIS

Considere:

p_i = probabilidad de que un elemento de la población i, escogido aleatoriamente pertenezca a la categoría I.

= proporción de observaciones de la población i que pertenecen a la categoría I.

Para probar si las dos poblaciones son parecidas verificando si hay alguna relación entre sus proporciones, existen tres formas. Si sólo se quiere saber si hay diferencias entre las poblaciones se puede utilizar la prueba A. Si se cree que las poblaciones son distintas y además que una en particular tiene una mayor proporción de ocurrencia en la categoría I, entonces es recomendable emplear las pruebas B y C de una cola.

Prueba A (de dos colas):

$$\begin{aligned} H_0 &: p_1 = p_2 \\ &\quad vs \\ H_1 &: p_1 \neq p_2. \end{aligned}$$

Prueba B (de una cola):

$$\begin{aligned} H_0 &: p_1 \leq p_2 \\ &\quad vs \\ H_1 &: p_1 > p_2. \end{aligned}$$

Prueba C (de una cola):

$$\begin{aligned} H_0 &: p_1 \geq p_2 \\ &\quad vs \\ H_1 &: p_1 < p_2. \end{aligned}$$

ESTADÍSTICO DE PRUEBA

Los estadísticos de prueba T y T_1 para las pruebas de dos y una cola son, respectivamente:

$$T = \frac{n(O_{11}O_{22} - O_{12}O_{21})^2}{n_1 n_2 C_1 C_2}. \tag{1.1}$$

y

$$T_1 = \frac{\sqrt{n}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}}. \tag{1.2}$$

La distribución exacta de T, bajo la hipótesis nula de la prueba A ($H_o : p_1 = p_2 = p$), se calcula de la siguiente manera: la probabilidad de sacar de la población 1, exactamente x_1 elementos que pertenezcan a la categoría I y por consiguiente $n_1 - x_1$ elementos de la categoría II está dada por la probabilidad binomial:

$$P\ (\ x_1 \in \text{categoría 1 y } n_1 - x_1 \in \text{categoría 2}) = \begin{pmatrix} n_1 \\ x_1 \end{pmatrix} p^{x_1}(1-p)^{n_1 - x_1}.$$

Similarmente para la población 2 se tiene:

$$P\ (\ x_2 \in \text{categoría 1 y } n_2 - x_2 \in \text{categoría 2}) = \begin{pmatrix} n_2 \\ x_2 \end{pmatrix} p^{x_2}(1-p)^{n_2 - x_2}$$

Como las *muestras son independientes*, se debe cumplir que para 2 eventos independientes E_1 y E_2, la probabilidad conjunta de los dos eventos es el producto de las probabilidades de ocurrencia de cada uno (Feller,1968). Por tanto,

$$P\begin{pmatrix} x_1 \in \text{categoría 1}, n_1 - x_1 \in \text{categoría 2}, \\ x_2 \in \text{categoría 1}, n_2 - x_2 \in \text{categoría 2} \end{pmatrix} = \begin{pmatrix} n_1 \\ x_1 \end{pmatrix}\begin{pmatrix} n_2 \\ x_2 \end{pmatrix} p^{x_1 + x_2}(1-p)^{n - x_1 - x_2}$$

En el caso simple donde $n_1 = 2$ y $n_2 = 2$, hay nueve diferentes combinaciones correspondientes a las nueve posibles tablas que aparecen en la Tabla 1.1.

Tablas de contingencia	Probabilidades si H_0 es verdadera general	(p=1/2)	(p=1)	T
2 0 2 0	p^4	1/16	1	Indefinido
2 0 1 1	$2p^3(1-p)$	1/8	0	4/3
2 0 0 2	$p^2(1-p)^2$	1/16	0	4
1 1 2 0	$2p^3(1-p)$	1/8	0	4/3
1 1 1 1	$4p^2(1-p)^2$	1/4	0	0
1 1 0 2	$2p(1-p)^3$	1/8	0	4/3
0 2 2 0	$p^2(1-p)^2$	1/16	0	4
0 2 1 1	$2p(1-p)^3$	1/8	0	4/3
0 2 0 2	$(1-p)^4$	1/16	0	Indefinido

Figura 1.1: Posibles tablas de contingencia para $n_1 = 2$ y $n_2 = 2$.

De igual forma, en la Tabla 1.1 se presentan las probabilidades asociadas a cada tabla de contingencia bajo la hipótesis nula. Por ejemplo, para la primera, la probabilidad conjunta bajo H_0 es

$$\text{Probabilidad conjunta} = \binom{2}{2}\binom{2}{2} p^4(1-p)^0 = p^4$$

y el estadístico $T = 0/0$ que es una forma indeterminada. Sin embargo, las dos tablas que resultan en valores indefinidos para T son fuertemente indicativas de que H_0 es verdadera, igual que la quinta tabla, por lo que se define arbitrariamente T como 0 para la primera y última tabla de acuerdo al resultado de la quinta tabla y T tiene la siguiente distribución de probabilidad para $p = 1/2$,

$$\begin{aligned} P(T &= 0) = p^4 + (1-p)^4 + 4p^2(1-p)^2 = 3/8, \\ P(T &= \frac{4}{3}) = 4p^3(1-p) + 4p(1-p)^3 = 1/2, \\ P(T &= 4) = 2p^2(1-p)^2 = 1/8. \end{aligned}$$

Para $p = 1$, la distribución de probabilidad es

$$P(T = 0) = 1.$$

De la misma manera, para cualquier tamaño de muestra de dos poblaciones n_1 y n_2 se pueden encontrar las distribuciones de probabilidad exactas después de definir apropiadamente los valores indefinidos de T. Sin embargo, la función de probabilidad no es única aún cuando H_0 se supone verdadera, como se vio en el ejemplo anterior, sino que depende de p.

El tamaño de la región crítica es máximo cuando $p = 1/2$. Si esta región corresponde al valor más grande de T (i.e. $T = 4$), entonces el nivel descriptivo p = $1/8$. Para valores de n_1 y n_2 pequeños se puede calcular la distribución exacta de T. En el caso de valores grandes de n_1 y n_2 la distribución exacta de T es difícil de tabular debido a todas las diferentes combinaciones de valores posibles de O_{11}, O_{12}, O_{21} y O_{22}. En este caso T, sigue una distribución asintótica de una variable χ^2 con 1 grado de libertad (Cramér, 1946).

Una corrección por continuidad fue introducida por Yates (1934), para compensar parcialmente la inexactitud producida por el uso de una distribución continua (la χ^2) para aproximar la función de distribución discreta de T. La corrección consiste en reducir

el valor absoluto de $O_{11}O_{22} - O_{12}O_{21}$ por la cantidad $n/2$ antes de elevarlo al cuadrado. Yates aconseja usar el siguiente estadístico en lugar de la ecuación 1.1

$$T = \frac{n\left[|O_{11}O_{22} - O_{12}O_{21}| - (n/2)\right]^2}{n_1 n_2 C_1 C_2}. \tag{1.3}$$

Sin embargo, algunos autores como Conover (1974), Pearson (1947) y Grizzle (1967) se oponen a esta corrección ya que, aluden que (1.3) tiende a ser demasiado conservadora, y (1.1) parece estar más de acuerdo a una χ^2 con un grado de libertad. Para el estadístico de una cola, la distribución normal es utilizada, puesto que el estadístico $T_1 = \sqrt{T}$, sigue una distribución normal (Mood et al, 1983).

REGLA DE DECISIÓN

Para un nivel de significancia α, se tiene para cada prueba, las siguientes reglas de decisión:

Prueba A (de dos colas):

$$\text{Utilice el estadístico } T \text{ : Rechace } H_0 \text{ si } T > \chi^{1-\alpha}_{(1)}.$$

Prueba B (de una cola):

$$\text{Utilice el estadístico } T_1 \text{ : Rechace } H_0 \text{ si } T_1 > N^{1-\alpha}_{(0,1)}.$$

Prueba C (de una cola):

$$\text{Utilice el estadístico } T_1 \text{ : Rechace } H_0 \text{ si } T_1 < N^{\alpha}_{(0,1)}.$$

Ejemplo 1.1 *En un experimento se aplicaron dos tratamientos químicos de germinación a dos muestras de 200 semillas cada una, obteniéndose los siguientes resultados:*

Tratamientos	*Germinación*		*Total*
	Si	*No*	
A	190	10	200
B	170	30	200
Total	360	40	400

Tabla 1.2 : Tabla de germinaciones con dos tratamientos.

Se desea saber si la proporción de germinación es igual para los dos tratamientos. Utilice un nivel de significancia $\alpha = .05$.

Se plantea la hipótesis nula :

H_0 : Los dos métodos tienen la misma proporción de germinación.

H_1 : Los métodos difieren en la proporción de germinación.

es decir,

$$H_o : p_1 = p_2$$
$$H_1 : p_1 \neq p_2.$$

Calculando el estadístico T

$$T = \frac{400(190 \times 30 - 10 \times 170)^2}{200 \times 200 \times 360 \times 40}$$
$$= 11.11.$$

Como $T > \chi^{95}_{(1)} = 3.84$, se rechaza H_0 y se puede decir que las proporciones de germinación son diferentes.

Si además, se desea saber si la proporción de germinación en la población A es menor

a la proporción en la población B, la hipótesis a probar sería

$$\begin{aligned} H_o &: p_1 \leq p_2 \\ H_1 &: p_1 > p_2. \end{aligned}$$

El estadístico para la prueba estaría dado por

$$T_1 = \sqrt{T} = \sqrt{11.11} = 3.33.$$

Se compara con $N_{(0,1)}^{975} = 1.96$, y dado que $T_1 > N_{(0,1)}^{975}$ entonces H_o es rechazada y se concluye que la proporción de la población A a la cual se le aplicó el primer tratamiento es más grande que la proporción de la población B con el segundo tratamiento.

Prueba χ^2 para independencia

Otro uso para las tablas de contingencia de 2×2 es para una población clasificada en dos variables, es decir, que cada observación de una muestra de n elementos es clasificada de acuerdo a dos propiedades y cada propiedad puede tomar una de dos formas. Esta prueba es un caso específico de la prueba χ^2 para probar independencia para una tabla de contingencia de $r \times c$, con $r = 2$ y $c = 2$, por esa razón, en esta sección sólo se ofrece un resumen de esta prueba, para mayor referencia ver la Sección 1.3.2.

Se obtiene una muestra aleatoria con n observaciones. Cada observación se clasifica de acuerdo a dos criterios o propiedades. Además, cada una de ellas es clasificada en una de dos categorías con respecto al primer criterio y similarmente con respecto a la segunda propiedad. Una manera conveniente de presentar estos datos es una tabla de contingencia de 2×2.

	Variable 1		
Variable 2	Categoría 1	Categoría 2	Total
Categoría A	O_{11}	O_{12}	n_1
Categoría B	O_{21}	O_{22}	n_2
Total	C_1	C_2	n

donde

O_{ij} = número de observaciones asignadas al renglón i, columna j.

n_i = número total de observaciones de la muestra en el renglón i.

C_j = número total de observaciones en la columna j.

n = número total de observaciones de la muestra.

SUPUESTOS

- La muestra es considerada aleatoria.
- Cada observación puede ser clasificada en sólo una clase de acuerdo al primer criterio y en sólo una clase de acuerdo al segundo criterio.

HIPÓTESIS

La hipótesis que está siendo probada plantea que existe independencia entre las variables de la población. La hipótesis nula se plantea como

H_0 : El que una observación esté en el renglón i es independiente de que la misma observación esté en la columna j, $\forall\ i, j$.

ESTADÍSTICO DE PRUEBA

Para probar esta hipótesis, se utiliza el estadístico T de la sección anterior. El estadístico de prueba T esta dado por :

$$T = \frac{n(O_{11}O_{22} - O_{12}O_{21})^2}{n_1 n_2 C_1 C_2}$$

REGLA DE DECISIÓN

Rechazar H_0 si $T > \chi^{1-\alpha}_{(1)}$. El nivel de significancia es α.

Ejemplo 1.2 *Un investigador realizó una encuesta con un grupo de 450 personas de ambos sexos, para determinar si existe una diferencia entre hombres y mujeres en las características para elegir su coche. Los encuestados respondieron a la pregunta ¿cuál es la característica de su automóvil que más le satisface? Las respuestas se clasificaron en dos categorías de acuerdo a si la respuesta correspondía a apariencia o a rendimiento. Las respuestas obtenidas son:*

	Coche		
Sexo	*Apariencia*	*Rendimiento*	***Total***
Hombre	75	125	200
Mujer	150	100	250
Total	225	225	450

Tabla 1.3: Tabla de características de elección de un coche en hombres y mujeres.

Con un nivel de significancia $\alpha = .05$ se probará la siguiente hipótesis nula.

H_0 : No hay diferencia entre las preferencias de los hombres y mujeres con respecto a las características para la elección de su coche.

H_1 : Sí existen diferencias entre las características deseables para un coche entre hombres y mujeres.

Se calcula el estadístico T

$$\begin{aligned} T &= \frac{450(75 \times 100 - 125 \times 150)^2}{200 \times 250 \times 225 \times 225} \\ &= 22.5. \end{aligned}$$

Como $T > \chi^{95}_{(1)} = 3.84$, se rechaza H_0 y se puede decir que hombres y mujeres muestran una base distinta para la preferencia de los automóviles

1.3.2 Tabla de contingencia $r \times c$

En general, una *tabla de contingencia de* $r \times c$ es un arreglo de frecuencias en r renglones y c columnas. Estas tablas se utilizan para presentar una tabulación de los datos (al menos en una escala nominal) contenidos en varias muestras. Se tienen r poblaciones que son acomodadas en los r renglones de la tabla y c categorías de una variable correspondientes a las c columnas. Otro uso de las tablas de contingencia es con una sola muestra. En este caso, cada elemento puede ser clasificado en uno de los r renglones de acuerdo a un primer criterio y clasificado en cualquiera de las c columnas de acuerdo a una segunda *propiedad*

Prueba χ^2 para diferencias entre poblaciones

Esta prueba puede utilizarse para determinar la igualdad de las distribuciones entre r grupos o poblaciones independientes.

Se tienen r poblaciones de las que se obtiene una muestra aleatoria de cada una. Cada elemento muestral puede ser clasificado en una de c categorías y se asume que las muestras son independientes entre ellas. Los datos se presentan en una tabla de contingencia de $r \times c$, en la cual, los renglones representan poblaciones y cada columna representa una categoría de la variable medida.

Poblaciones	Categorías				Total
	Categoría 1	Categoría 2	...	Categoría c	
Poblacion 1	O_{11}	O_{12}	...	O_{1c}	n_1
Población 2	O_{21}	O_{22}	...	O_{2c}	n_2
...	...	...	...	...	...
Población r	O_{r1}	O_{r2}	...	O_{rc}	n_r
Total	C_1	C_2	...	C_c	n

donde

O_{ij} = número de observaciones de la población i que se clasificaron en la categoría j.

n_i = número total de observaciones de la muestra de la población i.

= $O_{i1} + O_{i2} + \ldots + O_{ic}$.

C_j = número total de observaciones que pertenecen a la categoría j.

= $O_{1j} + O_{2j} + \ldots + O_{rj}$.

n = número total de observaciones de todas las muestras.

= $n_1 + n_2 + \ldots + n_r$.

i = $1, 2, \ldots, r$ y $j = 1, 2, \ldots, c$.

SUPUESTOS

- Independencia entre muestras.
- Independencia entre los elementos dentro de la muestra
- Cada observación es clasificada en sólo una de las c categorías de la variable medida.

HIPÓTESIS

La hipótesis plantea que los grupos difieren respecto a las probabilidades de algunas características y, por tanto, respecto a la frecuencia relativa con que sus elementos caen dentro de algunas categorías. Para probar esta hipótesis, se cuenta el número de casos de cada renglón que caen en las distintas columnas y se comparan las proporciones de casos de un grupo en las distintas variables, con las del otro grupo. Si las proporciones son las mismas, entonces las poblaciones son iguales.

Sea p_{ij} la probabilidad de que un elemento seleccionado aleatoriamente de la población i sea clasificado en la j-ésima categoría, para $i = 1, \ldots, r$ y $j = 1, \ldots, c$. La hipótesis nula se plantea como:

$$H_0 \ : \ p_{1j} = p_{2j} = \ldots = p_{rj}, \ \forall j.$$

$$H_1 \ : \ p_{ij} \neq p_{kj} \text{ para alguna } j, \text{ y para algún par } i \neq k$$

ESTADÍSTICO DE PRUEBA

El estadístico de prueba T está dado por :

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \qquad \text{con} \qquad E_{ij} = \frac{n_i C_j}{n}, \ i = 1, \ldots, r \text{ y } j = 1, \ldots, c.$$

donde

$$E_{ij} = \text{número esperado de observaciones en la celda } (i, j).$$

Este estadístico también se puede escribir en una forma reducida como

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - n.$$

Si la hipótesis nula es verdadera, entonces el número de observaciones en la celda (i, j) debería ser aproximado al tamaño de la i-ésima muestra multiplicado por la proporción

C_j/n de todas las observaciones en la categoría j.

Para dos poblaciones se puede llegar a una forma más simple de este estadístico:

$$T = \frac{1}{\frac{n_1}{n}\left(1 - \frac{n_1}{n}\right)} \sum_{j=1}^{c} \frac{(O_{1j} - C_j \frac{n_1}{n})^2}{C_j}.$$

Si hay dos categorías , i.e., $c = 2$, entonces el estadístico T se convierte en:

$$T = \frac{1}{\frac{C_1}{n}\left(1 - \frac{C_1}{n}\right)} \sum_{i=1}^{r} \frac{(O_{i1} - C_1 \frac{n_i}{n})^2}{n_i}.$$

Si $c = r = 2$, es decir, el caso de una tabla de 2×2, caso visto en la sección anterior, el estadístico se reduce a :

$$T = \frac{n(O_{11}O_{22} - O_{12}O_{21})^2}{n_1 n_2 C_1 C_2}.$$

REGLA DE DECISIÓN

Debido a las dificultades de tabular la distribución exacta de T como en el caso de tablas de 2×2, se utiliza la aproximación asintótica de la distribución de T, ya que, cuando las E_{ij} son grandes. la distribución de T sigue una χ^2 con $(r-1)(c-1)$ grados de libertad, (Cramér, 1946), en consecuencia, la regla de decisión es: rechazar H_0 si T es más grande que el cuantil $(1-\alpha)$ de una variable χ^2 con $(r-1)(c-1)$ grados de libertad.

Como se utiliza la distribución asintótica, el valor aproximado para α, como se calcula aquí, es una buena aproximación del valor real de α si las E_{ij} son suficientemente grandes (Conover, 1980). Sin embargo, si algunas de las E_{ij} son pequeñas, la aproximación puede ser muy pobre. Cochran (1952) sostiene que si cualquier E_{ij} es menor que 1 o si más del 20% de las E_{ij} son menores de 5, la aproximación puede ser bastante débil. Otros autores como Conover (1980) dicen que si r y c no son muy reducidos, las E_{ij} pueden ser

tan pequeñas como 1 sin dañar la validez de la prueba. Si algunas de las E_{ij} son muy chicas, se deben combinar varias categorías para eliminar esas E_{ij}.

Ejemplo 1.3 *Suponga que se desea probar que las mujeres y hombres difieren con respecto a sus cualidades de liderazgo. Se extrajo una muestra aleatoria de 43 mujeres y otra de 52 hombres. Cada individuo se clasificó como "líder", "seguidor" o "sin clasificación". Los resultados fueron:*

Frecuencias (O_{ij})	*Categorías*			
Poblaciones	*Líder*	*Seguidor*	*Sin clasificar*	***Total***
Mujeres	12	22	9	43
Hombres	32	14	6	52
Total	44	36	15	95

Tabla 1.4 : *Tabla de contingencia de líderes y seguidores en hombres y mujeres.*

¿Las mujeres tendrán igual clasificación que los hombres? Utilice un $\alpha = .05$

La hipótesis nula es que, el sexo es independiente de si una persona es líder o seguidor, es decir, la proporción de mujeres que son líderes es igual a la proporción de hombres, similarmente con los seguidores, etc.

Se calculan las frecuencias esperadas, por ejemplo,

$$E_{11} = \frac{n_1 C_1}{n} = \frac{43 \times 44}{95} = 19.9.$$

Haciendo cálculos similares para cada casilla, se obtiene la tabla de valores esperados.

Esperadas (E_{ij})	Categorías			
Poblaciones	Líder	Seguidor	Sin clasificar	**Total**
Mujeres	19.92	16.29	6.79	43
Hombres	24.08	19.71	8.21	52
Total	44	36	15	95

Tabla 1.5: Tabla de frecuencias esperadas para la tabla de seguidores y líderes.

El estadístico es

$$\begin{aligned} T &= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{ij}^2}{E_{ij}} - n \\ &= \frac{12^2}{19.92} + \frac{22^2}{16.29} + \frac{9^2}{6.79} + \frac{32^2}{24.08} + \frac{14^2}{19.71} + \frac{6^2}{8.21} - 95 \\ &= 10.71 \end{aligned}$$

Como $\chi^{95}_{(r-1)(c-1)} = \chi^{95}_{(2)} = 5.99 < T$ se rechaza la hipótesis nula y se puede decir que las poblaciones no son iguales y que sí existen diferencias entre ellas.

La prueba χ^2 para independencia

Para la segunda aplicación de la tabla de $r \times c$ se obtiene una muestra aleatoria con n observaciones. Cada observación se clasifica de acuerdo a dos variables. Hay r categorías (renglones) para la primera variable y c categorías (columnas) para la segunda variable. Con respecto a la primera variable, cada observación es asociada con uno y sólo uno de los r renglones y con respecto a la segunda variable, cada observación es asociada con sólo una de las c columnas. Los datos se presentan en una tabla de contingencia de $r \times c$:

	Variable 1				
Variable 2	Categoría 1	Categoría2	...	Categoría c	**Total**
Categoría 1	O_{11}	O_{12}	...	O_{1c}	R_1
Categoría 2	O_{21}	O_{22}	...	O_{2c}	R_2
...	...	...	...	...	..
Categoría r	O_{r1}	O_{r2}	...	O_{rc}	R_r
Total	C_1	C_2	...	C_c	n

donde

O_{ij} = número de observaciones asignadas al renglón i, columna j,

i = $1, . . , r$ y $j = 1, \ldots, c$.

R_i = número total de observaciones de la muestra en el renglón i

= $O_{i1} + O_{i2} + ... + O_{ic}$, $i = 1, \ldots, r$.

C_j = número total de observaciones en la columna j,

= $O_{1j} + O_{2j} + ... + O_{rj}, j = 1. . . . , c$.

n = número total de observaciones de la muestra.

Cabe notar que en esta prueba los totales de renglón R_i son aleatorios a diferencia de la prueba anterior en la que estaban fijos

SUPUESTOS

- La muestra es aleatoria.
- Cada observación es clasificada en exactamente una sola categoría de las r con respecto a la primera variable y también en una sola en relación con la segunda variable.

HIPÓTESIS

La hipótesis nula se plantea como:

H_0 : El evento "una observación está en el renglón i" es independiente al evento "que la misma observación esté en la columna j", $\forall\, i, j$.

Se sabe que la probabilidad de la intersección de dos eventos independientes es la multiplicación de cada una de las probabilidades de los dos eventos (Feller, 1968), por lo que, la hipótesis nula se puede volver a escribir como

$$H_o \;:\; P(\text{ renglón } i, \text{columna } j) = P(\text{ renglón } i) \times P(\text{ columna } j) \qquad \forall i, j.$$

$$H_1 \;:\; P(\text{ renglón } i, \text{columna } j) \neq P(\text{ renglón } i) \times P(\text{ columna } j) \text{ para alguna } i, j.$$

ESTADÍSTICO DE PRUEBA

El estadístico de prueba T está dado por :

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \qquad \text{con} \qquad E_{ij} = \frac{R_i C_j}{n},\ i = 1, \ldots, r \text{ y } j = 1, \ldots, c \qquad (1.4)$$

o equivalentemente una expresión reducida

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - n$$

donde

$$E_{ij} = \text{número esperado de observaciones en la celda } (i, j).$$

Nótese que, si las frecuencias observadas son muy similares a las frecuencias esperadas, las diferencias $(O_{ij} - E_{ij})$ serán pequeñas y, por lo tanto, el valor de T será pequeño. Con un valor pequeño de T no es posible rechazar la hipótesis nula de que las dos variables son independientes.

REGLA DE DECISIÓN

Rechazar H_0 si T sobrepasa el cuantil $(1-\alpha)$ de una variable aleatoria χ^2 con $(r-1)(c-1)$ grados de libertad. El nivel de significancia aproximado es α. Las mismas consideraciones acerca de la regla de decisión de la prueba anterior se aplican para ésta.

Ejemplo 1.4 *Una muestra aleatoria de estudiantes en una cierta universidad fueron clasificados de acuerdo a la carrera que estaban estudiando y a si provenían de una preparatoria del gobierno o privada. Los resultados se presentan en la siguiente tabla de contingencia de* 2×4.

	Carreras				
Preparatoria	*Ingeniería*	*Artes y Ciencias*	*Económicas*	*Otras*	***Total***
Pública	16	14	13	13	56
Privada	14	6	10	8	38
Total	30	20	23	21	94

Tabla 1.6: Tabla de carreras elegidas por estudiantes de preparatorias públicas y privadas.

Se debe probar la hipótesis de que la carrera que escogió cada estudiante es independiente de si proviene de una escuela pública o privada. Usando (1.4) se obtiene el valor T

$$T = 1.524$$

El cuantil de .95 para una χ^2 con 3 grados de libertad es 7.8147 por lo que se acepta la hipótesis nula Concluyéndose que la elección de carrera es independiente del tipo de preparatoria.

Prueba de cociente de verosimilitud y modelos loglineales

Los métodos comentados en esta sección se resumen con el uso de la estadística

$$T = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

donde

$$O_{ij} = \text{número registrado de observaciones en la celda } (i,j).$$
$$E_{ij} = \text{número esperado de observaciones en la celda } (i,j).$$

Sin embargo, no son los únicos para analizar tablas de contingencia. Existe un método de análisis diferente llamado **prueba de cociente de verosimilitud logarítmica (log likelihood ratio test)**, que emplea el siguiente estadístico:

$$T_2 = 2\sum_i \sum_j O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right),$$

en lugar de T, donde ln se refiere al logaritmo natural. La estadística T_2 es comparada con la distribución χ^2, igual que para T, con el mismo número de grados de libertad que los usados para T. Aunque los dos estadísticos T y T_2 tienen la misma distribución asintótica, sus valores difieren para una tabla de contingencia particular y esa diferencia puede ser grande.

Otro método popular de análisis son los **modelos loglineales**. Estos modelos se utilizan generalmente para analizar tablas de contingencia de 3 o más dimensiones y son muy populares para diseño de experimentos. Las mismas estadísticas T y T_2 son ocupadas en los modelos loglineales; la diferencia está en el método para encontrar las E_{ij}. Usualmente ocupan métodos iterativos que requieren de una computadora.

El nombre de modelos loglineales proviene de una razón muy simple, ya que, en las tablas de contingencia de 2×2, la hipótesis nula de independencia se puede expresar como:

$$H_0 : p_{ij} = p_{i+} \times p_{+j} \quad \forall i, j.$$

donde

p_{ij} = probabilidad de que una observación sea clasificada en la celda (i,j).

p_{i+} = probabilidad marginal de renglones.

p_{+j} = probabilidad marginal de columnas.

Tomando el logaritmo de p_{ij} da como resultado

$$H_0 : \log p_{ij} = \log p_{i+} + \log p_{+j}$$

que es una ecuación lineal. Entonces se prueban estas cantidades para ver si el modelo para los logaritmos de las probabilidades de las celdas es una función lineal de los logaritmos de las probabilidades marginales. Para mayor información ver Bishop, Fienberg y Holland (1975).

1.4 ÁRBOLES

Para comenzar esta sección, se darán algunas definiciones de teoría de gráficas a fin de entender el concepto de árbol. En particular esta sección está dedicada a la notación y metodología de los árboles de clasificación puesto que son los que utiliza el algoritmo CHAID.

Un *árbol* es un caso especial de lo que, en matemática combinatoria se llama una gráfica.

Definición 1.6 *Una gráfica G es una pareja ordenada (X, A) donde X es un conjunto finito de elementos llamados vértices o nodos y $A \in X \times X$ es un conjunto cuyos elementos (x, y) son llamados arcos o aristas.*

Definición 1.7 *Un camino es una sucesión alternada de vértices y arcos, tales*

que, para cada arco de la sucesión, el vértice que precede es su vértice inicial y el que lo sucede es su vértice final.

Definición 1.8 *Una cadena es una sucesión alternada de vértices y arcos, tal que, para cada arco de la sucesión los vértices que lo encuadran son sus vértices terminales (sin importar la dirección).*

Definición 1.9 *Un ciclo es una cadena cerrada.*

Definición 1.10 *Un circuito es un camino cerrado.*

Definición 1.11 *Una gráfica es conexa si para todo par de vértices, $x, y \in X$, existe una cadena que los une.*

Dadas las definiciones anteriores, ahora se puede definir un *árbol* como una gráfica $T = [\mathbf{X}, \mathbf{A}]$ conexa y sin ciclos, (para mayor referencia ver Serre, 1980). Ejemplos de árboles se presentan en la Figura 1.1.

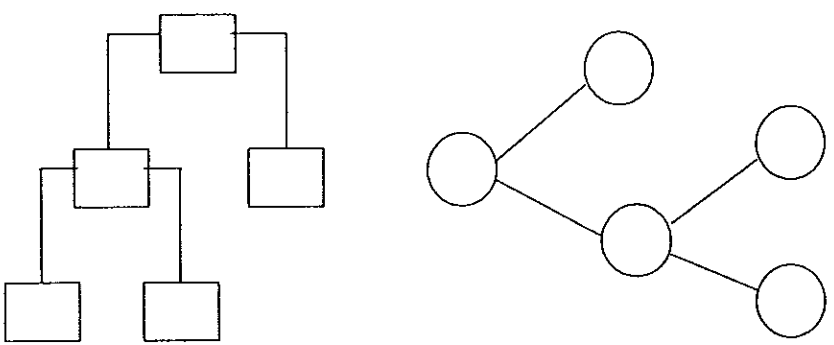

Figura 1.1: Ejemplos de árboles

1.4.1 Árboles de clasificación

Los árboles son muy recurridos en la computación (estructuras de datos), biología (clasificación), psicología (teoría de las decisiones) y muchos otros campos. En estadística, los árboles de clasificación y de regresión se utilizan para la predicción. En las dos últimas décadas, se han vuelto populares como alternativas a la regresión, al análisis discriminante y a otros procesos de clasificación basados en modelos algebraicos.

Un *árbol de clasificación* es una regla empírica para predecir la categoría a la que pertenece un objeto a partir de los valores de las variables predictoras o independientes. Se construyen dividiendo repetidamente un conjunto de n observaciones en subconjuntos ajenos y exhaustivos, que son representados como nodos descendientes a partir del nodo X_1 al que se le llama **nodo raíz (root node)**, ya que a partir de éste se ramifican los demás vértices (Figura 1.2).

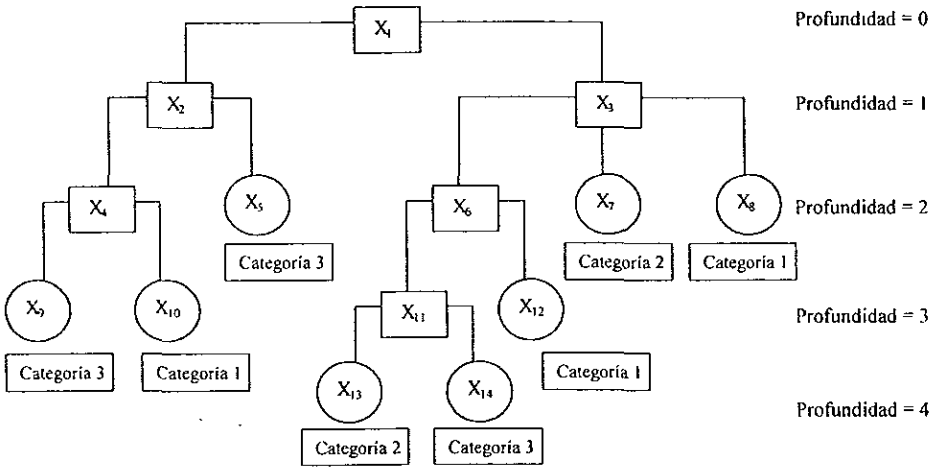

Figura 1.2: Ejemplo de árbol de clasificación.

Se llama **nodo paterno (parental node)**, al vértice al cual un nodo está relacionado *de manera ascendente* en el árbol. Los nodos que aparecen de manera descendente a partir de un nodo paterno, se les conoce como **nodos hijos (child nodes)**. Al número

de niveles entre el nodo raíz y los nodos terminales se le conocerá como **profundidad del árbol (tree depth).**

En la Figura 1.2, se observa que los nodos X_2 y X_3 son disjuntos o mutuamente exclusivos y $X_1 = X_2 \cup X_3$, es decir, contienen toda la población. Los nodos que no tienen ninguna división posterior son llamados **nodos terminales (terminal nodes)** y se indican con un círculo en la misma figura.

Los nodos terminales representan los subconjuntos finales en los que se dividió el conjunto original de datos, es decir, forman una partición de las n observaciones. Cada subconjunto o nodo terminal es designado por una **etiqueta de clase o categoría** (e.g. en la Figura 1.2, el nodo X_9 tiene asignada la categoría 3). Puede haber dos o más nodos terminales con la misma etiqueta de clase (e.g. el nodo X_{10} y el nodo X_{12} comparten la misma categoría). La partición correspondiente al clasificador se obtiene juntando todos los subconjuntos que tengan la misma etiqueta. En este ejemplo, las observaciones de los nodos X_8, X_{10} y X_{12}, pertenecen a la categoría 1, los nodos X_7 y X_{13} pertenecen a la categoría 2 y los nodos terminales X_5, X_9 y X_{14} tienen todas las observaciones pertenecientes a la categoría 3.

Las divisiones se forman por condiciones en los datos. El árbol clasificador predice una categoría para el vector de medidas $\mathbf{x}$ de esta manera: de la condición para la primera división se determina si $\mathbf{x}$ se va al nodo X_2 o X_3 . Si, por ejemplo, $\mathbf{x}$ se va al nodo X_3, entonces de la condición de división del nodo X_3 se determina si $\mathbf{x}$ cae en X_6, X_7 o X_8.

Cuando finalmente $\mathbf{x}$ se mueve a un nodo terminal, su categoría predicha está dada por la etiqueta de la clase correspondiente a ese subconjunto terminal.

En resumen, la construcción de un árbol involucra tres elementos principales:

1. La selección de las divisiones.

2. La decisión de cuándo declarar un nodo terminal o continuar dividiéndolo.

3. La asignación de un nodo terminal a una categoría o clase.

1.5 SEGMENTACIÓN

La *segmentación* es una herramienta relativamente reciente que persigue el objetivo de dividir una población en grupos homogéneos y excluyentes con respecto un criterio específico. El comercio fue el terreno en el que cobró especial importancia, pero no fue sino hasta los años cincuenta que los académicos la formalizaron (Smith, 1956). La segmentación es invaluable en una amplia gama de aplicaciones, sobre todo en campo de la mercadotecnia; por ejemplo, una compañía puede definir un perfil de clientes potenciales para comprar un producto y asignar recursos de promoción y de ventas para atacar a sus mejores prospectos. De esta manera, puede especializar sus ofertas de productos y estrategias comerciales con base en las características de cada segmento.

Hoy en día, una gran variedad de tecnologías estadísticas se ocupan para identificar segmentos que ofrezcan mayor oportunidad para un producto o servicio de una compañía, para identificar patrones de conducta, necesidades, actitudes, etc. Entre estas técnicas están los cúmulos de partición (partition clustering), cúmulos jerárquicos (hierarchical clustering) y análisis de factor de tipo-Q (Q-type factor analysis), el detector de interacción automática (Automatic Interaction Detector, AID), el detector de interacción automática χ^2 (Chi-squared Automatic Interaction Detector, CHAID), la técnica de árboles de clasificación y regresión (Classification and Regression Trees, CART) y el análisis de discriminante.

1.5.1 Segmentación de mercado

La proposición básica de la segmentación de mercado plantea que, la mayoría o probablemente todos los mercados, están conformados de submercados que son relativamente homogéneos en términos de lo que ellos necesitan o quieren de las compañías que ofrecen productos o servicios similares.

Los *segmentos de mercado* consisten en grupos de personas u organizaciones que responderán de manera similar a una mezcla particular de mercadotecnia (producto, precio,

promoción y distribución), o son similares en otras características que son significativas para propósitos de planeación mercadológica.

Existen dos grandes tipos de esquemas de clasificación que son los más usados para realizar la segmentación.

- Cliente contra producto/servicio.
- "A priori" contra "a posteriori".

Para el enfoque basado en el cliente, es importante observar las características específicas de los prospectos que los diferencian y que son significativas para los propósitos de *planeación mercadológica*; por ejemplo, características demográficas, valores, necesidades. En contraste, el enfoque basado en el producto/servicio trabaja con las características específicas de los productos o servicios, los tipos de beneficios que los clientes quieren, las tasas o patrones de uso, etc.

En el otro esquema de clasificación, una segmentación "a priori" designa grupos de consumidores que son similares en términos de algún factor o factores que son o se creen conocidos de antemano por estar relacionados con el consumo del servicio o producto de una compañía, por ejemplo, la demográfica, el volumen de compra o el área geográfica. Las variables de una segmentación "a priori" son seleccionadas antes de que el análisis comience. Los segmentos "a posteriori" son establecidos con base en respuestas que están disponibles sólo después de que ha sido llevada a cabo una recabación de datos, tales como, valores, necesidades, tasas o patrones de uso que pueden señalar segmentos de mercado útiles. Los dos esquemas no son excluyentes, por el contrario, en la mayoría de los estudios actuales se combinan y proporcionan flexibilidad en los análisis.

La mayoría de los trabajos de segmentación se caracterizan por ciertos pasos generales que a continuación se mencionan a fin de facilitar su aplicación:

- Decidir las variables de segmentación o también llamadas variables base (basis variables).

- Decidir la metodología de análisis para los datos.
- Aplicar la metodología para identificar varios segmentos.
- Seleccionar los segmentos objetivo.
- Describir el perfil de todos los segmentos usando las variables base y otras.
- Desarrollar una mezcla mercadológica para cada segmento.

Requerimientos para una segmentación efectiva

Sin fijarse en los objetivos o las tecnologías empleadas, cualquier segmento objetivo seleccionado puede poseer varios criterios útiles para una compañía. Debe ser de tamaño *suficiente* para justificar el gasto de los esfuerzos de mercadotecnia. Debe ser también *distinguible* de otros segmentos y de la población en total. Debe ser *accesible* para la promoción normal de la compañía y a los métodos de distribución. Y por último, normalmente debe ser *compatible* con los recursos y experiencia de la compañía. Frecuentemente, segmentos prometedores se deben desechar ya que no cumplen con uno o más de estos criterios.

Otra disyuntiva se presenta al elegir del número y tamaño de los segmentos para operar con ellos. Generalmente las grandes compañías operan con porciones grandes y viceversa con las pequeñas. Sin embargo, esta decisión depende de diversos factores, entre los que se encuentran: los objetivos y recursos de las organizaciones, la competencia, los requerimientos del mercado, etc. Incluso la decisión de enfocar esfuerzos a dos o más segmentos con diferentes características tampoco resulta evidente, ya que, puede generar conflicto que se trate de abarcar varios segmentos con el mismo producto, debido a la inconsistencia del mensaje de consumo y a la débil posición en las mentes de los consumidores.

Finalmente, también es difícil elegir un modelo de segmentación que sea eficiente (utilizar los menos segmentos) y efectivo (cualquiera y todas las diferencias significati-

vas, tanto estadística como comercialmente, son tomadas en cuenta). La significancia estadística se refiere a la diferencia que existe entre segmentos y la comercial alude a la capacidad de emplear estas diferencias entre segmentos para incrementar las ganancias o bajar el costo.

Beneficios de la segmentación

La habilidad de asignar recursos más eficientemente es uno de los beneficios de la segmentación. Usualmente esto significa reasignar gastos de ventas y de mercadotecnia, tales como llamadas de ventas por teléfono, disminuyéndolas en los segmentos y clientes de bajo potencial e incrementándolas en los de mayor probabilidad de éxito de venta.

El entendimiento y pronósticos de los segmentos puede beneficiar a otras áreas, tales como, producción Por ejemplo, si hay dos diferentes modelos de un producto, la gerencia de producción podrá saber qué cantidad de cada producto se debe manufacturar. Típicamente, el modelo preferido del cliente variará dependiendo del segmento de mercado. Con este pronóstico, la gerencia podrá proyectar mejor las combinaciones de productos y compaginar las decisiones de producción con los segmentos específicos que den las más altas ventas.

Otro beneficio se obtiene al desarrollar mejores pronósticos de demanda global. El pronóstico basado en suposiciones promedio, con incertidumbre promedio, acerca del mercado promedio global no es tan confiable como hacer suposiciones más específicas, acerca de los más importantes segmentos del mercado que han sido estudiados con mayor detalle.

Los pronósticos sin segmentación pueden caer en error de la idea de que inclusive si un nuevo producto sólo captura una pequeña parte de un gran mercado, el rendimiento o ganancia es enorme. Después de una cuidadosa segmentación, hay inclusive más expectativas de compartir penetración en los segmentos más relevantes del mercado. Pero al mismo tiempo se identifican segmentos grandes del mercado que pueden ofrecer muy bajo

potencial. El pronóstico basado en suposiciones específicas es más realista, y usualmente muy diferente a un pronóstico que no diferencia entre segmentos y usa suposiciones promedio.

En resumen, hay tres grandes beneficios de usar un buen modelo de segmentación:

- Enfocarse o atacar *más efectivamente* a aquellos que tienen más probabilidad de comprar el producto;
- Asignar más eficientemente recursos para satisfacer la demanda de cada segmento y
- Obtener un mejor pronóstico de demanda global.

Capítulo 2

CHAID

CHAID es una técnica que originalmente se desarrolló para realizar análisis de exploración para descubrir estructuras o interacciones en los datos, es decir, para investigar las relaciones potenciales en un conjunto de observaciones. Actualmente sirve primordialmente como una técnica de segmentación de poblaciones, particularmente utilizada en investigación de mercados, obteniendo perfiles de consumidores o usuarios.

CHAID considera una variable dependiente y al menos una variable independiente. Básicamente trata de predecir la primera a partir de las independientes, semejando una regresión múltiple. Sin embargo, la regresión no está diseñada para identificar grupos de casos que son similares en términos de factores que afectan la variable dependiente. Un análisis de regresión múltiple tradicional, a diferencia de CHAID, trabaja generalmente con variables cuantitativas (de intervalo o de razón). La técnica de CHAID fue diseñada para usarse con variables categóricas (e.g. sexo, nivel socioeconómico, religión, ocupación, raza, ciudad, etc...) y fue extendida a variables cuantitativas, las cuales son divididas en categorías (e.g. salario, educación, edad, etc...).

CHAID, cuyas siglas en inglés significan "Chi-Squared Automatic Interaction Detector", utiliza pruebas de significancia estadística χ^2 para tablas de contingencia para probar relación entre variables con el fin de particionar en forma repetitiva o automática una población en dos o más segmentos mutuamente excluyentes (cada caso pertenece a

sólo un segmento), y exhaustivos (cada caso pertenece al menos a un segmento), que mejor describan a la variable dependiente y difieran significativamente en términos de la variable dependiente u objetivo.

Aunque no es un procedimiento sofisticado, CHAID es una herramienta extremadamente versátil Puede ahorrar mucho tiempo al investigador, evitando que éste analice cientos de tablas de contingencia, sin encontrar alguna relación importante entre las variables. CHAID identifica rápida y fácilmente las relaciones significativas entre las predictoras o variables independientes gracias a la automatización del proceso de búsqueda de predictoras y de exploración de relaciones entre las variables; genera un resumen de los resultados de la segmentación en forma de diagramas de árbol de fácil comprensión, dividiendo el total de datos en ramas progresivamente más pequeñas, donde los últimos nodos son los segmentos finales; asimismo, genera "tablas de ganancias" y "matrices de clasificación" que contienen información relevante para calificar y validar la calidad de los resultados finales.

2.1 ANTECEDENTES HISTÓRICOS

Históricamente, CHAID tiene sus fundamentos en la técnica de análisis de datos cuantitativos estructurados en árbol "Automatic Interaction Detection" (AID) (Morgan y Sonquist, 1963). Ambas técnicas pertenecen a las *técnicas de detección de interacción.*

Las técnicas de detección de interacción separan el total de la muestra en segmentos que difieren en términos del factor o combinación de factores que guardan la relación más grande con la variable dependiente. Hacen esto dividiendo por separado los elementos de la muestra en subgrupos de acuerdo a cada variable independiente o predictora y entonces prueban la significancia estadística de las diferencias entre subgrupos, en términos de los valores de la variable dependiente.

El algoritmo realiza una división de la muestra a lo largo de varios pasos. En general, el procedimiento es el siguiente: primero se selecciona la variable que produce la diferencia

más grande entre grupos en la variable dependiente. A continuación, cada subgrupo obtenido se divide en subgrupos de acuerdo con las variables restantes y nuevamente se someten a prueba para encontrar diferencias en términos de la variable dependiente. Se elige aquella que muestra la mayor diferencia. Cada subgrupo se divide y se continúa hasta que sean tan pequeños que ya no puedan ser subdivididos o hasta que ya no haya una variable independiente que aumente significativamente la separación de grupos. El proceso entero es mostrado gráficamente, en algunos de estos métodos, en forma de árbol.

El término *interacción* proviene del hecho de que los subgrupos para cada factor *importante* son divididos independientemente en subgrupos del factor que muestra las diferencias más grandes en la variable dependiente. En el diagrama de árbol, la interacción se representa por ramificaciones del mismo nodo (o rama), mismas que a su vez cuentan con diferentes variables independientes más abajo.

AID es el más antiguo de los métodos de detección de interacción. Fue diseñado para ajustar árboles de manera que predijeran una variable cuantitativa. Esta técnica trabaja con una sola variable dependiente medida en una escala de intervalo o de razón y con un número razonable de variables independientes medidas en cualquier tipo de escala; sin embargo, mientras más grande sea el número de estas variables, más grande debe ser la muestra a fin de proporcionar resultados estables.

La técnica AID busca la mejor división binaria para cada variable. Esto requiere encontrar todas las posibles divisiones en dos grupos, y entonces, para cada parte, probar la significancia de las diferencias entre grupos en términos de sus valores promedio en la variable dependiente, que como está en escala métrica, se utiliza una prueba F. La variable con el valor más grande F identifica los dos primeros segmentos.

Dado que, en algunos estudios, la variable dependiente está en forma de categorías; por ejemplo, el tipo de automóvil preferido, actividades recreativas predilectas, respuestas a un programa de correo directo, etc, surgieron otras técnicas similares en perspectiva y metodología general que atacan este problema, tales como THAID, CHAID y CART.

THAID se dio a conocer en los años setentas en la Universidad de Michigan (ver Messenger y Mandel, 1972 y Morgan y Messenger, 1973). Su propósito era extender AID permitiendo que la variable dependiente fuera nominal. Similar a AID, sólo están permitidas divisiones binarias en las variables independientes. THAID utiliza dos pruebas estadísticas (llamadas *Theta* y *Delta*), para decidir si un nodo paterno debe ser dividido El estadístico Theta está definido como la proporción de la muestra clasificada correctamente cuando se utiliza una estrategia de "predicción óptima al modo". El estadístico Delta está basado en el objetivo de encontrar grupos divididos cuya probabilidad difiera máximamente del grupo original y, por lo tanto, de los demás. El analista puede seleccionar uno u otro para cada análisis

El profesor Gordon Kass propuso en 1980, una modificación a AID llamada CHAID para las variables dependientes e independientes categorizadas. Su algoritmo incorpora una fusión secuencial de categorías homogéneas, un procedimiento de separación basado en una estadística de χ^2, con divisiones múltiples en lugar de binarias y con un nuevo tipo de variable predictora o independiente para manejar valores faltantes. Kass estaba preocupado por el tiempo de cálculo, por lo que, decidió quedarse con una división cuasi-óptima en cada predictora, en vez de buscar todas las combinaciones posibles de las categorías.

Magidson en 1992, efectuó una modificación al algoritmo original de Kass para variables dependientes ordinales. En el algoritmo ordinal se utilizan puntajes de categoría en el cálculo de los valores p para probar más fuertemente la significancia de las predictoras.

El algoritmo CHAID ahorra tiempo en la computadora produciendo resultados razonables. sin embargo, no se debe olvidar que la regresión por pasos. no garantiza que se encuentren las mejores divisiones predictoras. Esto sólo se puede lograr con la regresión de todos los subconjuntos posibles o con una búsqueda exhaustiva de los subconjuntos de categoría. Aunque al principio se enfocó a predictoras categóricas. también se puede utilizar para modelos mixtos de manera indirecta. Particionando el rango de posibles valores de las variables cuantitativas en categorías, es posible aplicar el algoritmo para

este tipo de variables. El algoritmo CHAID es, a pesar de estas consideraciones, una forma efectiva y rápida de buscar heurísticamente a través de tablas amplias.

Para finalizar con esta recapitulación de técnicas, se debe hacer mención al programa "Classification and Regression Trees" (CART) de Breiman y colegas (1984). Esta técnica sigue el procedimiento general de división de grupos en términos de una sola variable dependiente nominal, similar al de CHAID y THAID, sin embargo, tiene algunas características que lo diferencian de estos métodos. La principal diferencia que los seguidores de CART esgrimen en su favor, es que éste sigue una secuencia que permite podaciones del árbol para construirlo del tamaño correcto, en contraposición a CHAID, el cual sigue una construcción descendente del árbol y no poda ramas. Por el contrario, entre sus desventajas, CART sólo permite divisiones binarias y no utiliza un valor ajustado de probabilidad por lo que está sujeto a sobreestimación.

2.2 USOS DE CHAID

CHAID se ocupa como una herramienta importante en múltiples campos y actualmente ha encontrado una amplia gama de utilidades en la mercadotecnia; tales como, reconocer segmentos de mercado, explicar las diferencias en estudios de satisfacción del cliente o conocer el perfil de aquellos a quienes les gustó un nuevo concepto, producto, empaque o publicidad. CHAID también se utiliza para descubrir los efectos de interacción entre las predictoras, por ejemplo, la edad puede tener un efecto diferente en la respuesta a una promoción en dos grupos con salario distinto.

En particular, sus usos en mercadotecnia son variados: estudios de mercado en el que se realiza un análisis de compradores y no compradores, delimitando el perfil del mercado objetivo más beneficioso así como las características de los mejores clientes potenciales; pronósticos de demanda para incrementar el ingreso gracias a que se concentran los esfuerzos de mercadotecnia y ventas en clientes de gran demanda; asignación de recursos más eficientemente para satisfacer la demanda de cada segmento; obtención de un mejor

pronóstico removiendo sesgos en estudios o muestras utilizadas para hacer pronósticos. CHAID también facilita la realización de sondeos con respuestas, lo que se conoce como correo directo (direct mailing), al determinar los perfiles de los clientes e identificar los individuos potenciales para responder.

Aunque el objetivo primario de CHAID es actualmente el análisis de segmentación de mercado, también tiene otros usos:

- *Análisis de créditos*: comparación de préstamos de alto riesgo frente a los de bajo riesgo, identificando clientes con un perfil de deudor.
- *Investigación biomédica*: evaluación de ensayos clínicos e interacción de los factores de riesgo, así como la descripción de los perfiles de pacientes.
- *Estudios de sondeos de opinión*: análisis de datos para determinar los perfiles de votación de una población.
- *Regresión logística*: identificación de los efectos de interacción para incluir en análisis de regresión logística.
- *Estadística no paramétrica*: simplificación de tablas de contingencia combinando las categorías que menos difieran significativamente.
- *Control de calidad*: analizar datos de producción para identificar los factores principales que repercuten en los defectos de producción.
- *Estudios políticos*: analizar datos de encuestas para tener una idea sobre las variables que más afectan a un gobierno o una elección.
- *Investigación científica*: analizar resultados de experimentos para determinar las variables que más afectan un fenómeno.
- *Tipologías*: mostrar los resultados de un estudio en términos de los grupos demográficos más relevantes.

- *Reducción de variables:* identificar información inútil de una base de datos a fin de poder eliminarla con seguridad.
- Complemento al *análisis de regresión,* ya que, permite identificar las variables e interacciones importantes en el análisis.

2.3 COMPONENTES DEL ANÁLISIS CHAID

Un análisis de CHAID tiene los siguientes componentes básicos:

1. El método de análisis (nominal u ordinal) para construir el modelo de segmentación. Este criterio depende de la variable dependiente.
2. Una o más variables predictoras cuyos valores son usados para definir los segmentos.
3. Valores para los diversos parámetros de CHAID.

2.3.1 Variables

Variable dependiente u objetivo

La *variable dependiente u objetivo* es aquella que se va a explicar o describir. En CHAID puede ser dicotómica (con dos categorías) o politómica (con más de dos categorías). En ambos casos, puede ser tratada como nominal u ordinal. Si la variable dependiente es nominal, el criterio de segmentación se basa en la distribución de probabilidad de ésta Si la variable es politómica ordinal, se pueden utilizar *puntajes de categoría* (Sección 2.8) para ordenarlas y proveer una medida de distancia relativa entre ellas y el criterio de segmentación es la media o el valor esperado, de los puntajes de categoría especificados. En el análisis de variables dependientes dicotómicas, tanto con el método ordinal como nominal, resultan en la misma segmentación.

Variable independiente o predictora

Es la variable cuyos valores servirán para predecir o explicar a la variable dependiente, puede ser dicotómica o politómica. CHAID permite clasificarlas en tres tipos: *monotónica, libre* o *flotante.* Esta elección afecta el algoritmo para unir categorías y al cálculo de niveles de significancia.

- **Predictora monotónica (monotonic predictor).** Una variable predictora monotónica es aquella cuyas categorías caen en una escala *ordinal* (predictoras que tengan un orden natural) ver Sección 1.2. Las categorías de una variable monotónica sólo se pueden combinar si son *adyacentes* .

- **Predictora libre (free predictor).** Una variable predictora libre es aquella cuyas categorías son puramente nominales. CHAID permite cualquier agrupamiento de categorías de este tipo de variables, sean contiguas o no. Las variables *nominales* deberán ser tratadas como libres.

- **Predictora flotante (floating predictor).** En muchos casos prácticos, las categorías de una predictora se encuentran a nivel ordinal, excepto la de una sola categoría que puede no pertenecer al resto o cuya posición en una escala ordinal es desconocida. Esta situación se presenta frecuentemente cuando una investigación permite una categoría para valores desconocidos o faltantes. Dado que estos datos no afectan el análisis CHAID, es usual asignarles una categoría especial denominada flotante. Excepto esta última, sólo está permitido agrupar categorías contiguas como las predictoras monotónicas. La flotante, sin embargo, puede estar sola o en combinación con cualquier otra categoría o grupo de categorías. Cuando está combinada, toma las características de aquélla con la cual comparte la distribución de la variable dependiente. Para facilitar la comprensión de este tipo de predictora supóngase que la edad de los entrevistados de una muestra está representada por la variable *edad.* Si no se tuviera la edad de algunos sujetos de la muestra, entonces se tendrían valores desconocidos y la predictora *edad* sería normalmente tratada

como monotónica excepto que la categoría final representa "desconocido". Por lo tanto, *edad* es definida como una variable flotante.

No es conveniente definir todas las variables como libres, ya que, se subestimará la significancia de las predictoras. Asimismo, cabe notar que, las predictoras dicotómicas serán tratadas como monotónicas aún cuando sean clasificadas como libres o flotantes. Por último, se debe tomar en consideración, que el tiempo de cálculo se incrementa exponencialmente para las predictoras libres, en la medida en que incrementan las categorías.

2.3.2 Parámetros de paro

Estos parámetros determinarán el tiempo de paro del algoritmo de CHAID, dado que éste continúa dividiendo cada subgrupo hasta que una de las siguientes reglas sea alcanzada:

- No hay más divisiones estadísticamente significativas.
- El tamaño de la muestra para el subgrupo es inferior al nivel mínimo para el tamaño del subgrupo.
- El nivel de profundidad del subgrupo es igual al número límite de divisiones permitido por CHAID.

Nivel de significancia descriptivo o valor p

El *valor* p es una medida de significancia estadística. En cada etapa del análisis, CHAID divide el árbol con la variable predictora que tenga el valor más bajo de probabilidad ajustado, o **valor p ajustado** (valor p multiplicado por el multiplicador de Bonferroni, Sección 2.6), siempre y cuando el valor p sea menor que el nivel de separación (generalmente 0.05). El valor p representa la probabilidad de que la relación de la muestra observada entre una predictora y la variable dependiente sea tanto o más extrema si las dos variables no estuvieran estadísticamente relacionadas o fueran independientes. Un valor p de 0.05 significa que la relación observada entre la predictora y la variable

dependiente sería tanto o más extrema solamente un 5 % de las veces si las variables fueran independientes. Por lo que la predictora con el valor p más bajo es la más probable para estar relacionada con la variable dependiente, es decir, entre más pequeño sea p, la predictora es más significativa estadísticamente.

Nivel de significancia para unión

Se refiere al nivel de dificultad para combinar categorías de una variable predictora, es decir, entre más alto sea este nivel, más difícil será que las categorías sean unidas. Si el nivel de unión es especificado como 1, ninguna categoría será unida. El valor más usual para el nivel de unión es de 0.05.

Nivel de significancia para separación

Se refiere al nivel α o error tipo I para que una variable sea considerada estadísticamente significativa. El valor más usual es de 0.05. Solamente las predictoras que tengan un valor p ajustado menor o igual a α son consideradas para dividir subgrupos. Un valor p de 0.05 significa que la relación entre la muestra observada entre una predictora y la variable dependiente solamente ocurrirá en un 5% de las veces, si las dos variables no están en realidad relacionadas. Mientras más bajo sea el valor p, más estadísticamente significativa es la relación.

Límite de profundidad

Se utiliza para delimitar el tamaño del árbol, es decir, cuántos niveles tendrá. Esta regla de paro puede permitir ahorrar bastante tiempo del algoritmo.

Tamaño de grupo antes de una división

Se refiere al mínimo número de casos u observaciones del subgrupo para permitir una división. Por ejemplo, si se asigna 50 a este parámetro, ningún subgrupo con menos de 50 casos será analizado, lo que lo convertirá en un nodo terminal del árbol.

Tamaño de grupo después de una división

Este parámetro asegura que los segmentos finales contengan al menos el número mínimo especificado de casos. Por ejemplo, si se asigna un valor de 40 a este parámetro, todos los nodos tendrán al menos 40 casos.

2.4 MÉTODO DE ANÁLISIS

El objetivo de CHAID estriba en particionar un conjunto de datos, formado por variables nominales u ordinales (cuantitativas categorizadas), una de las cuales es la variable dependiente en grupos más homogéneos. Los elementos restantes son predictoras y sus categorías pueden estar o no ordenadas.

El algoritmo CHAID asume que, la población representa un grupo heterogéneo con respecto a algún criterio de la variable dependiente y divide la población en dos o más grupos distintos basados en las categorías de la variable predictora más significante. El conjunto de segmentos mutuamente exclusivos y exhaustivos resultante, representa una segmentación de CHAID de la población. Dependiendo de si la variable dependiente es nominal u ordinal se elige el método de análisis. Si la variable dependiente tiene sólo dos categorías. los cálculos de los dos métodos son iguales.

El algoritmo general CHAID sigue el siguiente proceso para cada predictora:

1 Se forma una tabla de contingencia entre la predictora y la variable dependiente.

2. Para cada predictora se encuentra la mejor partición, uniendo categorías que sean homogéneas con el fin de llegar a la mejor agrupación de categorías

3. Se calcula el valor p ajustado de Bonferroni.

4. Las predictoras son comparadas y se escoge la que tenga el menor valor p ajustado, considerándola la mejor.

5. Los datos son subdivididos en dos o más grupos de acuerdo a la predictora elegida.

6. Cada uno de estos subgrupos es nuevamente analizado independientemente para producir otras subdivisiones para el análisis, mientras no se cumplan los parámetros de paro o no se encuentren predictoras significantes.

7. Cada nodo final se asigna a la clase que sea mayoritaria en ese nodo.

2.4.1 Método nominal

Cuando la variable dependiente es nominal, CHAID particiona los datos a través de pruebas de independencia χ^2, en subconjuntos que mejor describen a la variable dependiente. Esto permite la formación de divisiones múltiples. El algoritmo original de Kass también se puede usar con una variable dependiente ordinal, pero el orden es ignorado y los segmentos resultantes *podrían no ser los mejores*.

CHAID prueba todos los pares de categorías de cada variable predictora para ver si son homogéneas (i.e. que no son significativamente diferentes) con respecto a la variable dependiente. Las que son homogéneas son unidas en una sola. A continuación, se prueba si hay alguna categoría unida (categoría compuesta de 3 o más originales) que deba estar separada Estos procedimientos de unión y de división aseguran que, los casos pertenecientes al mismo segmento sean homogéneos y los casos en diferentes segmentos tiendan a ser heterogéneos con respecto al criterio de segmentación.

El tipo de cada predictora determina la agrupación permisible de sus categorías, así como la construcción de la tabla de contingencia con el nivel más alto de significancia de acuerdo a la prueba χ^2. Esto implica que se supone hay suficientes observaciones para la validación de esta prueba Si éste no es el caso, Kass (1980) recomienda utilizar otro criterio, como la prueba exacta de Fisher (Conover, 1980).

Sólo variables que tengan una relación estadísticamente significativa con la dependiente son viables para dividir un grupo paterno. El algoritmo CHAID selecciona la mejor predictora usando un mutiplicador de Bonferroni para ajustar un valor de probabilidad p que tome en cuenta la unión de categorías de las predictoras Se divide, primero, la

variable con mayor significancia estadística en términos de la variable dependiente, luego, cada uno de los grupos se fragmenta en cualquier número de categorías de acuerdo a la predictora. Este proceso continúa de manera independiente en cada grupo no analizado.

Los resultados de un análisis CHAID se resumen en forma de un diagrama de árbol y en tablas de ganancias. Los nodos del árbol corresponden a los subgrupos que se van formando. Al dividir los grupos en subgrupos más pequeños, el árbol expande nodos adicionales. Los nodos terminales del árbol son los segmentos finales.

2.4.2 Método ordinal

El método ordinal se utiliza para las variables dependientes que tienen categorías ordenadas. La segmentación se determina empleando los *puntajes (scores) de categoría* de la variable dependiente. El algoritmo ordinal de CHAID identifica segmentos que no sólo tienen una característica de interés sino que además son beneficiosos.

El algoritmo ordinal de CHAID fue desarrollado por Magidson en 1992 para considerar el valor relativo o puntaje de las categorías de la variable dependiente, siempre y cuando haya al menos tres. En este caso, la prueba de χ^2 para independencia es reemplazada por una prueba de χ^2 para asociación cero. Los segmentos buenos de un análisis ordinal CHAID diferirán de los malos con respecto a su valor esperado, en contraste con el algoritmo nominal que identifica segmentos basado en las diferencias en la distribución de la variable dependiente sin importar si éstas diferencias están relacionadas a la ganancia esperada

El método ordinal emplea un algoritmo iterativo de máxima verosimilitud para el cálculo de la significancia estadística (valor p) para cada predictora. Para estar seguro de que se obtiene la χ^2 correcta, deben hacerse suficientes iteraciones para que el algoritmo converja.

2.5 EL ALGORITMO

2.5.1 El algoritmo original

Para describir el algoritmo original, son necesarias algunas especificaciones de nomenclatura. Considere la variable dependiente (y) que se supone con un número de categorías $c >= 2$, se utilizan k variables predictoras ($x_1, x_2, \ldots, x_k$) y cada una tiene al menos dos categorías Los parámetros de paro son

$$\begin{aligned} \alpha_1 &= \text{nivel de significancia para unión.} \\ \alpha_2 &= \text{nivel de significancia para división.} \\ m &= \text{nivel de profundidad del árbol.} \\ n_1 &= \text{número de casos mínimo antes de dividir.} \\ n_2 &= \text{número de casos mínimo después de dividir.} \end{aligned}$$

Para cada una de las predictoras $x_1, x_2, \ldots, x_k$ hacer lo siguiente:

- **Paso 1:** Hacer una tabla de contingencia con las categorías de la variable predictora y con las de la variable dependiente

- **Paso 2:** Para cada par de categorías de la predictora (sólo considerando pares permitidos determinados por el tipo de predictora, ver Sección 2.3.1). calcular la estadística χ^2 para probar homogeneidad de proporciones entre este par y todas las categorías de la variable dependiente. Si se selecciona el método ordinal, se utiliza el procedimiento de la Sección 2.5.3 para calcular las χ^2.

- **Paso 3:** Encontrar el par de categorías de la predictora cuya subtabla $2 \times c$ sea la menos significativamente diferente, para esto. calcular el valor p para cada estadística χ^2. Si algunos pares no son significativos (su nivel de significancia descriptivo es mayor que el nivel de unión α_1), unir el par de categorías que tenga el valor χ^2 más bajo (o el valor p más alto) en una sola categoría. Considerar esta unión como

una sola categoría compuesta e ir al paso 4. Si todos los pares son significativos ir al paso 5.

- **Paso 4:** Para cada categoría compuesta (tres o más de las categorías originales unidas), probar si alguna de ellas debería estar separada, encontrando la división binaria más significante (restringida por el tipo de predictora), en la cual la agrupación puede ser separada Para lograrlo, utilizar el nivel de significancia descriptivo (valor p) de la estadística χ^2 que compara a ésta con las otras de la categoría unida. Si la χ^2es significativa (su valor p es menor que el nivel de unión α_1) separar dicha categoría de las demás. Si más de una cumple con esta condición, separar la que tenga la χ^2 más grande (o el valor p más pequeño). Si la significancia no es más grande que el nivel de unión, no efectuar ninguna división. Regresar al paso 2.

- **Paso 5:** Unir cualquier categoría que tenga menos observaciones que lo especificado por el tamaño mínimo de grupo después de dividir (n_2) con la categoría más similar de acuerdo a la menor estadística χ^2.

- **Paso 6:** Calcular el valor ajustado de Bonferroni (ver Sección 2.6), para cada predictora y aislar las significativas (el valor p ajustado menor que el nivel de separación α_2). Elegir la variable que tenga el valor p ajustado menor y subdividir los datos en las categorías (unidas) de la predictora elegida. Cada una de ellas forma un nuevo subgrupo del grupo paterno. Si ninguna predictora tiene un valor p significativo, no separar el grupo.

- **Paso 7:** Regresar al paso 1 para cada subgrupo de datos que todavía no haya sido analizado y que, contenga al menos tantas observaciones como lo especificado por el tamaño mínimo del subgrupo antes de dividir (n_1). Detener cuando todos los grupos hayan sido analizados o cuando contengan muy pocos casos.

El algoritmo de Kass busca reducir la tabla de contingencia original de $r \times c$ a la tabla más significante de $j \times c$ combinando categorías de la variable predictora (de acuerdo al

tipo de predictora), es decir, unir categorías de la predictora que sean homogéneas.

Kass planteó la solución conceptual a este problema como sigue: primero se calculan las estadísticas $T_j^{(i)}$, que son las estadísticas usuales χ^2 para el i-ésimo método de formar una tabla de $j \times c$, $j = 2, 3, ..., c$ (el rango de i depende del tipo de predictora). De esta manera, se construyen todas las tablas posibles en las que pudiera reducirse la tabla original. Entonces, se escoge $T_j^{(*)} = \max_i T_j^{(i)}$ que es la estadística χ^2 para la mejor tabla $j \times c$ para cada j posible. Se elige la más significante $T_j^{(*)}$. Para cada predictora se realiza el mismo procedimiento.

Kass observó que, en el caso de una predictora monotónica o una predictora dicotómica libre, la $T_j^{(*)}$ puede ser encontrada por el procedimiento de Fisher (1968). Este procedimiento de programación dinámica es computacionalmente de orden c^2. Cuando $c >= 3$, una predictora libre no puede tener sus categorías ordenadas en general, y el método de Fisher no puede ser aplicado. En su lugar, notó que la aplicación estándar de programación dinámica a problemas de tipo de permutaciones puede aplicarse, como lo hizo Dreyfus y Law (1977, p. 69). Esta solución es computacionalmente de orden 2^c.

Sin embargo, la programación dinámica no era computacionalmente práctica, por esta razón, Kass desarrolló una propuesta alternativa para encontrar la $T_j^{(*)}$. Consiste en buscar las estadísticas $T_j^{(*)}$ por pasos tal como se presenta en el algoritmo (pasos 2 y 3). Este procedimiento tiene paralelos en otros campos, como la regresión múltiple y en la regresión por pasos.

Para finalizar, se debe notar que el paso 4, aunque parece innecesario, no lo es, ya que, mientras el criterio para dividir una unión sea más estricto que el criterio para juntar dos categorías, el algoritmo impedirá una división que recree una situación previa. Esta condición también es necesaria para que converja el procedimiento a una solución estable en un número de pasos finitos. En la práctica, una unión es raramente dividida, pero debe estar permitido asegurar resultados más óptimos. En la Figura 2.1 se muestra el diagrama de flujo del algoritmo.

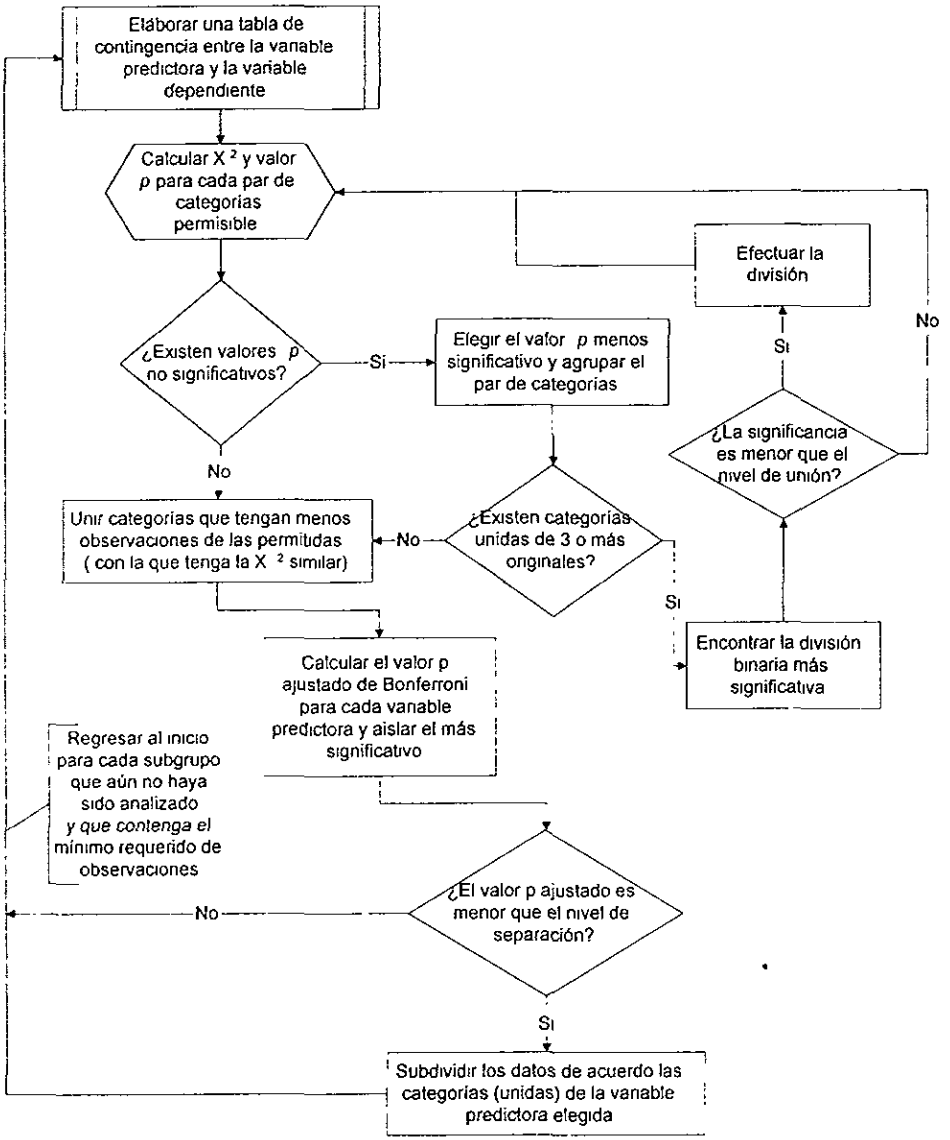

Figura 2.1: Diagrama de flujo del algoritmo de CHAID.

2.5.2 Estimación bajo el método nominal

Dada una tabla de contingencia entre una predictora A, el cual tiene I categorías y la variable dependiente B con J categorías, CHAID considera el siguiente modelo loglineal (Haberman, 1978; Agresti, 1984), véase Sección 1.3.2:

$$\mathrm{H}_1 : \ln(\frac{E_{ij}}{Z_{ij}}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB},$$

donde E_{ij} son los conteos esperados para la celda (i, j), bajo el modelo con restricciones, $Z_{ij} = 1/W_{ij}$; donde W_{ij} es el peso promedio muestral y las λ son parámetros del modelo que están sujetos a las identificaciones usuales

$$\sum_i \lambda_i^A = \sum_j \lambda_j^B = 0$$

y

$$\sum_i \lambda_{ij}^{AB} = \sum_j \lambda_{ij}^{AB} = 0.$$

Bajo el método nominal, CHAID considera el valor p asociado con la hipótesis nula de independencia para el modelo contra la hipótesis alternativa H_1

$$H_0 : \lambda_{ij}^{AB} = 0, \tag{2.1}$$

donde $i = 1, \ldots, I$ y $j = 1, \ldots, J$.

A continuación se probará que, la hipótesis nula del modelo loglineal equivale a la relación de independencia entre los renglones y las columnas de la tabla de contingencia, para lo cual, primero, se tomará el caso de muestreo multinomial, es decir, que un mismo individuo es clasificado en dos criterios. Se observa que si las columnas y renglones fueran independientes

$$p_{ij} = p_{i+}p_{+j},$$

donde

$$\begin{aligned} p_{ij} &= \text{probabilidad de que una observación caiga en el } i\text{-ésimo} \\ &\quad \text{renglón y en la } j\text{-ésima columna.} \\ p_{i+} &= \text{probabilidad de que una observación caiga en el renglón } i. \\ p_{+j} &= \text{probabilidad de que una observación caiga en la columna } j, \end{aligned}$$

por lo que los conteos esperados en la tabla son

$$E_{ij} = np_{i+}p_{+j}$$

donde n es el número de observaciones totales de la tabla, entonces

$$\ln E_{ij} = \ln n + \ln p_{i+} + \ln p_{+j}.$$

Tomando apropiadamente los valores de los parámetros λ, se obtiene el modelo loglineal. En otras palabras, si los renglones y las columnas son independientes, un modelo loglineal de la forma

$$\ln E_{ij} = \lambda + \lambda_i^A + \lambda_j^B \tag{2.2}$$

es válido. Sin embargo, si se basa el análisis en un modelo loglineal, es muy importante saber si el modelo (2.2) implica independencia entre renglones y columnas.

Teorema 2.1 *Para muestreo multinomial en una tabla* $I \times J$, $\ln E_{ij} = \lambda + \lambda_i^A + \lambda_j^B$, $i = 1, \ldots, I$, $j = 1, \ldots, J$ *si y sólo si* $p_{ij} = p_{i+}p_{+j}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$.

Demostración. Ya se demostró que independencia implica el modelo loglineal. Ahora dado el modelo loglineal se tiene que

$$E_{ij} = e^{\lambda + \lambda_i^A + \lambda_j^B}.$$

Sea $a = e^{\lambda}$, $a_i^A = e^{\lambda_i^A}, a_j^B = e^{\lambda_j^B}$. Sea $a_+^A = \sum_{i=1}^{I} a_i^A$ y $a_+^B = \sum_{j=1}^{J} a_j^B$. Note que

$$\begin{aligned} p_{ij} &= E_{ij}/n = aa_i^A a_j^B/n, \\ p_{i+} &= aa_i^A a_+^B/n, \\ p_{+j} &= aa_+^A a_j^B/n, \end{aligned}$$

y

$$1 = p_{++} = aa_+^A a_+^B/n.$$

Sustituyendo

$$\begin{aligned} p_{i+}p_{+j} &= aa_i^A a_+^B aa_+^A a_j^B/n \\ &= (aa_i^A a_j^B/n)(aa_+^A a_+^B/n) \\ &= (aa_i^A a_j^B/n) \\ &= p_{ij} \end{aligned}$$

Por lo tanto el modelo loglineal implica independencia. ■

Para un producto de multinomiales, es decir, cuando se tienen varias poblaciones clasificadas con respecto a un criterio

$$E_{ij} = n_{i+}p_{ij} \tag{2.3}$$

donde n_{i+} es el número de observaciones en el renglón i, y el modelo loglineal

$$\ln E_{ij} = \ln n_{i+} + \ln p_{ij}$$

se obtiene fácilmente. Ahora, considerando el modelo bajo H_0 se observa que si $\pi_j =$

$p_{1j} = \ldots = p_{Ij}$ para toda $j = 1, \ldots, J$ entonces

$$E_{ij} = n_{i+}\pi_j$$

Teorema 2.2 *Para un producto de muestreo multinomial en una tabla de $I \times J$ donde los renglones son muestras independientes,* $\ln E_{ij} = \lambda + \lambda_i^A + \lambda_j^B$, $i = 1, \ldots, I$, $j = 1, \ldots, J$ *si y sólo si* $p_{1j} = \ldots = p_{Ij}$, $j = 1, \ldots, J$.

Demostración. Si para cada j las probabilidades p_{ij} son iguales, se tiene que $E_{ij} = n_{i+}\pi_j$ y $\ln E_{ij} = \ln n_{i+} + \ln \pi_j$. Tomando $\lambda = 0$, $\lambda_i^A = \ln n_{i+}$ y $\lambda_j^B = \ln \pi_j$ se muestra que el modelo loglineal es válido.

De manera inversa, si $\ln E_{ij} = \lambda + \lambda_i^A + \lambda_j^B$ entonces $E_{ij} = a a_i^A a_j^B$ donde $a = e^{\lambda}$, $a_i^A = e^{\lambda_i^A}$, $a_j^B = e^{\lambda_j^B}$. Nótese que $p_{i+} = 1$ así que de (2.3), $E_{i+} = n_{i+}$ y

$$n_{i+} = a a_i^A a_+^B.$$

Como $p_{ij} = E_{ij}/n_{i+}$

$$\begin{aligned} p_{ij} &= a a_i^A a_j^B / n_{i+} \\ &= a a_i^A a_j^B / a a_i^A a_+^B \\ &= a_j^B / a_+^B \end{aligned}$$

Esto es verdad para cualquier i, así que $a_j^B/a_+^B = p_{1j} = p_{2j} = \ldots = p_{Ij}$, $j = 1, \ldots, J$. ■

Cuando no se utiliza una variable de peso, entonces las Z_{ij} son iguales a uno y los conteos de celdas esperados en CHAID, son calculados usando el algoritmo IPF (Iterative Proportional-Fitting) de Deming y Stephan (1940), el cual converge en una iteración (Magidson, 1993b) y los estimados tiene una expresión explícita. Si se ocupa una variable de peso, el algoritmo IPF es modificado para usar las Z_{ij} como valores iniciales. Este algoritmo es conocido como WLM (Weighted Loglinear Modeling, véase Magidson, 1987;

Clogg y Eliasin, 1987), el cual requiere de varias iteraciones para alcanzar la convergencia (Magidson, 1993b).

Si la variable dependiente es estratificada (por ejemplo, si se utiliza un 1% muestral de personas que no responden pero un 100 % de los que sí responden y se asigna una variable de peso igual a 100 para los que no respondieron y 1 para los que respondieron), el algoritmo WLM lleva a los mismos estimados para los conteos esperados que si no fuera utilizado ningún peso (Magidson, 1987).

Después de que se alcanzó la convergencia, se calcula una estadística χ^2 para asegurar la bondad de ajuste entre los conteos esperados y los observados. Las estadísticas que se utilizan son la χ^2 de Pearson y el cociente de verosimilitud (G^2). La estadística de cociente de verosimilitud es igual a

$$G^2 = 2\sum_i \sum_j O_{ij} \ln\left(\frac{O_{ij}}{\hat{E}_{ij}}\right),$$

donde O_{ij} denotan los conteos de las celdas y $\hat{E}_{ij}$ denota los conteos esperados estimados.

Alternativamente, la estadística χ^2 de Pearson es

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}.$$

El valor p no ajustado se obtiene de la distribución χ^2 con $(I-1)(J-1)$ grados de libertad.

2.5.3 Estimación bajo el método ordinal

Cuando están disponibles puntajes para cuantificar las categorías de la variable dependiente B, Magidson (1992) desarrolló una alternativa a la hipótesis de independencia del método nominal. En lugar de probar la hipótesis nula (H_0) de independencia, CHAID

utiliza el *modelo de asociación Y* (Agresti, 1984; Magidson, 1992, 1993a, 1993b)

$$H_1' : \ln(\frac{E_{ij}}{Z_{ij}}) = \lambda + \lambda_i^A + \lambda_j^B + x_i(y_j - \bar{y}) \tag{2.4}$$

donde y_j denota el puntaje para la categoría j de B, x_i denota coeficientes desconocidos para las y_j, y $\bar{y}$ denota el puntaje promedio para la variable dependiente.

Este modelo se sujeta a las condiciones de identificación del modelo de independencia para el método nominal

$$\sum_i \lambda_i^A = \sum_j \lambda_j^b = 0$$

y

$$\sum_i \lambda_{ij}^{AB} = \sum_j \lambda_{ij}^{AB} = 0$$

y además la siguiente condición de identificación para los parámetros x_i

$$\bar{x} = 0 \tag{2.5}$$

donde $\bar{x}$ denota el promedio de las x_i. Bajo el modelo H_1', la hipótesis nula de independencia (2.1) se convierte en

$$H_0' : x_1 = x_2 = .. \ = x_I \tag{2.6}$$

ya que si $x_1 = x_2 = \ldots = x_I = x$, entonces $\bar{x} = x$ y por (2.5) $x = 0$, por lo que en (2.4), la expresión $x_i(y_j - \bar{y})$ es igual a cero y se obtiene el modelo de independencia de la sección de estimación nominal (2.2).

La hipótesis nula (2.6) es equivalente a

$$H_0' : \mu_1 = \mu_2 = . \ . = \mu_I$$

donde

$$\mu_i = \sum_j y_j P_{j \mid i}$$

y $P_{j|i}$, es la probabilidad condicional de pertenecer a la categoría j de A, dado que una observación está clasificada en la categoría de A.

Para probar la equivalencia de la ecuación B obsérvese que bajo el modelo de independencia

$$p_{ij} = p_{i+}p_{+j}$$

entonces

$$P_{j|i} = \frac{p_{ij}}{p_{i+}} = \frac{p_{i+}p_{+j}}{p_{i+}} = p_{+j}.$$

con lo que

$$\begin{aligned} \mu_i &= \sum_j y_j P_{j|i} \\ &= \sum_j y_j p_{+j}. \end{aligned} \tag{2.7}$$

La expresión (2.7) no depende de i por lo que $\sum_j y_j p_{+j} = \mu_1 = \mu_2 = \ldots = \mu_I$.

La estimación bajo el método ordinal involucra dos pasos. En el primero, CHAID calcula con el algoritmo de máxima verosimilitud (Goodman 1979,1985; Magidson, 1992), los estimados para los conteos de celdas esperados bajo el modelo H_1'. Entonces, para estos conteos estimados se prueba independencia usando la siguiente estadística χ^2 de cociente de verosimilitud:

$$G^2 = 2\sum_i \sum_j O'_{ij} \ln\left(\frac{O'_{ij}}{\hat{E}_{ij}}\right) \tag{2.8}$$

donde O'_{ij} denota los conteos estimados bajo el modelo H_1' y E'_{ij} denota los conteos esperados estimados bajo independencia (i. e., bajo H_0). Los grados de libertad para la estadística χ^2 de cociente de verosimilitud en (2.8) son $I - 1$. La estadística G^2 mide sólo la cantidad de no independencia que es relevante a los puntajes y_j, esto es, sólo la dependencia que representa desviaciones de la hipótesis nula.

2.6 SIGNIFICANCIA DE LAS PREDICTORAS

Cuando dos o más categorías de la variable predictora se hayan unido, se utiliza el ajuste de Bonferroni en el cálculo del valor p o significancia de cada predictora. CHAID calcula un valor p no ajustado siguiendo los pasos antes mencionados en el algoritmo y lo modifica multiplicándolo por el multiplicador Bonferroni B^* para estimar un valor p ajustado (Kass, 1980).

El procedimiento consiste en determinar el número de formas en que una predictora de un tipo dado con I categorías puede ser reducida a J grupos $(1 < J < I)$ y utilizarlo en la desigualdad de Bonferroni *(Apéndice A)* para obtener un límite para el nivel de significancia; a este número se le conoce como *multiplicador de Bonferroni.* Este multiplicador hace un ajuste de la inferencia simultánea asociada con el hecho de que CHAID examina diferentes maneras de juntar categorías y seleccionar la que tenga la más alta significancia en la tabla. Sin embargo, si no se unieron categorías, el multiplicador es igual a 1 y el valor ajustado de Bonferroni es igual al valor p no ajustado.

Por ejemplo, considere que, A es una variable predictora libre de cinco categorías que se reduce a dos y que se emplea el método nominal y B es la variable dependiente. Sea α el error tipo I deseado, asociado con la prueba de independencia en una tabla bivariada formada por una tabulación cruzada entre A y B. Esto es, α denota la probabilidad de rechazar la hipótesis nula de independencia cuando de hecho, debía aceptarse. Hay 15 formas de dicotomizar las categorías de X. Si se efectuaran las pruebas para cada una de estas combinaciones y éstas fueran independientes unas de otras, la probabilidad de cometer un error tipo I en una o más de estas pruebas es igual a:

$$1 - (1 - \alpha)^{15}, \tag{2.9}$$

la cuál es más grande que α. Por la desigualdad de Bonferroni, la probabilidad en (2.9) es aproximadamente 15α cuando α es pequeño. En general, si B^* iguala el número de maneras de terminar con $J \leq I$ categorías, entonces para una α se tiene que

$$1 - (1 - \alpha)^{B^*} \leq B^* \alpha.$$

Por lo tanto, si se realizan 15 pruebas, el valor p debe ser al menos $\frac{\alpha}{15}$ para ser considerado significativo al nivel α

El ajuste de Bonferroni se aplica en los métodos de análisis nominal y ordinal. La cantidad de ajuste depende del tipo de predictora. Las fórmulas para calcular estos multiplicadores para los tres tipos de predictoras permitidas por CHAID son:

Predictoras Monotónicas. La variable predictora pasa de I a J categorías, por lo que se deben acomodar I categorías originales en J nuevas categorías (recuérdese que sólo se pueden unir categorías contiguas, puesto que la predictora es monotónica) que no deben estar vacías. Nótese que dividiendo las J nuevas categorías hay $J - 1$ espacios entre ellas y similarmente para separar las I categorías originales hay $I - 1$ espacios o divisiones. Si sólo se pueden unir categorías adyacentes, entonces los espacios entre ellas deben desaparecer. La condición de que ninguna celda esté vacía impone la restricción de que no haya espacios adyacentes. De los $J - 1$ espacios dejados por las J categorías nuevas, $I - 1$ son ocupados por los espacios de las I categorías originales; de modo que se tienen $\begin{pmatrix} I-1 \\ J-1 \end{pmatrix}$ maneras de acomodar $I - 1$ espacios originales en $J - 1$ espacios nuevos Por lo que el multiplicador para las variables monotónicas es el multiplicador binomial:

$$B_{monotónico} = \begin{pmatrix} I-1 \\ J-1 \end{pmatrix}$$

Predictoras Libres. La determinación del multiplicador de Bonferroni para predictoras libres es equivalente a la solución de un problema de ocupación en el cual se desea acomodar r bolas en n celdas sin dejar una celda vacía. Bajo este planteamiento, Feller

desarrolló una fórmula para calcular el número de distribuciones posibles (véase Feller, 1968, p.101):

$$\text{Número de Distribuciones (ND)} = \sum_{i=0}^{n-1}(-1)^i \binom{n}{i}(n-i)^r. \tag{2.10}$$

En el caso de las predictoras libres, se tienen que acomodar I categorías originales en J categorías nuevas, sin dejar ninguna vacía. Entonces (2.10) toma la forma de

$$\text{ND} = \sum_{i=0}^{J-1}(-1)^i \binom{J}{i}(J-i)^I. \tag{2.11}$$

La expresión (2.11) toma en cuenta el orden de las categorías, por ejemplo, si se tenían 3 categorías originales numeradas del 1 al 3 y se unieron para formar 2; el arreglo (1-2. 3) es considerado diferente al (3, 1-2). Sin embargo, para el análisis CHAID es irrelevante el orden de las categorías finales por lo que se divide el lado derecho de (2.11) entre $J!$. Por tanto, el número de formas de acomodar I categorías en J categorías nuevas para una predictora libre es

$$B_{libre} = \sum_{i=0}^{I-1}(-1)^i \frac{(J-i)^I}{i!(J-i)!}.$$

Predictoras Flotantes. El multiplicador Bonferroni se obtiene como una extensión del caso monotónico. La categoría flotante puede permanecer sola o unirse a otras. Si no se combina, entonces está fija y ocupa una de las J categorías disponibles y quedan por acomodar $I-1$ categorías originales en las $J-1$ categorías restantes que como siguen una escala ordinal, se sigue el mismo razonamiento del multiplicador de Bonferroni para variables monotónicas, y se obtienen $\binom{I-2}{J-2}$ formas de acomodarlas.

Cuando la categoría flotante se combina con otra(s), considere primero acomodarla en cualquiera de la J categorías nuevas. Las restantes $I-1$ categorías (sin la flotante) habrá que acomodarlas en J categorías también. Puesto que la categoría flotante se va

a unir con otra, siguiendo el razonamiento de las variables monotónicas, para este caso, hay $J \times \begin{pmatrix} I-2 \\ J-1 \end{pmatrix}$ posibilidades de unión. Tomando en cuenta los dos casos descritos se obtiene el siguiente multiplicador:

$$\begin{aligned} B_{flotante} &= \begin{pmatrix} I-2 \\ J-2 \end{pmatrix} + J \begin{pmatrix} I-2 \\ J-1 \end{pmatrix} \\ &= \frac{(J-1)}{(I-1)} \begin{pmatrix} I-1 \\ J-1 \end{pmatrix} + J\frac{(I-J)}{(I-1)} \begin{pmatrix} I-1 \\ J-1 \end{pmatrix} \\ &= \left[\frac{(J-1)}{(I-1)} + \frac{J \times (I-J)}{(I-1)}\right] \times \begin{pmatrix} I-1 \\ J-1 \end{pmatrix} \\ &= \left[\frac{(J-1)+J \times (I-J)}{(I-1)}\right] \times B_{monotonico} \end{aligned}$$

Si no hay unión de categorías, el multiplicador de Bonferroni para cualquier variable predictora es igual a 1, y el valor p ajustado es igual al valor p no ajustado. Si todas las categorías de una variable predictora se unen en una sola, al valor p ajustado se le asigna el valor de 1 para evitar que se tome en cuenta a ésta en el análisis, ya que no sirve para identificar más diferencias en términos de la variable dependiente.

2.7 TABLAS DE GANANCIAS

Otras de las extensiones actuales de CHAID, son las *tablas de ganancias.* En ellas se muestran los resultados de la segmentación y proporcionan información relevante de este proceso. Son de gran utilidad en la validación del análisis CHAID gracias a que, presentan las características y virtudes del análisis final. Existen dos tipos de tablas de ganancias: *las detalladas* y *las resumen.*

2.7.1 Tabla de ganancias detallada

Este tipo de tablas contienen un renglón para cada nodo terminal o segmento final, del diagrama de árbol. Se pueden generar *la tabla de ganancias detallada nodo por nodo* y *la tabla de ganancias detallada acumulativa*. En ambas se tienen las mismas columnas, pero en la primera la información es individual, mientras que, en la segunda es acumulada. Las columnas que contiene la tabla de ganancias detallada nodo por nodo son:

Nodo	Número de identificación que corresponde al nodo terminal.
Nodo: n	Número de casos en el nodo o segmento.
Nodo: %	Porcentaje de casos del nodo en relación con el total.
Resp: n	Número de respuestas que caen en este nodo.
Resp: %	Porcentaje de respuesta del nodo en relación al total de respuestas.
Ganancia (%)	Porcentaje de respuesta en el nodo.
Indicador (%)	Puntaje o ganancia promedio de respuesta para ese segmento en relación al puntaje o ganancia promedio del total de la muestra.

El número de identificación de cada nodo está dado por el número del nodo en el árbol, numerado de izquierda a derecha a partir del nodo raíz, el cual es numerado como "0". El porcentaje de casos del nodo en relación con el total se obtiene de dividir el número de casos en el nodo (*Nodo: n*) entre el total de la muestra. Las columnas *Resp: n* y *Resp: %* hacen referencia a las "respuestas" de la muestra. Las *respuestas* se entienden como los casos que pertenecen a la categoría objetivo o de interés de la variable dependiente. La ganancia se obtiene dividiendo el número de respuestas en el nodo (*Resp: n*) entre el tamaño del nodo (*Nodo: n*). La ganancia promedio del total de la muestra es el resultado de la división de la respuestas totales de la muestra entre el tamaño de la misma. Por lo que, el *Indicador (index)* proviene de dividir la ganancia del nodo (*Ganancia: %* o *gain*) entre la ganancia promedio del total de la muestra. El indicador no tiene ningún significado cuando la ganancia promedio para toda la muestra es menor o igual a 0. Un

indicador de, por ejemplo, 152% indica que la tasa de respuesta para este nodo es 52% mayor que el promedio.

La tabla de ganancias acumulativa ofrece la misma información que la anterior, sólo que, acumulada. Los datos se calculan igual a como se describió previamente (un ejemplo de este tipo se presenta en la Sección 3.2). En suma en este tipo de tablas se encuentra:

Nodo	Número de identificación que corresponde al nodo terminal.
Nodo: n	Número de casos acumulados hasta el nodo o segmento.
Nodo: %	Porcentaje de casos acumulados en relación con el total.
Resp: n	Número de respuestas acumuladas hasta el nodo.
Resp: %	Porcentaje de respuestas acumuladas en relación al total de respuestas.
Ganancia (%)	Porcentaje de respuesta combinada hasta el nodo.
Indicador (%)	Puntaje o ganancia promedio acumulada de respuesta en los segmentos en relación al puntaje o ganancia promedio del total de la muestra.

Para el método nominal, la columna *Ganancia* refleja la distribución de porcentaje para la categoría 1 de la variable dependiente.

La información que proporcionan las tablas de ganancias detallada permiten escoger los segmentos con mejor respuesta, determinar su respuesta combinada y establecer una mejor estrategia para aprovechar los recursos disponibles.

2.7.2 Tabla de ganancias resumen

Este tipo también se conoce como tabla de ganancias de cuantiles. Resume los segmentos del árbol en la forma de un arreglo de cuantiles, es decir, presenta resultados acumulativos en porcentajes fijos de igual tamaño con respecto al total de la muestra (conocidos aquí como cuantiles). Los cuantiles se van formando acumulando casos hasta llegar al porcentaje fijo por cuantil, respetando el orden de mayor ganancia en un nodo. La tabla

de resumen contiene

Nodos	Número de identificación de los nodos que forman el cuantil.
Percentil	Porcentaje fijo acumulado del cuantil.
Percentil: n	Número de casos acumulados para el cuantil.
Resp: n	Número de respuestas acumuladas hasta ese cuantil.
Resp: %	Porcentaje del total de respuestas acumuladas hasta ese cuantil.
Ganancia (%)	Porcentaje de respuesta acumulado hasta el cuantil.
Indicador (%)	Puntaje o ganancia promedio de respuesta acumulada de un cuantil en relación al puntaje o ganancia promedio del total de la muestra.

La *tabla de ganancias resumen* se puede utilizar como una herramienta de búsqueda dinámica que juega las veces de la estadística incremental R^2 en la regresión por pasos. Para este propósito, es necesario construir una tabla de ganancias resumen en cada nivel del árbol. De esta manera, se podrán evaluar los pasos de la segmentación hasta llegar al árbol final (veáse el ejemplo de clasificación de créditos, Sección 3.2).

2.8 PUNTAJES DE CATEGORÍAS

Los puntajes de categoría son calificaciones numéricas que se asignan a las diferentes categorías de la variable dependiente, con el fin de sopesar sus costos relativos o sus ganancias. Estos puntajes afectan el análisis CHAID, porque están ligados con el aseguramiento de la significancia estadística. Sin embargo, esto sólo se aplica si la variable dependiente es ordinal y politómica. Si la variable dependiente es tratada como nominal, el empleo o no de puntajes no afecta al análisis CHAID. Si la variable dependiente es dicotómica, no se utilizan los puntajes asignados de categorías para determinar el criterio de segmentación. En la siguiente sección se estudiarán los dos casos más detalladamente.

En CHAID, los puntajes de categorías son utilizados para:

- Producir tablas de ganancias basadas en diferentes conjuntos de puntajes y escoger

la mejor estrategia de acción.

- Reflejar beneficios numéricos de los costos o ganancias asociados con las categorías y cuantificar el valor de una estrategia, si la variable es ordinal.
- Contrastar o incrementar la distancia entre dos categorías.
- Unir u omitir categorías de la variable dependiente.

2.8.1 Método nominal

Los puntajes asignados a una variable dependiente nominal no afectan la determinación de la significancia estadística, por lo que, la segmentación bajo el método nominal es la misma con o sin puntajes. Sin embargo, se pueden utilizar los puntajes de categorías para construir tablas de ganancias promedio que se esperarían obtener en cada nodo terminal.

Las tablas de ganancias promedio ofrecen la siguiente información:

Nodo por nodo	
Nodo	Número de identificación que corresponde al nodo terminal.
Nodo: n	Número de casos en el nodo o segmento.
Nodo: %	Porcentaje de casos del nodo en relación con el total.
Ganancia ($)	Beneficio o pérdida promedio de cada elemento en el nodo.
Indicador (%)	Puntaje o ganancia promedio para ese segmento en relación al puntaje o ganancia promedio del total de la muestra

Acumulativa	
Nodo: n	Número de casos acumulados hasta el nodo o segmento.
Nodo: %	Porcentaje de casos acumulados en relación con el total.
Ganancia ($)	Beneficio o pérdida combinada hasta el nodo.
Indicador (%)	Puntaje o ganancia promedio acumulada para los segmentos en relación al puntaje o ganancia promedio del total de la muestra.

Tanto las tablas de ganancias promedio como las de ganancias detallada ofrecen información semejante, excepto en las columnas *Resp(n y %)*, ya que, como no se está interesado en el número de respuestas, no se despliega. Por el contrario, dado que el interés está en la ganancia promedio, en lugar de la información anterior, se presenta *Ganancia($)* en términos monetarios. Los valores de esta columna se obtienen al ponderar los puntajes de categoría con los porcentajes de participación de cada categoría en el nodo. El indicador se calcula de igual forma que en las tablas detalladas al dividir la ganancia promedio de cada nodo entre la ganancia promedio del total de la muestra, la cual se obtiene al ponderar los puntajes de categoría entre el porcentaje de participación de las categorías en la muestra total.

2.8.2 Método ordinal

Si la variable ordinal es dicotómica, los puntajes de categoría no afectan el criterio de segmentación; sin embargo, los puntajes promedio pueden ser utilizados en las tablas de ganancias. En cambio, si la variable es politómica, la significancia estadística está determinada por estos puntajes. Al alterar esos puntajes, el criterio de segmentación cambia y la segmentación resultante puede ser diferente. Las tablas de ganancias cuando se trata de una variable ordinal están basadas en los puntajes promedio de categoría. Si no se está seguro de los puntajes a utilizar, es recomendable realizar análisis separados utilizando diferentes puntajes y comparar las segmentaciones resultantes.

2.9 MATRIZ DE CLASIFICACIÓN

Una de las extensiones que ha tenido CHAID es la *matriz de clasificación* de datos, que apareció con el programa de "AnswerTree de SPSS[1]".

La matriz de clasificación de datos tabula los casos que fueron mal clasificados al final o en cada etapa del análisis CHAID, con el fin de calcular el error de clasificación que se

[1] AnswerTree es marca registrada de SPSS, 1998

cometió y determinar o evaluar si el modelo es pobre o no. Recuérdese que cada nodo se clasifica en la categoría de la variable dependiente con mayor frecuencia en el nodo. Esta matriz es una tabla cuyos renglones representan las categorías predichas o asignadas por CHAID a los valores de la muestra y en las columnas aparecen las categorías reales de los valores (Tabla 2.1).

		Categoría real				
		Cat. 1	*Cat. 2*	.	*Cat. m*	**Total**
Categoría predicha	*Cat. 1*	F_{11}	F_{12}		F_{1m}	d_1
	Cat 2	..		.	..	...
		..		.	...	...
	Cat m	F_{21}	.	...	F_{mm}	d_m
	Total	r_1	...	...	r_m	n
Estimador de Riesgo		0.0000				
ES del Estimador de Riesgo		0.0000				

Tabla 2.1: Matriz de clasificación.

donde

F_{ij} = número de observaciones que se clasificaron en la categoría i cuando en realidad están en la categoría j.

d_i = número de observaciones que fueron clasificadas en la categoría i por el algoritmo.

r_i = número de observaciones que en la realidad están clasificadas en la categoría i.

n = número de observaciones en la muestra.

m = número de categorías de la variable dependiente.

$i = 1, \ldots, m,\ j = 1, \ldots, m$

La diagonal de esta tabla representa las observaciones bien clasificadas, es decir,

aquellas que fueron clasificadas en una categoría y que pertenecían a ella en la realidad. Las malas clasificaciones son todas aquellas que están fuera de la diagonal y, por tanto, se presenta un desacuerdo entre lo asignado por CHAID y la realidad.

También en esta tabla aparecen dos estimadores que indican qué tan bien resultó el análisis: el *estimador de riesgo* y el *error estándar del estimador de riesgo*. El primero indica en que porcentaje falló el análisis y se calcula como la proporción de casos en la muestra que fueron clasificados incorrectamente por el algoritmo:

$$\text{Estimador de Riesgo (ER)} = \frac{\sum_{i \neq j} F_{ij}}{\mathbf{n}}.$$

El error estándar del estimador de riesgo se utiliza para conocer la dispersión que tendrá el estimador de riesgo si se utilizaran otros datos y se calcula considerando p igual al Estimador de Riesgo. Entonces se tiene

$$\begin{aligned} \text{ES del Estimador de Riesgo} &= \sqrt{\frac{p(1-p)}{n}} \\ &= \sqrt{\frac{\text{ER}(1-\text{ER})}{\mathbf{n}}}. \end{aligned}$$

Capítulo 3

APLICACIONES

3.1 CLASIFICACIÓN DE CRÉDITOS

Una de las aplicaciones más importantes de CHAID ha sido la clasificación de crédito o toma de decisiones acerca de quién probablemente pagará un préstamo y quién no. Gracias a este enfoque en la resolución del problema de crédito, se pueden identificar perfiles de grupos homogéneos con alto y bajo riesgo así como clasificar futuras solicitudes de acuerdo al perfil del solicitante y determinar las reglas para predicciones individuales.

En este ejemplo, se analizará un grupo de datos de solicitantes de crédito con el fin de clasificarlos de acuerdo a si representan un riesgo de pérdida desde el punto de vista del prestamista. Para este propósito se utilizará el conjunto de datos contenido en el archivo *credit.sav* que está disponible en los discos de instalación del paquete "AnswerTree de SPSS".

Este ejemplo está dividido en tres secciones principales: en la primera, correspondiente a la construcción del árbol, se sigue paso a paso el *algoritmo* de CHAID presentado en la Sección 2.5, con el fin de ejemplificar y explicar detalladamente su ejecución. La segunda, se enfoca en la construcción y uso de las *matrices de clasificación* para la evaluación del árbol y, la tercera sección, ocupa las *tablas de ganancias* para resumir los resultados de la segmentación.

Datos

Los datos representan una encuesta con 323 solicitantes de crédito. Las variables para este estudio se enlistan en la Tabla 3.1:

Nombre	Descripción	No. de categorías
Ocupación	Ocupación del solicitante.	5
Frec_pago	Pago del salario: semanal o mensual.	2
Edad	Edad del solicitante.	3
Amex	¿Posee una tarjeta American Express?	2
Tipo_crédito	Clasificación del tipo de crédito	2

Tabla 3.1: Tabla de variables de un estudio de créditos.

En la siguiente tabla se muestran las categorías de cada una de las variables, así como la frecuencia de cada categoría en el conjunto de datos. La variable *edad*, aún cuando es una variable continua, se categoriza dividiendo la escala de posibles valores en tres intervalos (0-24, 25-35, >35).

Variable	Categorías	Frecuencia
Ocupación	Gerencial	39
	Profesional	158
	Oficinista	47
	Técnico	41
	Obrero	38
Frec_pago	Semanal	165
	Mensual	158
Edad	Joven (<25 años)	187
	Maduro (25-35)	79
	Viejo (>35)	57
Amex	No	167
	Si	156
Tipo_crédito	Malo	168
	Bueno	155

Tabla 3.2: Características de las variables del estudio de créditos.

El archivo contiene una variable dependiente u objetivo (*tipo_crédito)* y cuatro variables predictoras: *edad, amex, frec_pago y ocupación.* La variable *edad* es ordinal,

por lo que, en este análisis es considerada como monotónica, mientras que, la variable *ocupación* es nominal o libre. Cabe notar que, las variables *frec_pago* y *amex* pueden ser tomadas como monotónicas o libres sin que esto afecte el comportamiento del análisis puesto que tienen 2 categorías. Sin embargo, sí existe una diferencia para el algoritmo en escoger si la variable dependiente es ordinal o nominal, siempre y cuando, la variable dependiente tenga más de dos categorías. Como *tipo_crédito* sólo tiene dos categorías, el análisis es el mismo si la variable es tomada como nominal u ordinal.

3.1.1 Construcción del árbol

Después de definir la variable objetivo y las variables predictoras, el siguiente paso es fijar los parámetros de paro que CHAID utilizará para este análisis. El número de niveles de profundidad que tendrá el árbol se fijará en 3 y es importante considerar que dado que la muestra es relativamente pequeña es necesario ajustar los criterios de tamaño de los grupos; el tamaño mínimo de un subgrupo antes de una división, es decir, el número de casos mínimo del nodo paterno será de 25 y el tamaño mínimo de un subgrupo después de una división o tamaño del nodo hijo se fijará en 1. En resumen, los parámetros de paro para el algoritmo son:

Nivel de profundidad	=	3
Tamaño mínimo del subgrupo antes de una división	=	25
Tamaño mínimo del subgrupo después de una división	=	1

También se deben especificar ciertos parámetros para el funcionamiento del algoritmo. Estos parámetros son:

Nivel de significancia para unión (α_1)	=	0.05
Nivel de significancia para separación (α_2)	=	0.05

Nivel cero

Una vez establecidos los criterios de paro y especificaciones para el algoritmo, se procederá a construir el árbol siguiendo los pasos descritos en el capítulo anterior (Sección 2.5). En la Figura 3.1 se muestra el nodo raíz del árbol con que comienza el análisis. Para este nodo, se presentará con detalle la ejecución del algoritmo.

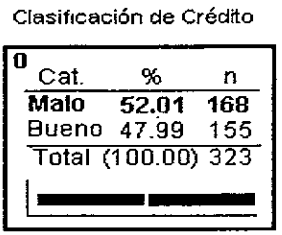

Figura 3.1: Nodo raíz del árbol de clasificación de crédito.

El diagrama de la Figura 3.1 procede del paquete AnswerTree de SPSS y es una representación gráfica valiosa que facilita la comprensión de la segmentación. La información que se despliega en este diagrama se compone de los siguientes elementos: en la primera columna, se presentan las categorías de la variable dependiente, la segunda columna muestra el porcentaje de casos clasificados en cada categoría con respecto al total de cada nodo y, finalmente, en la tercera columna, se muestra el número de casos correspondiente a cada categoría. En la esquina superior izquierda se encuentra un número que indica la enumeración de los nodos del árbol para facilitar su localización. Se comienza asignando "0" al nodo raíz y se continúa enumerando cada nodo de acuerdo a su posición en el árbol, guardando el orden de izquierda a derecha y de arriba hacia abajo.

La información del nodo raíz representa la división de los casos del total de la muestra, con base en la variable dependiente; dicho en cifras, hay 155 casos clasificados como créditos buenos (representan un bajo riesgo para el prestamista), que corresponden al

47.99% del total de casos y 168 clasificados como créditos de alto riesgo o malos, sumando 323 casos totales.

El nodo también presenta una gráfica (Figura 3.2), para visualizar la distribución de las categorías en el nodo. En este conjunto de datos, están igualmente distribuidos los casos de riesgos de créditos buenos (47.99%) y los de riesgos de créditos malos (52.01%). La gráfica 3.3 ejemplifica también esta proporción. Si se asignasen todos los casos de la muestra a la clase mayoritaria (crédito malo), se caería en un porcentaje de error del 47.99% referente al porcentaje de aquellos solicitantes que se estarían clasificando como malos cuando en realidad son buenos.

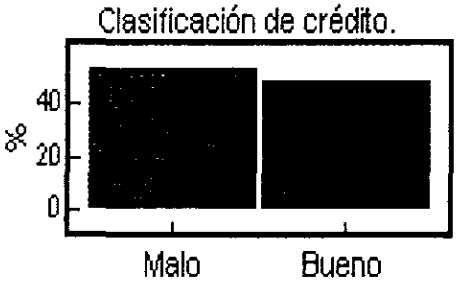

Figura 3.2: Distribución de crédito en el nodo raíz.

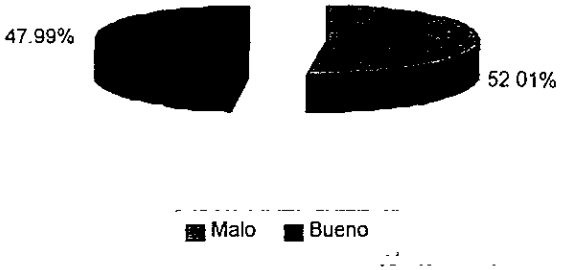

Figura 3.3: Clasificación de créditos del total de la muestra.

Nivel uno

Para cada una de las predictoras de este análisis, se debe encontrar la tabla reducida más significativa con respecto a la variable dependiente. El algoritmo muestra un método iterativo para lograr este propósito, el cuál está comprendido del paso 1 al paso 6. Este procedimiento se debe realizar con cada una de las predictoras.

Si por ejemplo, se toma la variable predictora *edad*, el primer paso del algoritmo consiste en realizar una tabla de contingencia de todas las categorías de esta variable predictora *edad* con todas las categorías de la variable dependiente (*tipo_crédito*) (Paso 1). Esta tabla se muestra en la Tabla 3.3.

	Clasificación de Crédito				
	Malo		*Bueno*		
Edad	*Casos*	*%*	*Casos*	*%*	Total
Joven	151	80 75%	36	19.25%	**187**
Maduro	16	20 25%	63	79.75%	**79**
Viejo	1	1.75%	56	98.25%	**57**
Total	**168**	**52.01%**	**155**	**47.99%**	**323**

Tabla 3.3: Tabla de contingencia de las variables *edad* y *tipo_crédito*.

Como se vio en el capítulo 1, Sección 1.3.2, referente a la prueba de independencia para tablas de $r \times c$, la estadística de prueba de la hipótesis de independencia es la χ^2 de Pearson. Para la Tabla 3 3, esta estadística se calcula a continuación

$$\begin{aligned} T &= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{ij}^2}{E_{ij}} - n \\ &= \sum_{i=1}^{3}\sum_{j=1}^{2}\frac{O_{ij}^2}{E_{ij}} - 323 \\ &= (234.426 + 14.442 + 6.230 + 104.695 + 0.034 + 114.649) - 323 \\ &= 151.4763. \end{aligned}$$

También se puede ocupar como estadístico de prueba el cociente de máxima verosimilitud

(ver Sección 1.3.2) y los cálculos entonces serían de la siguiente manera:

$$\begin{aligned} T &= 2\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \ln \frac{O_{ij}}{E_{ij}} \\ &= 2\sum_{i=1}^{3}\sum_{j=1}^{2} O_{ij} \ln \frac{O_{ij}}{E_{ij}} \\ &= 2 \times (66\,4188 - 32.8810 + -15.0907 + 31.9986 - 3.3894 + 40.1255) \\ &= 174.3636. \end{aligned}$$

La prueba de independencia indica que los grados de libertad de la χ^2, son el resultado de calcular la siguiente operación $(r-1)(c-1)$, donde r es el número de renglones en la tabla de contingencia y c el número de columnas. Esta expresión se reduce a $(3-1)(2-1) = 2$. con lo que, para el estadístico T que se distribuye como una χ^2 con 2 grados de libertad, se tiene que la probabilidad de dependencia o valor de probabilidad descriptivo p es de 1 28E−33. Para el estimador de verosimilitud, el valor p es de 1 37E−38. Ambas pruebas conducen a la conclusión de que la dependencia entre las variables es significativa al 5%. Una vez ejemplificados estos cálculos. de ahora en adelante se calculará sólo el estimador de verosimilitud para hacer más fácil el seguimiento de este ejemplo.

El siguiente paso (Paso 2), consiste en probar igualdad de distribuciones para cada par de categorías de la variable predictora que puedan ser unidas (de acuerdo al tipo de variable predictora, ver Sección 2.3.1), con respecto a las categorías de la variable dependiente. La variable *edad* tiene tres categorías. *joven*, *maduro* y *viejo*. *Edad* está definida como ordinal o monotónica, por tanto, sólo se pueden unir categorías contiguas, es decir. sólo las combinaciones *joven-maduro* o *maduro-viejo* son consideradas. En la Tabla 3.4 se muestra la tabla de contingencia del primer par de categorías (*joven-maduro*) con la variable dependiente *tipo_crédito*.

Edad	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Joven	151	80.75%	36	19 25%	**187**
Maduro	16	20.25%	63	79 75%	**79**
Total	**167**	**62.78%**	**99**	**37.22%**	**266**

Tabla 3.4: Tabla de contingencia de las categorías *joven y maduro* de la variable *edad.*

La χ^2 correspondiente a esta tabla está dada por

$$\begin{aligned} T &= 2\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \ln \frac{O_{ij}}{E_{ij}} \\ &= 2\sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij} \ln \frac{O_{ij}}{E_{ij}} \\ &= 2 \times (38.0027 - 23.7317 - 18.1017 + 48.01) \\ &= 88.35868. \end{aligned}$$

Hay un grado de libertad para la χ^2 asociada a esta tabla y el valor de probabilidad *descriptivo* p que se obtiene con estos datos es igual a 5.46E−21. Como este valor p es menor que α_1 (el nivel de significancia para unión de categorías), el cual es igual a 0.05, entonces se determina que estas categorías son significativas y por tanto el algoritmo indica que no se debe unir este par.

Este mismo procedimiento se debe aplicar para todo par permisible de categorías. En la Tabla 3.5 se muestra la tabla de contingencia del segundo par de categorías (*maduro-viejo*) con la variable dependiente.

Edad	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Maduro	16	20.25%	63	79.75%	**79**
Viejo	1	1.75%	56	98.25%	**57**
Total	**17**	**12.50%**	**119**	**87.50%**	**136**

Tabla 3.5: Tabla de contingencia de las categorías *maduro* y *viejo* de la variable *edad.*

El estimador correspondiente a esta tabla es

$$\begin{aligned} T &= 2\sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij} \ln \frac{O_{ij}}{E_{ij}} \\ &= 2 \times (7.7213 - 5.8452 - 1.9636 + 6.4866) \\ &= 12.7981. \end{aligned}$$

El valor de probabilidad descriptivo p para una χ^2 de 12.7981 con un grado de libertad es de 0.00035. Como el valor p es inferior al nivel de unión ($\alpha_1 = 0.05$), tampoco se unen estas categorías. Los dos pares permisibles de unir son significativos (Paso 3), por lo que, la tabla original no se reduce y el algoritmo indica ir directamente al paso 5, el cual señala que se debe unir la o las categorías que tengan menos observaciones que lo especificado por el tamaño mínimo de grupo después de dividir (en este ejemplo es igual a 1), con la categoría más similar de acuerdo a la menor estadística χ^2. En la Tabla 3.3 se observa que, la categoría *joven* tiene 187 casos, la categoría *maduro* 79 y la categoría *viejo* 57, sobrepasando todas el tamaño mínimo y, por esa razón, no se unen categorías bajo este paso. Para finalizar, se calcula el valor ajustado de Bonferroni,

$$\begin{aligned} \text{Valor } p \text{ ajustado} &= B_{monotónico} \times \text{valor } p \\ &= \begin{pmatrix} I-1 \\ J-1 \end{pmatrix} \times \text{valor } p \\ &= \begin{pmatrix} 3-1 \\ 3-1 \end{pmatrix} \times \text{valor } p \\ &= 1 \times 1.37\text{E-}38 \\ &= 1.37\text{E-}38. \end{aligned}$$

No hubo unión de categorías, por lo que, el multiplicador de Bonferroni es igual a 1 y el valor p ajustado de Bonferroni es igual al valor p de la tabla original, es decir, 1.37E−38.

Ahora se analizará la variable *ocupación*, siguiendo los mismos pasos del algoritmo de la Sección 2.5. Las categorías de esta variable son *gerencial, profesional, oficinista, técnico* y *obrero*. La tabla de contingencia de la variable *ocupación* con la variable dependiente aparece en la Tabla 3.6 (Paso 1).

	Clasificación de Crédito				
	Malo		*Bueno*		
Ocupación	*Casos*	*%*	*Casos*	*%*	Total
Gerencial	1	2.56%	38	97 44%	**39**
Profesional	56	35.44%	102	64.56%	**158**
Oficinista	40	85.11%	7	14.89%	**47**
Técnico	40	97 56%	1	2.44%	**41**
Obrero	31	81 58%	7	. 18.42%	**38**
Total	**168**	**52.01%**	**155**	**47.99%**	**323**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	147.2323	7.975E-31	significante
Pearson	123.5976	9.101E-26	significante

Tabla 3.6: Tabla de contingencia de las variables *ocupación* y *tipo_crédito*.

El estadístico T asociado con la tabla de contingencia de *ocupación* con *tipo de crédito* es de 147.2323 con un valor p de 7.98E−31. La variable *ocupación* está definida como libre por lo que, sus categorías pueden unirse unas con otras sin importar su orden.

Se empezará por estudiar la unión de las dos primeras categorías *(gerencial-profesional)* (Paso 2), Tabla 3.7.

	Clasificación de Crédito				
	Malo		*Bueno*		
Ocupación	*Casos*	*%*	*Casos*	*%*	Total
Gerencial	1	2.56%	38	97.44%	**39**
Profesional	56	35.44%	102	64.56%	**158**
Total	**57**	**28.93%**	**140**	**71.07%**	**197**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	22.2671	2.372E-06	significante
Pearson	16.4445	5.010E-05	significante

Tabla 3.7: Tabla de contingencia de las categorías *gerencial* y *profesional* de la variable *ocupación*.

El estimador χ^2 de cociente de verosimilitud es de

$$\begin{aligned} T &= 2\sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij} \ln \frac{O_{ij}}{E_{ij}} \\ &= 2 \times (-2.4234 + 11.9923 + 11.3629 - 9.7982) \\ &= 22.267099. \end{aligned}$$

El valor p asociado a este valor es de 2.37E−06 para una χ^2 con un grado de libertad, lo que lleva a decidir no unir estas categorías. Es necesario realizar todas las uniones de categorías posibles para encontrar las que no sean significativas y de ellas escoger la mejor (Paso 3).

Para ahorrar tiempo y espacio en esta evaluación de uniones de pares se tomarán casos considerando primero a la categoría *gerencial* y sus posibles uniones, siguiendo con la de profesional y sus posibles uniones pero, sin considerar aquéllas ya estudiadas en el caso de la gerencial, procediendo de la misma manera para las de oficinista, técnico y obrero. De esta manera, se analizarán todos los casos y no se duplicarán tablas de contingencia.

En el primer caso, entonces, se considera todas las uniones posibles de la categoría *gerencial* con las categorías restantes de la variable *ocupación*. En la Tabla 3.8 se muestran las tablas de contingencia, así como los valores del estadístico y su respectivo valor p. Se observa que ninguna de estas uniones son aprobadas por el algoritmo, porque todas son significantes, en consecuencia, no hay uniones.

	Clasificación de Crédito				
	Malo		*Bueno*		
Ocupación	*Casos*	*%*	*Casos*	*%*	Total
Gerencial	1	2.56%	38	97.44%	**39**
Oficinista	40	85.11%	7	14.89%	**47**
Total	**41**	**47.67%**	**45**	**52.33%**	**86**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	70.1732	5.432E-17	significante
Pearson	58 2127	2.353E-14	significante

	Clasificación de Crédito				
	Malo		*Bueno*		
Ocupación	*Casos*	*%*	*Casos*	*%*	Total
Gerencial	1	2.56%	38	97.44%	**39**
Técnico	40	97.56%	1	2.44%	**41**
Total	**41**	**51.25%**	**39**	**48.75%**	**80**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	92.1497	8.036E-22	significante
Pearson	72.1952	1.949E-17	significante

	Clasificación de Crédito				
	Malo		*Bueno*		
Ocupación	*Casos*	*%*	*Casos*	*%*	Total
Gerencial	1	2.56%	38	97.44%	**39**
Obrero	31	81 58%	7	18.42%	**38**
Total	**32**	**41.56%**	**45**	**58.44%**	**77**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	58.9315	1.633E-14	significante
Pearson	49.4759	2.008E-12	significante

Tabla 3.8: Tablas de contingencia de los pares entre la categoría *gerencial* y las restantes de la variable *ocupación.*

El siguiente caso tomará en consideración la categoría *profesional* y sus posibles uniones con las demás categorías que forman la variable *ocupación*, salvo la categoría *gerencial*, cuya unión ya se estudió en el caso anterior. Como se observa en la Tabla 3.9, todas las uniones son significativas por lo que, tampoco aquí se encontró una unión posible entre estas categorías.

Ocupación	Clasificación de Crédito				Total
	Malo		*Bueno*		
	Casos	%	*Casos*	%	
Profesional	56	35 44%	102	64.56%	**158**
Oficinista	40	85 11%	7	14.89%	**47**
Total	**96**	**46.83%**	**109**	**53.17%**	**205**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	38.3584	5.887E-10	significante
Pearson	35.8825	2.096E-09	significante

Ocupación	Clasificación de Crédito				Total
	Malo		*Bueno*		
	Casos	%	*Casos*	%	
Profesional	56	35.44%	102	64 56%	**158**
Técnico	40	97.56%	1	2 44%	**41**
Total	**96**	**48.24%**	**103**	**51.76%**	**199**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	60.7776	6.390E-15	significante
Pearson	50.3060	1.315E-12	significante

Ocupación	Clasificación de Crédito				Total
	Malo		*Bueno*		
	Casos	%	*Casos*	%	
Profesional	56	35 44%	102	64.56%	**158**
Obrero	31	81 58%	7	18 42%	**38**
Total	**87**	**44.39%**	**109**	**55.61%**	**196**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	27.4863	1.582E-07	significante
Pearson	26.4137	2.756E-07	significante

Tabla 3.9: Tablas de contingencia de los pares formados con *profesional* y las restantes categorías de la variable *ocupación.*

Los últimos casos corresponden a las categorías *oficinista* y *técnico.* Se descartan las uniones con las categorías *gerencial* y *profesional* porque ya se consideraron anteriormente. La Tabla 3.10 muestra los resultados.

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Oficinista	40	85.11%	7	14.89%	**47**
Técnico	40	97.56%	1	2.44%	**41**
Total	**80**	**90.91%**	**8**	**9.09%**	**88**

	χ^2	Valor p	Estado
Cociente Verosimilitud	4.6526	3.101E-02	significante
Pearson	4.1100	4.263E-02	significante

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Oficinista	40	85.11%	7	14 89%	**47**
Obrero	31	81.58%	7	18 42%	**38**
Total	**71**	**83.53%**	**14**	**16.47%**	**85**

	χ^2	Valor p	Estado
Cociente Verosimilitud	0.1892	6.636E-01	no significante
Pearson	0.1900	6.629E-01	no significante

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Técnico	40	97 56%	1	2.44%	**41**
Obrero	31	81.58%	7	18.42%	**38**
Total	**71**	**89.87%**	**8**	**10.13%**	**79**

	χ^2	Valor p	Estado
Cociente Verosimilitud	6.0920	1.358E-02	significante
Pearson	5.5349	1 864E-02	significante

Tabla 3.10: Tablas de contingencia de los pares formados con las categorías *oficinista y técnico* de la variable *ocupación*.

En esta última tabla, se observa que hay un par que no es significante (*oficinista-obrero*), ya que el valor p resultante es de .663, el cual es superior al nivel de significancia para unión (α_1) que se definió como 0.05 al inicio de este ejemplo y dado que. es el único par no significante encontrado, se procede a su unión La tabla reducida se presenta en la Tabla 3.11:

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Gerencial	1	2.56%	38	97.44%	**39**
Profesional	56	35 44%	102	64.56%	**158**
Ofic - Obre	71	83 53%	14	16.47%	**85**
Técnico	40	97.56%	1	2.44%	**41**
Total	**168**	**52.01%**	**155**	**47.99%**	**323**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	147.0432	1.144E-31	significante
Pearson	123 4929	1.365E-26	significante

Tabla 3 11: Tabla de contingencia reducida de la variable *ocupación* con la variable *tipo_crédito.*

Después de unir un par de categorías, el algoritmo procede al paso 4 que, sólo se aplica si existe alguna categoría unida formada con 3 o más categorías originales. Este paso prueba si algún elemento de una categoría unida debería estar separado. Sin embargo, en este caso, este paso no se aplica debido a que la categoría conjunta tan sólo tiene dos categorías

El siguiente paso consiste en verificar si es posible unir más categorías para obtener una tabla más reducida. Considerando a la categoría unida o conjunta como una sola, se vuelven a ejecutar los pasos del algoritmo (del 1 al 5) que permitieron unir el par *oficinista-obrero.* La Tabla 3.12 muestra las combinaciones que restan a estudiar, considerando a la categoría *oficinista-obrero* como una sola.

Como ninguna categoría resultó no significativa, ya no hay más uniones que realizar, por lo que, se continúa con el paso 5 del algoritmo. Para este paso se observa que todas las categorías finales sobrepasan el tamaño mínimo de subgrupo después de dividir que es de 1, por lo tanto, no se unen más categorías a causa de este parámetro. Por último, se calcula el valor ajustado p de Bonferroni, nótese que *ocupación* es una variable libre, *por tanto, el valor p ajustado de Bonferroni (ver Sección 2.6), está dado* de multiplicar el multiplicador de Bonferroni por el valor p de la tabla reducida final:

Ocupación	Clasificación de Crédito: Malo Casos	Malo %	Bueno Casos	Bueno %	Total
Gerencial	1	2.56%	38	97.44%	39
Ofic - Obre	71	83.53%	14	16 47%	85
Total	72	58.06%	52	41.94%	124

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	83 3027	7.040E-20	significante
Pearson	71 9726	2 182E-17	significante

Ocupación	Clasificación de Crédito: Malo Casos	Malo %	Bueno Casos	Bueno %	Total
Profesional	56	35.44%	102	64 56%	158
Ofic - Obre	71	83.53%	14	16 47%	85
Total	127	52.26%	116	47.74%	243

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	54.8687	1.289E-13	significante
Pearson	51.2230	8 245E-13	significante

Ocupación	Clasificación de Crédito: Malo Casos	Malo %	Bueno Casos	Bueno %	Total
Ofic - Obre	71	83.53%	14	16 47%	85
Técnico	40	97 56%	1	2 44%	41
Total	111	88.10%	15	11.90%	126

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	6.5267	1 063E-02	significante
Pearson	5 1924	2 269E-02	significante

Tabla 3.12: Tablas de contingencia de pares de categorías de la variable *ocupación*.

$$\text{Valor } p \text{ ajustado} = B_{libre} \times \text{valor } p = \sum_{i=0}^{J-1} (-1)^i \frac{(J-i)^I}{i!(J-i)!} \times \text{valor } p.$$

En este caso $I = 5, J = 4$ y valor p = 1.1444E−31, entonces,

$$\begin{aligned} \text{Valor } p \text{ ajustado} &= \sum_{i=0}^{3} (-1)^i \frac{(4-i)^5}{i!(4-i)!} \times 1.1444\text{E-}31 \\ &= \left(\frac{4^5}{4!} - \frac{3^5}{3!} + \frac{2^5}{2!2!} - \frac{1}{3!}\right) \times 1.1444\text{E-}31 \\ &= 1.1444\text{E-}30. \end{aligned}$$

Falta considerar las predictoras *frec_pago* y *amex*, bajo el mismo procedimiento al aplicado a las variables *edad* y *ocupación*. Para la variable *frec_pago*, la Tabla 3.13 muestra la tabla original a reducir en menos categorías (Paso 1).

	Clasificación de Crédito				
	Malo		*Bueno*		
Frec_Pago	*Casos*	*%*	*Casos*	*%*	Total
Semanal	143	86 67%	22	13.33%	**165**
Mensual	25	15.82%	133	84.18%	**158**
Total	**168**	**52.01%**	**155**	**47.99%**	**323**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	179 6638	5 739E-41	significante
Pearson	162.2958	3.565E-37	significante

Tabla 3 13: Tabla de contingencia de la variable *frec_pago* y la variable *tipo_crédito*.

Como se puede observar en la Tabla 3.13 no se unen las dos categorías de la variable *frec_pago* ya que el par es significante (Pasos 2 y 3). El número de casos de cada categoría de esta variable es mayor al tamaño mínimo de casos por lo que, se termina el estudio de esta variable (Paso 5). El multiplicador de Bonferroni al no haber unión de categorías es igual a 1 por lo que, el valor p ajustado queda igual al valor p (5.739E-41).

El siguiente paso del algoritmo es analizar la variable *amex*. Para este propósito se tomará como base la tabla de contingencia de *amex* contra la variable dependiente *tipo_crédito*.

	Clasificación de Crédito				
	Malo		*Bueno*		
Amex	*Casos*	*%*	*Casos*	*%*	Total
No	89	53 29%	78	46.71%	**167**
Si	79	50.64%	77	49.36%	**156**
Total	**168**	**52.01%**	**155**	**47.99%**	**323**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	0 2274	6 335E-01	no significante
Pearson	0.2273	6.335E-01	no significante

Tabla 3.14: Tabla de contingencia de la variable *amex* y la variable *tipo_crédito*.

Se observa en la Tabla 3.14 que, la unión de estas categorías es recomendable, por lo que, se unen, dando como resultado una tabla reducida como la que se muestra en la Tabla 3.15. El valor p ajustado de Bonferroni para esta variable es 1 puesto que el número de categorías se redujo a una.

	Clasificación de Crédito				
	Malo		*Bueno*		
Amex	*Casos*	*%*	*Casos*	*%*	Total
No - Si	168	52.01%	155	47.99%	**323**
Total	**168**	**52.01%**	**155**	**47.99%**	**323**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	0.0000	1.000E+00	no significante
Pearson	0.0000	1.000E+00	no significante

Tabla 3.15: Tabla de contingencia reducida de la variable *amex* y la variable *tipo_crédito*.

El siguiente paso (Paso 6), una vez que ya se unieron las categorías posibles de todas las variables, es comparar las predictoras y elegir la mejor; para este efecto, sólo se tomarán en cuenta aquellos que tengan un valor ajustado p estadísticamente significante, es decir, menor a 0 05, y de ellos se seleccionará aquel que tenga un valor p estadísticamente más pequeño, considerándolo como la mejor variable predictora.

La Tabla 3.16 muestra un resumen de la información obtenida hasta ahora por el algoritmo. Están listadas las variables involucradas en este análisis con la siguiente información: el número inicial y final de categorías, los valores de la χ^2 de la tabla inicial y de la reducida final, los grados de libertad para la tabla reducida, los valores p de la tabla original y de la tabla reducida, así como el valor p ajustado de Bonferroni para la tabla reducida final.

Variables	Cat.	$\chi 2$ Inicial	$\chi 2$ Reducida	g.l.	*p* inicial	*p* reducida	*p* ajustado
Frec_Pago	2	179.6638	179.6638	1	5.739E-41	5.739E-41	5.739E-41
Edad	3	174.3636	174.3636	2	1.372E-38	1 372E-38	1.372E-38
Ocupación	5 → 4	147 2323	147 0432	3	7 975E-31	1 144E-31	1 144E-30
Amex	2 → 1	0.2274	0 0000	0	0.6335	1.0000	1 0000

Tabla 3 16: Tabla resumen de las variables predictoras para la clasificación de créditos.

En primer término, se observa que la variable *amex* posee un valor p ajustado superior a 0.05, en consecuencia no es considerada como significativa y, por lo tanto, queda fuera de la elección. La variable *frec_pago* tiene un valor p ajustado inferior a *edad* y *ocupación* por lo que se escoge a *frec_pago* como mejor predictora. Por lo tanto, se divide el nodo raíz en los casos correspondientes a las categorías de la tabla final o reducida de la variable *frec_pago*, la cual no tuvo uniones y permaneció con sus dos categorías iniciales: *semanal* y *mensual*. Consecuentemente, el árbol crece un nivel con dos nodos más (Figura 3.4), que exceden el número de casos mínimo antes de dividir, que es igual a 50 (Paso 7).

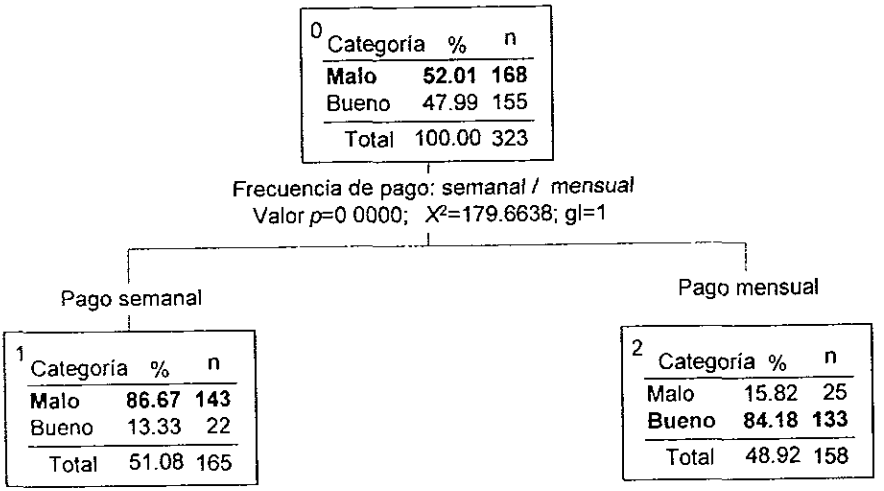

Figura 3.4: Árbol de un nivel para la clasificación de créditos.

El primer nivel del árbol se ha terminado, la mayoría de los casos en el grupo con pagos semanales tienen riesgos de crédito muy altos, mientras que aquellos en el grupo con pago mensual son más probables para tener riesgos de crédito bajos. Basados en esta variable, se pudo mejorar ampliamente la identificación de los riesgos de crédito buenos de los malos. Así que parece que un poco más de información y análisis ayudarán para este estudio.

Nivel dos

El mismo procedimiento que se ha empleado con el nodo raíz, se aplicará a todos los nodos que se vayan formando hasta que lo suspenda alguna regla de paro. Nuevamente se seguirán detalladamente los pasos del algoritmo, a fin de completar la explicación y ejecución del mismo y aclarar dudas que todavía persistan en el lector. Se tomará el nodo formado por el subgrupo correspondiente a aquellos a quiénes se les paga semanalmente. Este subgrupo está formado con 165 casos. En el diagrama del árbol (Figura 3.4) se identifica con el número "1".

Como en el nodo anterior, se debe construir una tabla de contingencia por cada variable predictora con el fin de reducirla en dimensión. La primera variable será *edad* y la tabla de contingencia de ésta con la variable dependiente *tipo_crédito* se muestra en la Tabla 3.17 (Paso1).

	Clasificación de Crédito				
	Malo		*Bueno*		
Edad	*Casos*	*%*	*Casos*	*%*	Total
Joven	127	92 03%	11	7.97%	**138**
Maduro	16	80.00%	4	20 00%	**20**
Viejo	0	0.00%	7	100.00%	**7**
Total	**143**	**86.67%**	**22**	**13.33%**	**165**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	32 8217	7.462E-08	significante
Pearson	49.7032	1.611E-11	significante

Tabla 3 17: Tabla de contingencia de *edad* y la variable *tipo_crédito* en el subgrupo de *pago semanal*.

Se probará a continuación la significancia para los pares posibles de esta variable (Paso 2). En la Tabla 3.18 se observa que el único par no significante es *joven-maduro*.

Edad	Clasificación de Crédito				Total
	Malo		*Bueno*		
	Casos	%	*Casos*	%	
Joven	127	92 03%	11	7 97%	**138**
Maduro	16	80.00%	4	20.00%	**20**
Total	**143**	**90.51%**	**15**	**9.49%**	**158**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	2 4041	1.210E-01	no significante
Pearson	2.9417	8 632E-02	no significante

Edad	Clasificación de Crédito				Total
	Malo		*Bueno*		
	Casos	%	*Casos*	%	
Maduro	16	80.00%	4	20 00%	**20**
Viejo	0	0.00%	7	100 00%	**7**
Total	**16**	**59.26%**	**11**	**40.74%**	**27**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	16.4826	4.910E-05	significante
Pearson	13 7455	2.093E-04	significante

Tabla 3.18: Tablas de contingencia de los pares de categorías de la variable *edad.*

El valor p para el par *joven-maduro* es de 0.1210, por lo que no es significante, en consecuencia las categorías *joven* y *maduro* se unen y se obtiene un tabla reducida (Paso 3), como la que se muestra en la Tabla 3.19.

Edad	Clasificación de Crédito				Total
	Malo		*Bueno*		
	Casos	%	*Casos*	%	
Jov - Mad	143	90 51%	15	9.49%	**158**
Viejo	0	0.00%	7	100.00%	**7**
Total	**143**	**86.67%**	**22**	**13.33%**	**165**

	$\chi 2$	Valor p	Estado	Bonferroni
Cociente Verosimilitud	30 4176	3.484E-08	significante	6 967E-08
Pearson	47 5158	5 456E-12	significante	1.091E-11

Tabla 3.19: Tabla de contingencia reducida de la variable *edad* con la variable *tipo_crédito.*

La variable *edad* ya no tiene más uniones posibles y se observa que las categorías

formadas no tienen más de tres categorías conjuntas para que sea posible su división (Paso 4), de igual forma ninguna categoría tiene menos observaciones que el tamaño mínimo después de dividir (Paso 5). Lo que resta por hacer para esta variable, es calcular el valor ajustado de Bonferroni para la tabla reducida:

$$\begin{aligned}
\text{Valor } p \text{ ajustado} &= B_{monotonico} \times \text{valor } p \\
&= \begin{pmatrix} I-1 \\ I'-1 \end{pmatrix} \times \text{valor } p \\
&= \begin{pmatrix} 3-1 \\ 2-1 \end{pmatrix} \times \text{valor } p \\
&= 2 \times 3.48\text{E}-08 \\
&= 6.97\text{E}-08.
\end{aligned}$$

La siguiente variable a estudiar será *ocupación*. Su tabla de contingencia con relación al subgrupo de pagos semanales se despliega en la Tabla 3.20, así como, su estadístico y el valor p (Paso 1).

	Clasificación de Crédito				
	Malo		*Bueno*		
Ocupación	*Casos*	*%*	*Casos*	*%*	Total
Gerencial	1	33.33%	2	66 67%	**3**
Profesional	31	73 81%	11	26 19%	**42**
Oficinista	40	97 56%	1	2.44%	**41**
Técnico	40	97.56%	1	2 44%	**41**
Obrero	31	81.58%	7	18.42%	**38**
Total	**143**	**86.67%**	**22**	**13.33%**	**165**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	22.3484	1 708E-04	significante
Pearson	22.6662	1.476E-04	significante

Tabla 3.20: Tabla de contingencia de la variable *ocupación* y la variable *tipo_crédito*.

Con el fin de encontrar el menor par no significativo, en la Tabla 3.21 se muestran los pares con la categoría *gerencial*, en la Tabla 3.22 aquellos con *profesional* y, finalmente, la Tabla 3.23 presenta los pares con las categorías *oficinista* y *técnico* (Paso 2).

Ocupación	Clasificación de Crédito				
	Malo		*Bueno*		
	Casos	%	*Casos*	%	Total
Gerencial	1	33.33%	2	66 67%	3
Profesional	31	73.81%	11	26.19%	**42**
Total	**32**	**71.11%**	**13**	**28.89%**	**45**

	χ^2	Valor p	Estado
Cociente Verosimilitud	1 9814	1 592E-01	no significante
Pearson	2.2330	1.351E-01	no significante

Ocupación	Clasificación de Crédito				
	Malo		*Bueno*		
	Casos	%	*Casos*	%	Total
Gerencial	1	33 33%	2	66.67%	3
Oficinista	40	97.56%	1	2.44%	**41**
Total	**41**	**93.18%**	**3**	**6.82%**	**44**

	χ^2	Valor p	Estado
Cociente Verosimilitud	8 6825	3.213E-03	significante
Pearson	18.1508	2.041E-05	significante

Ocupación	Clasificación de Crédito				
	Malo		*Bueno*		
	Casos	%	*Casos*	%	Total
Gerencial	1	33.33%	2	66 67%	3
Técnico	40	97.56%	1	2 44%	**41**
Total	**41**	**93.18%**	**3**	**6.82%**	**44**

	χ^2	Valor p	Estado
Cociente Verosimilitud	8.6825	3.213E-03	significante
Pearson	18 1508	2.041E-05	significante

Ocupación	Clasificación de Crédito				
	Malo		*Bueno*		
	Casos	%	*Casos*	%	Total
Gerencial	1	33.33%	2	66 67%	3
Obrero	31	81.58%	7	18 42%	**38**
Total	**32**	**78.05%**	**9**	**21.95%**	**41**

	χ^2	Valor p	Estado
Cociente Verosimilitud	3.0301	8 173E-02	no significante
Pearson	3.7776	5.194E-02	no significante

Tabla 3.21: Tablas de los pares de la categoría *gerencial* y las restantes categorías de la variable *ocupación* contra la variable *tipo_crédito*.

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Profesional	31	73.81%	11	26.19%	**42**
Oficinista	40	97.56%	1	2 44%	**41**
Total	**71**	**85.54%**	**12**	**14.46%**	**83**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	10 8833	9.703E-04	significante
Pearson	9.4635	2 096E-03	significante

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Profesional	31	73.81%	11	26.19%	**42**
Técnico	40	97 56%	1	2.44%	**41**
Total	**71**	**85.54%**	**12**	**14.46%**	**83**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	10 8833	9 703E-04	significante
Pearson	9.4635	2 096E-03	significante

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Profesional	31	73.81%	11	26.19%	**42**
Obrero	31	81 58%	7	18.42%	**38**
Total	**62**	**77.50%**	**18**	**22.50%**	**80**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	0.6963	4.040E-01	no significante
Pearson	0.6906	4 060E-01	no significante

Tabla 3.22: Tablas de los pares de la categoría *profesional* y las restantes categorías de la variable *ocupación* contra la variable *tipo_crédito*.

Ocupación	Clasificación de Crédito				Total
	Malo		*Bueno*		
	Casos	*%*	*Casos*	*%*	
Oficinista	40	97.56%	1	2 44%	**41**
Técnico	40	97.56%	1	2 44%	**41**
Total	**80**	**97.56%**	**2**	**2.44%**	**82**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	0.0000	1 000E+00	no significante
Pearson	0 0000	1 000E+00	no significante

Ocupación	Clasificación de Crédito				Total
	Malo		*Bueno*		
	Casos	*%*	*Casos*	*%*	
Oficinista	40	97.56%	1	2 44%	**41**
Obrero	31	81.58%	7	18.42%	**38**
Total	**71**	**89.87%**	**8**	**10.13%**	**79**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	6 0920	1.358E-02	significante
Pearson	5.5349	1 864E-02	significante

Ocupación	Clasificación de Crédito				Total
	Malo		*Bueno*		
	Casos	*%*	*Casos*	*%*	
Técnico	40	97.56%	1	2.44%	**41**
Obrero	31	81.58%	7	18 42%	**38**
Total	**71**	**89.87%**	**8**	**10.13%**	**79**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	6.0920	1.358E-02	significante
Pearson	5 5349	1.864E-02	significante

Tabla 3.23: Tablas con las categorías *oficinista* y *técnico*, y la variable *tipo_crédito*.

De estas tablas de contingencia, nótese que hay 4 pares no significantes:

Pares de categorías	Valor p
gerencial-profesional	1.59E-01
gerencial-obrero	8.17E-02
profesional-obrero	4.04E-01
oficinista-técnico	1

Se escoge el par con el menor valor p (Paso 3), que es *gerencial-obrero*, y la tabla resultante se muestra en la Tabla 3.24.

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Geren-Obrer	32	78.05%	9	21 95%	**41**
Profesional	31	73.81%	11	26 19%	**42**
Oficinista	40	97.56%	1	2 44%	**41**
Técnico	40	97.56%	1	2.44%	**41**
Total	**143**	**86.67%**	**22**	**13.33%**	**165**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	19 3184	2.349E-04	significante
Pearson	17 0655	6 852E-04	significante

Tabla 3.24: Tabla de contingencia reducida de *ocupación* y la *tipo_crédito*.

Nuevamente, se realizan todas las combinaciones de las categorías nuevas. Estas se muestran en las Tablas 3.25 y 3.26 (Paso 2).

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Geren-Obrer	32	78 05%	9	21.95%	**41**
Profesional	31	73 81%	11	26 19%	**42**
Total	**63**	**75.90%**	**20**	**24.10%**	**83**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	0 2042	6 514E-01	no significante
Pearson	0 2039	6 516E-01	no significante

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Geren-Obrer	32	78 05%	9	21.95%	**41**
Oficinista	40	97.56%	1	2 44%	**41**
Total	**72**	**87.80%**	**10**	**12.20%**	**82**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	8 2520	4.071E-03	significante
Pearson	7 2889	6.938E-03	significante

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Geren-Obrer	32	78 05%	9	21 95%	**41**
Técnico	40	97 56%	1	2 44%	**41**
Total	**72**	**87.80%**	**10**	**12.20%**	**82**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	8 2520	4.071E-03	significante
Pearson	7.2889	6.938E-03	significante

Tabla 3.25: Tablas de los pares de la categoría conjunta *gerencial-obrero* y *tipo_crédito*.

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Profesional	31	73 81%	11	26.19%	**42**
Oficinista	40	97 56%	1	2 44%	**41**
Total	71	**85.54%**	12	**14.46%**	**83**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	10.8833	9.703E-04	significante
Pearson	9.4635	2 096E-03	significante

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Profesional	31	73.81%	11	26 19%	**42**
Técnico	40	97.56%	1	2 44%	**41**
Total	71	**85.54%**	12	**14.46%**	**83**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	10 8833	9 703E-04	significante
Pearson	9.4635	2 096E-03	significante

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Oficinista	40	97.56%	1	2 44%	**41**
Técnico	40	97.56%	1	2 44%	**41**
Total	80	**97.56%**	2	**2.44%**	**82**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	0.0000	1 000E+00	no significante
Pearson	0.0000	1 000E+00	no significante

Tabla 3.26. Tablas de los pares de las categorías *profesional, oficinista* y *técnico.*

Se observa ahora que, el par *(gerente-obrero)-profesional* es el que posee un valor p más pequeño por abajo de *oficinista-técnico* que es el otro par no significativo. De esta manera, la tabla reducida queda de la forma presentada en la Tabla 3.27 (Paso 3).

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Ger - Pro - Obr	63	75 90%	20	24.10%	**83**
Oficinista	40	97.56%	1	2 44%	**41**
Técnico	40	97.56%	1	2 44%	**41**
Total	**143**	**86.67%**	**22**	**13.33%**	**165**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	19 1142	7.070E-05	significante
Pearson	16 7428	2 314E-04	significante

Tabla 3.27: Tabla reducida de *ocupación* y la variable *tipo_crédito.*

El siguiente paso consiste en verificar que todas las categorías conjuntas deban estar unidas (Paso 4). Para este propósito, algoritmo señala que se deben encontrar todas las posibles (de acuerdo al tipo de predictora) divisiones binarias de las categorías predictoras que componen a la conjunta y obtener el estadístico χ^2 que compara a cada una con las demás que componen a la conjunta. Si se encuentra a la χ^2 significante, entonces se divide la categoría.

	Clasificación de Crédito				
	Malo		*Bueno*		
Ocupación	*Casos*	*%*	*Casos*	*%*	Total
Gerente	1	33.33%	2	66 67%	**3**
Pro-Obr	62	77.50%	18	22.50%	**80**
Total	**63**	**75.90%**	**20**	**24.10%**	**83**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	2 5380	1.111E-01	no significante
Pearson	3 0840	7.907E-02	no significante

	Clasificación de Crédito				
	Malo		*Bueno*		
Ocupación	*Casos*	*%*	*Casos*	*%*	Total
Ger - Pro	32	71.11%	13	28.89%	**45**
Obrero	31	81 58%	7	18.42%	**38**
Total	**63**	**75.90%**	**20**	**24.10%**	**83**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	1.2528	2.630E-01	no significante
Pearson	1.2343	2.666E-01	no significante

	Clasificación de Crédito				
	Malo		*Bueno*		
Ocupación	*Casos*	*%*	*Casos*	*%*	Total
Ger - Obr	32	78 05%	9	21.95%	**41**
Profesional	31	73 81%	11	26.19%	**42**
Total	**63**	**75.90%**	**20**	**24.10%**	**83**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	0.2042	6.514E-01	no significante
Pearson	0.2039	6.516E-01	no significante

Tabla 3.28: Tabla de divisiones binarias para la categoría conjunta *ger-pro-obr* de la variable *ocupación*.

En el presente análisis, como se puede observar en la Tabla 3.28, ninguna división llega a ser significativa por lo que, la categoría conjunta permanece sin cambios, ya que tampoco se sobrepasa el nivel mínimo después de dividir.

Se repite el procedimiento para las categorías nuevas y se obtienen los datos que se muestran en la Tabla 3.29. Existe un par no significante formado por las categorías *oficinista y técnico*; dado esto, se procede a unirlas y la tabla reducida se presenta en la Tabla 3.30 que, también, muestra su respectivo valor p ajustado de Bonferroni.

	Clasificación de Crédito				
	Malo		*Bueno*		
Ocupación	*Casos*	*%*	*Casos*	*%*	Total
Ger - Pro - Obr	63	75.90%	20	24.10%	**83**
Oficinista	40	97.56%	1	2 44%	**41**
Total	**103**	**83.06%**	**21**	**16.94%**	**124**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	11.7399	6 117E-04	significante
Pearson	9.1504	2 487E-03	significante

	Clasificación de Crédito				
	Malo		*Bueno*		
Ocupación	*Casos*	*%*	*Casos*	*%*	Total
Ger - Pro - Obr	63	75.90%	20	24.10%	**83**
Técnico	40	97.56%	1	2.44%	**41**
Total	**103**	**83.06%**	**21**	**16.94%**	**124**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	11.7399	6.117E-04	significante
Pearson	9.1504	2.487E-03	significante

	Clasificación de Crédito				
	Malo		*Bueno*		
Ocupación	*Casos*	*%*	*Casos*	*%*	Total
Oficinista	40	97.56%	1	2 44%	**41**
Técnico	40	97.56%	1	2.44%	**41**
Total	**80**	**97.56%**	**2**	**2.44%**	**82**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	0.0000	1.000E+00	no significante
Pearson	0.0000	1.000E+00	no significante

Tabla 3.29: Tablas de los pares de categorías de la tabla reducida de *ocupación*.

Ocupación	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
Ger - Pro - Obr	63	75.90%	20	24.10%	**83**
Ofic - Tec	80	97.56%	2	2.44%	**82**
Total	**143**	**86.67%**	**22**	**13.33%**	**165**

	$\chi 2$	Valor p	Estado	Bonferroni
Cociente Verosimilitud	19 1142	1 231E-05	significante	1 847E-04
Pearson	16 7428	4.280E-05	significante	6 421E-04

Tabla 3.30: Tabla reducida final para la variable *ocupación* y la variable dependiente.

Se detiene el proceso de unión para esta variable porque, como se observa en la Tabla 3 30, ya no hay pares no significantes.

La variable siguiente y última es *amex*, la tabla de contingencia (Tabla 3.31) muestra que sí deberían unirse sus categorías, resultando en una tabla con una sola (Tabla 3 32).

Amex	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
No	79	85.87%	13	14.13%	**92**
Si	64	87.67%	9	12.33%	**73**
Total	**143**	**86.67%**	**22**	**13.33%**	**165**

	$\chi 2$	Valor p	Estado
Cociente Verosimilitud	0.1149	7.346E-01	no significante
Pearson	0.1143	7.353E-01	no significante

Tabla 3.31: Tabla de contingencia de *amex* y *tipo_crédito*.

Amex	Clasificación de Crédito				Total
	Malo		Bueno		
	Casos	%	Casos	%	
No - Si	143	86.67%	22	13.33%	**165**
Total	**143**	**86.67%**	**22**	**13.33%**	**165**

	$\chi 2$	Valor p	Estado	Bonferroni
Cociente Verosimilitud	0.0000	1.000E+00	no significante	1 000E+00
Pearson	0.0000	1.000E+00	no significante	1 000E+00

Tabla 3.32: Tabla reducida de la variable *amex* y la variable *tipo_crédito*.

Finalmente. con toda la información obtenida y que, se muestra en la tabla resumen de la Tabla 3.33, se decidirá cuál es la mejor predictora.

Variables	Cat.	χ^2 Inicial	χ^2 Reducida	g.l.	p inicial	p reducida	p ajustado
Edad	3 → 2	32 8217	30 4176	1	7 462E-08	3.484E-08	6 967E-08
Ocupación	5 → 2	22.3484	19 1142	1	1.708E-04	1.231E-05	1 847E-04
Amex	2 → 1	0.1149	0 0000	0	0 7346	1.0000	1.0000

Tabla 3.33: Tabla resumen para el segundo nivel de la rama izquierda.

Se selecciona la variable *edad* y el subgrupo de pago semanal se divide en dos grupos: los jóvenes y maduros forman una categoría y aquellos con más de 35 años o viejos forman la segunda El árbol crece un nivel en este nodo, véase la Figura 3.5.

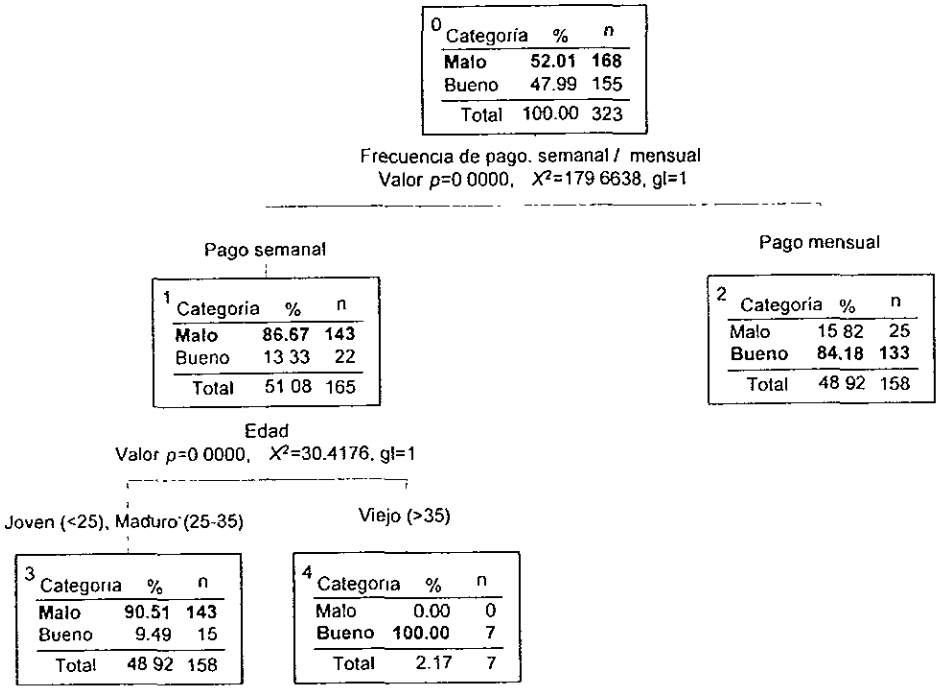

Figura 3.5: Árbol expandido en la rama de pago semanal.

Examinando esta rama izquierda, se observa que, para aquellos solicitantes de crédito a quienes se les paga semanalmente, los solicitantes viejos (>35 años) tienden a tener buen crédito (todos los casos en este ejemplo). Por el otro lado, los jóvenes (<25) y maduros (25-35) fueron más probables a tener un crédito pobre.

Una vez explicado el procedimiento del algoritmo, el análisis para los siguientes nodos se realizará con base en las tablas resumen de información sin la ejecución detallada del algoritmo, con el propósito de agilizar el término de este ejemplo. El siguiente nodo en el nivel uno es el dedicado a aquellos cuyo salario es mensual. La Tabla 3.34 presenta los datos finales, una vez que ya se redujeron las tablas.

Variables	Cat.	$\chi 2$ Inicial	$\chi 2$ Reducida	g.l.	p inicial	p reducida	p ajustado
Edad	3 → 2	60.2914	58.7219	1	8.089E-14	1 816E-14	3.632E-14
Ocupación	5 → 2	17.0907	17 0907	1	1 856E-03	3.564E-05	5.346E-04
Amex	2 → 1	0.6692	0 0000	0	0.4133	1.0000	1.0000

Tabla 3.34: Tabla resumen en la rama de pago mensual, primer nivel.

También en esta rama, la mejor predictora resulta ser *edad*, y, por tanto, este nodo se divide de acuerdo a esta variable que, como se observa (Figura 3.6) pasó de 3 a 2 categorías (*jóvenes y maduros-viejos*), sin embargo, éstas son diferentes que las del nodo uno.

Esta rama derecha parece un poco diferente a la rama izquierda. Se mantiene evidente que los solicitantes más viejos son más valiosos en crédito que los jóvenes. Sin embargo, para los solicitantes a quienes se les paga mensualmente, las personas maduras están agrupadas con los viejos porque representan menor riesgo de crédito que los jóvenes. Nuevamente, se observa que, la mayoría de estos solicitantes tienen buen crédito. En el grupo joven, los casos están divididos similarmente, la mitad de los solicitantes en este grupo tienen créditos buenos y la mitad tiene malos créditos.

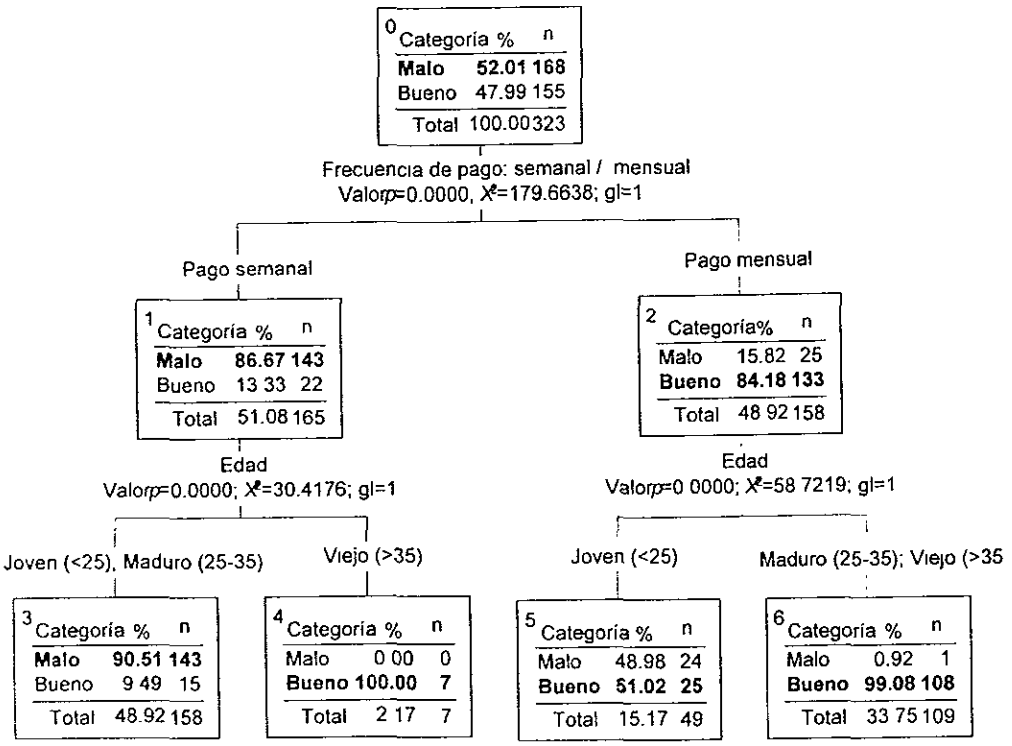

Figura 3.6: Árbol con dos niveles.

Nivel tres

Añadiendo un nivel más, quizá se pueda identificar la diferencia entre los buenos y malos créditos en este subgrupo. De acuerdo con los datos de la Tabla 3.35, el subgrupo de jóvenes con pagos mensuales se divide en dos grupos basados en la variable *ocupación* (Figura 3.7): en una rama están *gerencial-oficinista* y en la otra está *profesional* (no hubo trabajadores técnicos ni obreros en este nodo). En el nodo *gerencial-oficinista*, todos los casos tienen crédito bueno, mientras que el nodo *profesional* tiene tanto riesgos buenos como malos (41.46% y 58.54%, respectivamente).

Variables	Cat.	χ^2 Inicial	χ^2 Reducida	g.l.	*p* inicial	*p* reducida	*p* ajustado
Ocupación	5 → 2	12.2709	12.2709	1	0.0154	4.601E-04	0.0014
Amex	2 → 1	0.0272	0.0000	0	0.8690	1 0000	1 0000

Tabla 3.35: Tabla resumen de las variables predictoras para el subgrupo de jóvenes con pagos mensuales.

En el lado izquierdo del árbol, se dividió el subgrupo de jóvenes y maduros con pagos semanales en dos grupos, basados en la variable *ocupación* (Tabla 3.36). Sin embargo, como se observa en la Figura 3.7, esta división es innecesaria puesto que, las dos divisiones obtenidas se clasifican como malos créditos y no proporcionan información relevante acerca del perfil de los casos que componen estos nodos de la segmentación.

Variables	Cat.	χ^2 Inicial	χ^2 Reducida	g.l.	*p* inicial	*p* reducida	*p* ajustado
Ocupación	5 → 2	18 7980	15 5342	1	8.611E-04	8.103E-05	1 215E-03
Amex	2 → 1	0 7369	0 0000	0	0 3907	1 0000	1.0000

Tabla 3 36: Tabla resumen de las variables predictoras para nodo 3.

Dos nodos no se dividieron más, el subgrupo de solicitantes con más de 35 años con pagos semanales y el subgrupo de solicitantes con 25 años o más y pagos mensuales. El primer subgrupo no alcanzaba el tamaño mínimo de subgrupo antes de dividir que está definido como 25, por lo que, el algoritmo no analizó el nodo. Y, para el segundo subgrupo, la decisión de no dividir más se debió a que no hubo predictoras significantes, como se observa en las Tabla 3.37.

Variables	Cat.	χ^2 Inicial	χ^2 Reducida	g.l.	p	p reducida	p ajustado
Ocupación	5 → 1	0 7519	0 0000	0	0.9448	1 0000	1 0000
Amex	2 → 1	1.3771	0.0000	0	0.2406	1 0000	1 0000

Tabla 3.37: Tabla resumen de las variables predictoras para nodo 6.

Con este paso se termina el análisis, pues se ha alcanzado el nivel tres impuesto como regla de paro. El árbol ofrece suficiente información de cómo está constituida la población.

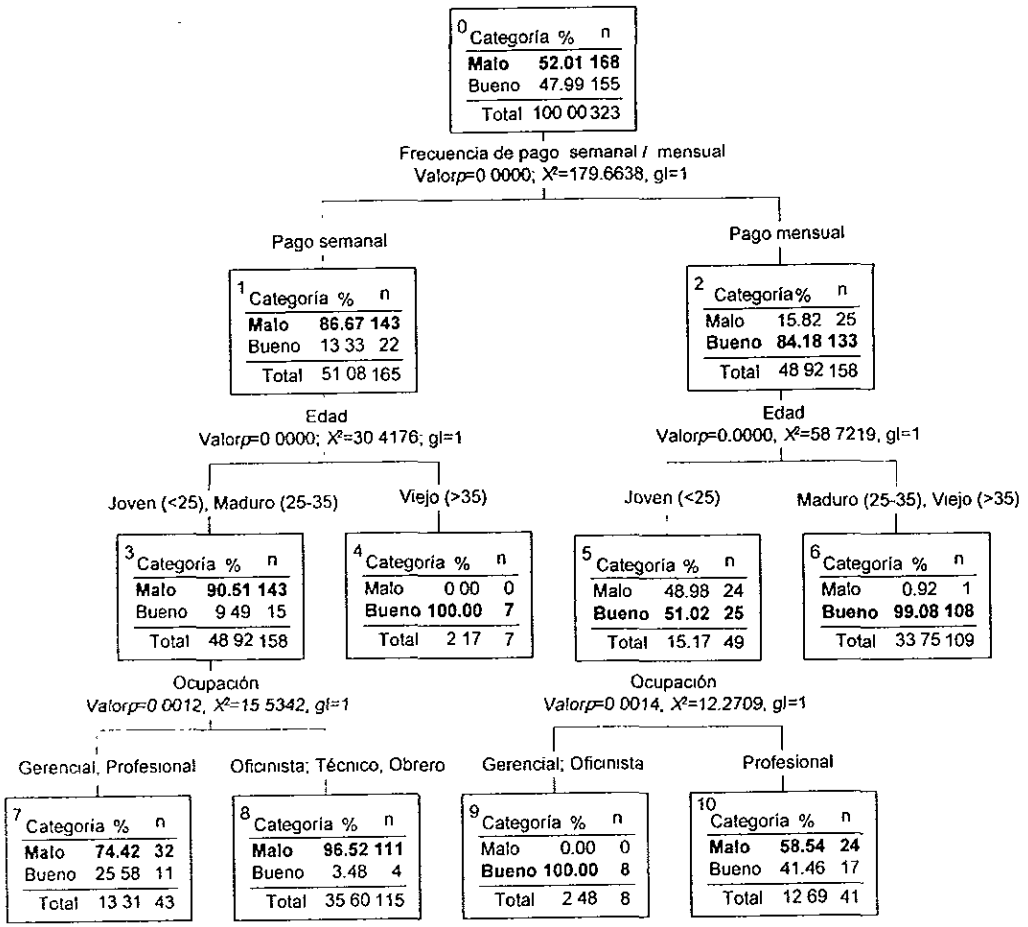

Figura 3.7: Diagrama de árbol con una rama de tres niveles

3.1.2 *Evaluando el árbol*

Una vez terminada la construcción del árbol, la tarea continúa con la evaluación e interpretación de los resultados que se obtuvieron en este análisis

		Categoría real		
		Malo	*Bueno*	**Total**
Categoría predicha	*Malo*	168	155	**323**
	Bueno	0	0	**0**
	Total	**168**	**155**	**323**
Estimador de Riesgo		0.479876		
ES del Estimador de Riesgo		0.027798		

Tabla 3.38: Matríz de clasificación del nivel cero.

Para determinar que tan certero fue el árbol para clasificar los datos, se utilizará la matriz de clasificación (Sección 2.9) como la que se muestra en la Tabla 3.38. Para el nivel cero del árbol (Figura 3.1), donde todavía no se había dividido el nodo raíz, se *tomaron los 323 casos por malos* puesto que son la mayoría, con lo que 155 casos buenos serían mal clasificados. Dado este hecho el estimador de riesgo o porcentaje de error es igual a la tasa de respuesta de créditos buenos o proporción de casos buenos del total de la muestra

$$\begin{aligned} \text{Estimador de Riesgo} &= \frac{155}{323} \\ &= 0\ 47988. \end{aligned}$$

Y el error estándar de este estimador es

$$\begin{aligned} \text{ES del Est. de Riesgo} &= \sqrt{\frac{0.47988 \times (1 - 0.47988)}{323}} \\ &= 2.7798 \times 10^{-2}. \end{aligned}$$

Examinando el primer nivel de división del árbol (Figura 3.4), se observa que existe una mejoría para distinguir entre buenos y malos créditos, y la estimación del riesgo para el primer nivel de división (Tabla 3.39) refuerza esta conclusión. Para hacer el cálculo de este estimador, se ocupará la matriz de clasificación que aparece en la Tabla 3.39; en ésta se indican los números de casos que corresponden a errores específicos de predicción.

		Categoría real		
		Malo	*Bueno*	**Total**
Categoría predicha	*Malo*	143	22	**165**
	Bueno	25	133	**158**
	Total	**168**	**155**	**323**
Estimador de Riesgo		0.145511		
ES del Estimador de Riesgo		0 01962		

Tabla 3.39: Matriz de clasificación del nivel uno.

La matriz de clasificación muestra exactamente qué tipos de errores se cometieron. Los elementos diagonales (izquierda arriba y derecha abajo) de la tabla representan las clasificaciones correctas. Los otros elementos (abajo a la izquierda y arriba a la derecha) representan las malas clasificaciones, es decir, 22 casos de buen crédito, se clasificaron como que tenían un crédito malo y 25 casos se ubicaron como buenos cuando en realidad son malos. El estimador de riesgo es calculado como la proporción de casos en la muestra que fueron clasificados incorrectamente por el algoritmo. Sumando los casos de los errores $(22 + 25 = 47)$ y dividiéndolo entre el total de la muestra (323) el estimador resulta en

$$\begin{aligned} \text{Estimador de riesgo} &= \frac{47}{323} \\ &= 0\ 14551. \end{aligned}$$

El estimador del riesgo es de 0.1455, indicando que si se ocupa una regla de decisión basada en el árbol actual, se podrá clasificar $(100 - 14.55)\% = 85.45\%$ de los casos correctamente. Finalmente, para conocer la dispersión del estimador de riesgo si se utilizaran otros datos, se calculó el error estandar,

$$\begin{aligned} \text{ES del Estimador de Riesgo} &= \sqrt{\frac{0.145511 \times (1 - 0.145511)}{323}} \\ &= 0.0\,1962. \end{aligned}$$

Este resultado se muestra en la Tabla 3.39, y se puede decir que el estimador es estadísticamente confiable.

La Tabla 3.40 muestra la matriz de clasificación para el nivel 3, donde se observa que el árbol final clasifica casi al 90% de los casos correctamente, ya que el estimador de riesgo resultó ser de 0.10217. Para este árbol, primero se nota que el error de clasificar a una persona en crédito malo cuando realmente pertenece al grupo de buen crédito, sólo se cometió una vez. Sin embargo, a 32 casos de buen crédito se les clasificó como que tenían un crédito pobre o riesgoso.

		Categoría real		
		Malo	*Bueno*	**Total**
Categoría predicha	*Malo*	167	32	**199**
	Bueno	1	123	**124**
	Total	**168**	**155**	**323**

Estimador de Riesgo	0.102167
ES del Estimador de Riesgo	0.016852

Tabla 3.40: Matriz de clasificación para el tercer nivel.

3.1.3 Tablas de ganancias

Las tablas de ganancias también ayudan a realizar un análisis útil del árbol. En ellas se muestran cuáles nodos tienen las proporciones más altas y más bajas de una categoría objetivo (categoría que se interesa estudiar) en el nodo. Esta tabla facilita identificar que los subgrupos de solicitantes (nodos) tienen mayor probabilidad de ofrecer buenos riesgos de crédito mediante los porcentajes de ocurrencia para la categoría objetivo o de buen crédito en cada nodo.

Tabla de ganancias detallada

La primera columna de una tabla de ganancias muestra los números de nodos, que corresponden a los que se encuentran en la esquina superior izquierda de cada nodo del árbol final (Figura 3.7). Por ejemplo, el nodo 4 corresponde a solicitantes a quienes se

los paga por semana y tienen más de 35 años. Las siguientes dos columnas (*Nodo: n* y *Nodo: %*) presentan el número de casos y el porcentaje de los mismos que están en el nodo con respecto al total. Las columnas *Resp: n* y *Resp: %*, contienen la frecuencia y el porcentaje con la respuesta objetivo que están en ese nodo. Con base en la segmentación final de este ejemplo, estas columnas representan el número y porcentaje de solicitantes en el nodo con crédito bueno (Tabla 3.41). La columna *Ganancia* indica la proporción de casos en el nodo que tienen la respuesta objetivo (buen crédito), y constituye el criterio para ordenar descendentemente a los nodos en la tabla. En este ejemplo, tanto el nodo 9 (pagos mensuales, menores de 25 años y con ocupación gerencial u oficinista) como el nodo 4, casos con pagos semanales y mayores de 35 años, no tienen casos con créditos malos, así que, los dos tienen un valor de ganancia del 100%. Por el contrario, el nodo 8 (pagos semanales, jóvenes y maduros y con ocupación de oficinista, técnico u obrero) sólo clasificó a 4 casos como crédito bueno lo que da un valor de ganancia de solamente 3.4783

La columna *Indicador* presenta una medida del porcentaje de créditos buenos o de bajo riesgo en este nodo, en comparación con el promedio de la muestra completa. En este ejemplo, como el porcentaje de casos de buenos créditos de la muestra total fue de 47.99%, y en los nodos 9 y 4 fue del 100%, que representa más del doble de lo obtenido en la muestra completa, el indicador de ganancia es de 208.39%. Claramente, éstos son los casos que se quiere buscar.

	Nodo por nodo					
Nodo	Nodo n	Nodo: %	Resp: n	Resp %	Ganancia (%)	Indicador (%)
9	8	2.48	8	5.16	100.0000	208 3871
4	7	2.17	7	4.52	100 0000	208 3871
6	109	33.75	108	69.68	99 0826	206.4753
10	41	12 69	17	10 97	41.4634	86 4044
7	43	13.31	11	7.10	25.5814	53 3083
8	115	35.60	4	2.58	3 4783	7.2483
Total	323	100.00	155	100.00		

Tabla 3.41: Tabla de ganancias detallada nodo por nodo.

La tabla de ganancias detallada acumulativa (Tabla 3.42) presenta las mismas estadísticas que la tabla de ganancias anterior con la salvedad de que son acumuladas.

	Acumulativas					
Nodo	Nodo n	Nodo %	Resp n	Resp %	Ganancia (%)	Indicador (%)
9	8	2.48	8	5 16	100.0000	208 3871
4	15	4.64	15	9.68	100 0000	208 3871
6	124	38.39	123	79.35	99.1936	206.7066
10	165	51.08	140	90.32	84.8485	176.8133
7	208	64.40	151	97 42	72 5962	151 2810
8	323	100.00	155	100.00	47.9876	100.0000

Tabla 3.42: Tabla de ganancias detallada acumulativa.

En el nodo 8, el indicador es del 100% porque el porcentaje de casos es equivalente al de la muestra total. Como este nodo tiene las proporciones más bajas de créditos buenos, suena razonable evitar préstamos a solicitantes que cumplan con este perfil.

Tabla de ganancias resumen

A continuación, se ejemplificará el uso de las tablas de ganancias resumen como una herramienta de búsqueda dinámica. Con este fin, en cada nivel del árbol se necesita construir la tabla correspondiente para evaluar los pasos de la segmentación.

La tabla de ganancias resumen presenta los resultados de la segmentación en deciles, ordenando la muestra de acuerdo a las tasas de respuesta de los segmentos. La que corresponde al nivel cero de este ejemplo de clasificación de créditos se muestra en la Tabla 3.43. La primera columna de una tabla de ganancias resumen, presenta los nodos del árbol ordenados de mayor a menor con respecto a su ganancia o puntaje. Como todavía no se ha dividido el nodo raíz, la tabla sólo muestra este nodo.

La columna *Percentil* muestra en cuantos percentiles se dividió la muestra, para este análisis se utilizan deciles por lo que, la muestra total de 323 se dividió en 10 partes iguales. La siguiente columna *Percentil: n* muestra de cuántos casos está compuesto cada decil. Las columnas *Resp: n* y *Resp: %* indican el número y el porcentaje de casos

de créditos buenos en cada decil. Las dos columnas finales muestran las estadísticas acumuladas *Ganancia* e *Indicador* que representan el porcentaje de créditos buenos en los nodos y la ganancia de cada nodo con base en la ganancia promedio (ver sección 2.7.2).

Nodos	Percentil	Percentil: n	Resp: n	Resp: %	*Ganancia (%)*	*Indicador (%)*
0	10	32	15	10 0	47.988	100.000
0	20	65	31	20.0	47.988	100.000
0	30	97	46	30.0	47 988	100.000
0	40	129	62	40.0	47.988	100 000
0	50	162	77	50 0	47.988	100.000
0	60	194	93	60.0	47.988	100.000
0	70	226	108	70.0	47 988	100 000
0	80	258	124	80.0	47.988	100.000
0	90	291	139	90.0	47.988	100.000
0	100	323	155	100.0	47.988	100.000

Figura 3.43: Tabla de ganancias resumen del nodo raíz.

Para el nodo raíz, la columna de *Ganancia* o también conocida como puntaje *(gain* o *score)* muestra que bajo el modelo base, el cual no considera ninguna variable para predecir el tipo de crédito de un solicitante, la tasa de créditos buenos es del 47.99% sin importar si 10%, 20%, 90% o 100% de la muestra fuera utilizada. Del último renglón se observa que, el número total de créditos buenos en los 323 casos que forman la muestra completa es de 155. Cada decil se incrementa en un número constante de créditos buenos (15.5), ya que hasta ahora no se tienen más elementos que permitan decidir en que nodos se concentran más los créditos buenos.

Se seleccionó *frec_pago* como primera predictora (Figura 3.4). Los 323 créditos se dividen en dos grupos: *semanal* y *mensual.* La tabla de ganancias resumen resultante (Tabla 3.44) muestra que el mejor decil contiene 27 casos o 17.5% de 155 créditos buenos. Para este decil, se predice un porcentaje de créditos buenos del 84.17%. Este decil está formado por una parte de los solicitantes de crédito a quienes se les paga mensualmente (32 de 158).

Nodos	Percentil	Percentil: n	Resp: n	Resp: %	Ganancia (%)	Indicador (%)
2	10	32	27	17 5	84 177	175.414
2	20	65	54	35.1	84 177	175.414
2	30	97	81	52.6	84 177	175.414
2	40	129	108	70.2	84 177	175 414
2;1	50	162	133	86.1	82 642	172 215
1	60	194	137	88.9	71 090	148 143
1	70	226	142	91.7	62.839	130 949
1	80	258	146	94.4	56 651	118 054
1	90	291	150	97.2	51.838	108.024
1	100	323	155	100.0	47 988	100 000

Tabla 3.44: Tabla de ganancias resumen del primer nivel.

Al segmentar nuevamente la población, se dividieron los grupos de solicitantes con pago de salario mensual y semanal bajo la variable *edad*. La tasa de créditos buenos para el mejor decil es ahora de 99.28% (Tabla 3.45). Este decil contiene todos los solicitantes con pago semanal de salario y con más de 35 años (7 casos), más una parte de los solicitantes con pago mensual y con edad mayor a los 25 años (25 casos).

Nodos	Percentil	Percentil: n	Resp: n	Resp: %	Ganancia (%)	Indicador (%)
4;6	10	32	32	20 7	99.281	206.890
6	20	65	64	41.3	99.182	206.682
6	30	97	96	62 0	99.149	206.613
6;5	40	129	121	78.5	94.222	196.346
5	50	162	138	89 2	85.582	178 341
5;3	60	194	142	92 1	73.650	153 478
3	70	226	145	94.1	64.485	134.378
3	80	258	148	96 0	57 611	120 054
3	90	291	151	98.0	52.265	108 913
3	100	323	155	100.0	47.988	100.000

Tabla 3.45: Tabla de ganancias resumen del segundo nivel.

Finalmente, la Tabla 3.46 resume los resultados de la segmentación final. El árbol se dividió en 6 nodos, ordenados en la tabla de acuerdo a su ganancia, es decir, al porcentaje de créditos buenos predichos en un nodo. La ganancia del mejor decil es ahora de 99.509%. De esta tabla, también se observa que el 40% de la muestra (correspondiente al percentil 40), se forma con los segmentos 9, 4 y 6 y una porción del 10. Hasta este percentil se acumula un porcentaje promedio de créditos buenos (ganancia) del 96.87% representando

el 80.7% del total.

Nodos	Percentil	Percentil: n	Resp: n	Resp: %	Ganancia (%)	Indicador (%)
9;4;6	10	32	32	20 7	99.509	207.363
6	20	65	64	41 4	99.296	206.919
6	30	97	96	62.0	99.225	206 771
6;10	40	129	125	80.7	96.870	201.865
10	50	162	138	89.4	85 789	178.773
10;7	60	194	147	95.1	76.041	158.460
7;8	70	226	151	97.8	67 063	139 751
8	80	258	152	98.6	59.115	123 188
8	90	291	153	99.3	52.933	110 306
8	100	323	155	100.0	47 988	100.000

Figura 3.46: Tabla de ganancias resumen de la segmentación final.

3.2 SEGMENTACIÓN DE MERCADO

Actualmente, CHAID se aplica ampliamente en la segmentación de mercado. El ejemplo de esta sección servirá para ilustrar y comparar los resultados de tres diferentes segmentaciones de CHAID en el análisis de una promoción por correo para la suscripción a una revista

La finalidad de este estudio es identificar a los mejores y peores segmentos que responden a un correo para suscribirse a una revista para que, las futuras campañas de mercadotecnia y ventas atiendan fundamentalmente a los buenos prospectos. En la primera segmentación, los datos se analizan usando el algoritmo original de CHAID, explicando el uso y construcción de las tablas de ganancias. La segunda y tercera segmentación, se emplearán para explicar y comparar los algoritmos nominal y ordinal utilizando puntajes de categoría.

Datos

El conjunto de datos se tomará del archivo de SPSS que está incluido en la instalación del paquete CHAID de SPSS[1] y su nombre es *subscrib.sav*. La muestra está compuesta por

[1] CHAID es marca registrada de SPSS, 1996.

81,040 observaciones con nueve variables para estos datos y una variable de frecuencia que cuenta el número de casos iguales en la muestra, con el fin de disminuir el tamaño del archivo:

Nombre	Descripción	No. de categorías
edad	Edad del jefe de familia.	7
sexo	Sexo del jefe de familia.	2
niños	Presencia de niños en el hogar.	2
ingreso	Ingreso del hogar.	8
tarjeta	Posesión de una tarjeta bancaria.	2
num_personas	Número de personas en el hogar	6
ocupación	Ocupación del jefe de familia.	4
resp2	Respuesta dicotómica a la promoción	2
resp3	Respuesta politómica a la promoción.	3

Tabla 3.47: Variables de un estudio de segmentación de mercado.

La Tabla 3.48 muestra la codificación y las categorías de las variables de este conjunto de datos. Las variables *resp2* y *resp3* proporcionan la respuesta de las personas a la promoción de la suscripción. Para explicar el uso de las tablas de ganancias se utilizará a la variable *resp2* como variable dependiente y, todas las demás variables, exceptuando *resp3*. serán consideradas como variables predictoras.

Variable	Categorías	Frecuencia
edad	18-24	677
	25-34	3,940
	35-44	4,566
	45-54	3,231
	55-64	2,898
	65 o más	7,121
	Desconocido	58,607
sexo	Masculino	61,313
	Femenino	19,727
niños	Si	6,311
	No	74,729
ingreso	Menos de $ 8,000	11,029
	$8,000-$9,999	4,856
	$10,000-$14,999	10,205
	$15,000-$19,999	9,703
	$20,000-$24,999	7,852
	$25,000-$34,999	15,321
	$35,000-$49,999	13,602
	$50,000 o mayor	8,472

Variable	Categorías	Frecuencia
tarjeta	Si	6,083
	No	74,957
num_personas	1	25,384
	2	11,240
	3	4,892
	4	3,187
	5 o más	3,011
	Desconocido	33,326
ocupación	Ejecutivo	4,231
	Obrero	3,910
	Otra	24,029
	Desconocido	48,870

resp2	Respondió	931
	No respondió	80,109
resp3	Respondió y pagó	478
	Respondió y no pagó	453
	No respondió	80,109

Tabla 3.48: Características de las variables de un estudio de segmentación de mercado.

De acuerdo a sus características, cada variable predictora se clasifica en monotónica, libre y flotante. Las variables *sexo, hijos* y *tarjeta* tienen dos categorías lo que, hace innecesaria su clasificación puesto que, CHAID trata a las variables dicotómicas como monotónicas Las variables cuantitativas *edad* y *num_personas* se hicieron categóricas al clasificar las observaciones de acuerdo a intervalos definidos en su escala. Estas variables contienen una categoría de valores faltantes y sus demás categorías siguen una escala ordinal por lo que, serán clasificadas como flotantes, en contraste con la variable *ocupación* que, también posee una categoría de valores faltantes pero, las demás no siguen un orden, por lo que, será considerada libre. Finalmente, la variable *ingreso* es una variable ordinal o monotónica por las características de sus valores. Se utilizará como estadístico *el cociente de verosimilitud* (Sección 1.3.2) y los parámetros para este ejemplo se definen con los siguientes valores,

Nivel de profundidad	=	3
Tamaño mínimo del subgrupo antes de dividir	=	4500
Tamaño mínimo del subgrupo después de dividir	=	1500
Nivel de significancia para unión (α_1)	=	0.05
Nivel de significancia para separación (α_2)	=	0.05

3.2.1 El árbol

En la Figura 3.8 se muestra el árbol resultante del análisis CHAID realizado. El nodo inicial o nodo raíz del árbol contiene 81,040 casos que representan el total de individuos a quienes fue enviada la promoción. De éstos, 931 casos respondieron a la promoción, es decir, el 1.15% de los individuos en la muestra. Este porcentaje es la tasa global de respuesta para el correo de promoción. El porcentaje de casos que no respondieron es de 98.85%. Este dato es complemento del anterior por ser la variable dependiente dicotómica, de tal forma que, al sumarlas se obtienen el 100% de las observaciones.

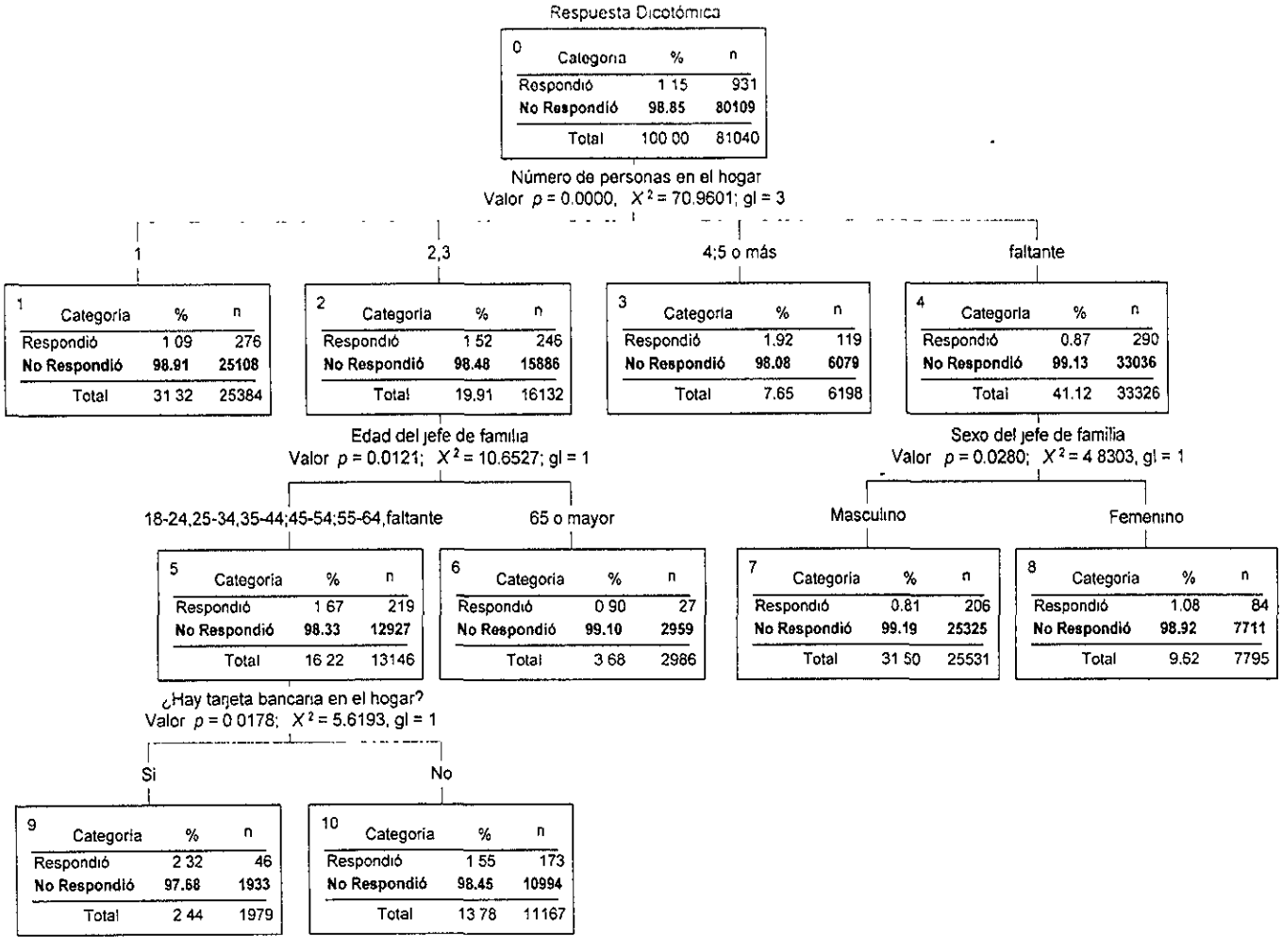

Figura 3.8: Árbol final de la suscripción a una revista con la variable respuesta dicotómica y nominal.

La mejor variable predictora encontrada por CHAID para saber si un individuo respondió o no respondió a la promoción es *num_personas,* por lo que, el nodo raíz se divide en cuatro nodos hijos. Estos nodos se forman con las categorías de la variable *num_personas,* algunas de las cuáles se unen para formar categorías conjuntas por no ser significativamente diferentes. *Num_personas* originalmente tiene 6 categorías, CHAID une las categorías cuyas tasas de respuesta fueron estadísticamente indistinguibles, por esta razón, los hogares de 2 y 3 personas se unieron en una sola categoría y los hogares de 4 personas fueron combinados con hogares de 5 o más personas. Por lo tanto, después de unir, *num_personas* contiene 4 categorías.

En el primer nodo (numerado así en la esquina superior izquierda), que representa hogares con una persona, se encuentran 25,384 casos y la tasa de respuesta para este grupo es de 1.09%. El nodo identificado por "2;3" representa a los hogares con dos y tres personas, y tuvo una respuesta de 1.52% del total de la muestra, lo que, representa la proporción de 249 casos de 16,132 totales de ese nodo. Los dos últimos grupos en los que se dividió la muestra son los hogares con 4, 5 o más personas y aquellos con un número desconocido de ocupantes, con tasas de respuesta de 1.92% y 0.87% respectivamente. Por estos resultados se infiere que, a mayor número de número de personas en un hogar, mayor es la respuesta.

En el siguiente nivel del árbol, se dividieron los hogares de 2 y 3 personas en dos grupos más pequeños. Uno corresponde a individuos de 18 a 64 años y el otro con los de 65 años o más. Los dos tamaños de muestra de estos dos nodos (2,986 y 13,146) suman el tamaño muestral de su nodo paterno, el de hogares de dos y tres personas con 16,132 casos.

En este mismo nivel, se dividen los hogares con un número desconocido de ocupantes con base en el sexo del jefe de familia. Se observa que el 1.08% de las mujeres respondió a la promoción y sólo el 0.81% de los hombres. La diferencia de estas dos tasas de respuesta fue estadísticamente significativa. En un tercer nivel, el grupo de hogares de 2 y 3 personas y cuya edad está entre 18 y 64 años, se divide en hogares que cuentan con

una tarjeta bancaria y en aquellos que no poseen una.

Como el nivel de profundidad es tres, CHAID detiene el análisis. No se realiza ninguna otra división en los demás nodos, porque ninguna predictora es estadísticamente significativa. Los nodos terminales del árbol son los segmentos finales. Los segmentos son identificados por los números ubicados en la esquina superior izquierda de cada nodo.

La matriz de clasificación para este árbol (Tabla 3.49), indica que, el árbol clasificó todos los datos como que no respondieron, ya que la muestra está dominada por estos casos (80,109 casos de un total de 81,040). El estimador de riesgo es de 1.15%, correspondiente al porcentaje de 931 que sí respondieron en relación con el total de la muestra (81,040).

		Categoría real		
		Respondió	*No respondió*	**Total**
Categoría predicha	*Respondió*	*0*	*0*	***0***
	No respondió	931	80109	**81040**
	Total	**931**	**80109**	**81040**
Estimador de Riesgo		0.0114882		
ES del Estimador de Riesgo		0.0003743		

Tabla 3.49: Matriz de clasificación.

La clasificación está correcta en un 98.85% pero, no ayuda a identificar prospectos buenos de malos como es el propósito de este estudio. Para este fin, la siguiente sección se ocupará de las tablas de ganancias.

Adicionalmente se pueden utilizar los resultados del análisis realizado para elaborar un cuestionario que facilite la labor del encuestador y ayude a determinar el mercado objetivo. Un ejemplo de este cuestionario se encuentra en el Apéndice B.

3.2.2 Tabla de ganancias detallada

Una vez descrito y evaluado el árbol resultante de segmentación, se resumirán los resultados obtenidos con la construcción de *tablas de ganancias detalladas*. Como el análisis está

basado en una variable dependiente dicotómica, las tablas de ganancias estarán basadas en el porcentaje de respuesta.

El análisis CHAID generó 7 segmentos finales que se muestran en la Tabla 3.50.

Segmento	Características	Tamaño
9	Hogares con 2 y 3 personas, jefe de familia entre 18 y 64 años y con tarjeta bancaria	1,979
3	Hogares con 4, 5 o más personas.	6,198
10	Hogares con 2 y 3 personas, jefe de familia entre 18 y 64 años y sin tarjeta bancaria.	11,167
1	Hogares con una sola persona.	25,384
8	Hogares con un número desconocido de personas cuyo jefe es mujer.	7,795
6	Hogares con 2 y 3 personas cuyo jefe de familia tenga más de 64 años.	2,986
7	Hogares con un número desconocido de personas cuyo jefe es hombre.	25,531

Tabla 3.50: Segmentación final.

La tabla de ganancias detallada nodo por nodo para el ejemplo se muestra en la Tabla 3.51. En ésta, los segmentos están ordenados de mayor a menor con respecto a la tasa de respuesta o *ganancia* (*gain*) de cada grupo. Para cada segmento, se presenta un renglón con estadísticas. La primera columna etiquetada como *Nodo* muestra los números de identificación de los segmentos o nodos en el árbol. La columna *Nodo: n* indica el tamaño de cada segmento, por ejemplo, el segmento 9 está compuesto por 1,979 observaciones de la muestra. La suma de esta columna es el total de la muestra.

	Nodo por nodo					
Nodo	Nodo: n	Nodo: %	Resp: n	Resp: %	Ganancia (%)	Indicador (%)
9	1,979	2 44	46	4.94	2.32441	202.3307
3	6,198	7.65	119	12.78	1.91997	167.1264
10	11,167	13 78	173	18.58	1.54921	134.8526
1	25,384	31.32	276	29.65	1.08730	94 6452
8	7,795	9 62	84	9.02	1.07761	93.8022
6	2,986	3.68	27	2.90	0 90422	78.7089
7	25,531	31.50	206	22.13	0 80686	70.2343
Total	81,040	100 00	931			

Figura 3.51: Tabla de ganancias detallada nodo por nodo.

El porcentaje que representa el tamaño del segmento con respecto al total, se muestra en la columna, *Nodo: %*. El 2 44 que se designa para el nodo 9, se obtiene al dividir el tamaño del segmento (1,979) entre el tamaño del total de la muestra (81,040) y multiplicarlo por 100. Es decir,

$$\frac{1,979}{81,040} \times 100 = 2.44$$

La columna *Resp: n* muestra el número de casos que respondieron en cada segmento. La suma de todos los segmentos de esta columna es el número total de casos que respondieron en la muestra. La estadística *Resp: %* es el porcentaje del total de casos que respondieron en cada segmento. Los valores de *Resp: %* se obtienen al dividir cada uno de los valores de *Resp: n* entre el total de quienes respondieron; por ejemplo, para el nodo 9 hay 46 casos y el total de respuestas es de 931; entonces

$$\frac{46}{931} \times 100 = 4.94.$$

La *Ganancia* (*gain*) se refiere al porcentaje de casos que responden en un segmento. Ésta se calcula dividiendo el número de respuestas entre el tamaño del segmento, por ejemplo, para el segmento 9, la ganancia es $\frac{46}{1,979} \times 100 = 2.32\%$.

Por último, la columna *Indicador* mide la ganancia promedio de respuesta para el segmento relativo a la ganancia promedio de la muestra. Para el segmento 9, el puntaje de respuesta se acaba de calcular y es 2.32% y el porcentaje total ($\frac{931}{81040}$) es de 1.15%. Realizando la división y multiplicando por 100 se obtiene un resultado de 202. Esto quiere decir que, la tasa de respuesta para el segmento fue 102% más alta que el promedio. También se puede calcular el índice dividiendo el porcentaje de respuesta de ese nodo con respecto al total de respuestas (*Resp: %*) entre el porcentaje de casos de ese nodo con respecto a toda la muestra (*Nodo: %*). Para el nodo 9, los cálculos serán : $\frac{4\,94}{2\,44} = 202\%$. Similarmente, para los demás segmentos se realizan los mismos cálculos.

De esta tabla de ganancias detallada se observa que, el segmento 9 tiene la tasa de

respuesta más alta con 2.32% y le sigue el segmento 3 con 1 92%. Los hogares con tarjeta bancaria probablemente responderán mejor a promociones por correo por sus respuestas altas (2.32% y 1.55%). Por el contrario, los hogares con un número desconocido de ocupantes y con jefe de familia masculino ocuparon el último lugar, ya que, la tasa de respuesta para este segmento fue de 0.81% que, en comparación con la tasa global de respuesta (1.15%), es significativamente baja. De estos resultados, se puede concluir que los hogares con 2 y 3 personas cuyo jefe de familia tiene un empleo de oficina son los prospectos más potenciales a considerarse en una nueva promoción.

	Acumulativas					
Nodo	Nodo: n	Nodo: %	Resp: n	Resp: %	Ganancia (%)	Indicador (%)
9	1,979	2 44	46	4.94	2.3244	202.3307
3	8,177	10.09	165	17.72	2.0179	175.6466
10	19,344	23.87	338	36.31	1.7473	152.0968
1	44,728	55 19	614	65.95	1.3727	119.4919
8	52,523	64.81	698	74 97	1.3289	115.6793
6	55,509	68.50	725	77.87	1.3061	113.6905
7	81,040	100.00	931	100.00	1.1488	100.0000

Tabla 3.52: Tabla de ganancias detallada acumulativa.

Es posible hacer, de manera similar, una tabla de ganancias acumulativas como la que se muestra en la Tabla 3.52. En esta tabla se representan las mismas estadísticas que en la de ganancias detallada nodo por nodo elaborada anteriormente, con la única diferencia de que en ésta, los valores representados son acumulados.

Para el segmento 3, se realizarán los cálculos para ver la similaridad de éstos con los de la tabla pasada. En la columna *Nodo: n* y *Resp: n*, se acumulan los valores de los tamaños de los segmentos superiores con base en el orden de la tabla de ganancias marginal. Para obtener el valor correspondiente al segmento 3, en esta tabla, se suma el tamaño del segmento 9 (1,979) con el tamaño del 3 (6,198), lo que da un total de 8,177. Para el *Nodo: %* se divide $\frac{8,177}{81,040} \times 100 = 10.09\%$ y *Resp* : % $= \frac{165}{931} \times 100 = 17\,72$. La *ganancia* $= \frac{165}{8,177} \times 100 = 2.0179$ y el *indicador* $= \frac{2\,0179}{1.15} \times 100 = 175.65\%$.

En la tabla acumulativa se puede observar el porcentaje acumulado de los segmentos

con mayor respuesta. En este ejemplo, los cuatro segmentos con respuesta más alta conforman el 55.19% del total de la muestra y tienen una respuesta combinada de 1.37%. El *indicador* mide la puntuación de respuesta acumulativa para los segmentos, relativa al puntaje promedio para el total de la muestra. Juntando los tres primeros segmentos, la tasa de respuesta es 52% más alta que el promedio.

Las tablas de ganancias son de gran ayuda para la segmentación de mercado. Por ejemplo, si la cantidad de correo a mandar fuera preestablecida a 40,000 correos, entonces se podrían mandar a los mejores 40,000 hogares de los segmentos con tasas más altas de respuesta, para ahorrar dinero y esfuerzo en la estrategia de mercado. Para más usos, ver Derrick y Magidson (1992) y Magidson (1993a).

3.2.3 Tabla de ganancias resumen

La *tabla de ganancias resumen* presenta los resultados que se hubieran obtenido si la promoción hubiera sido enviada a los segmentos con más alta respuesta. En la Tabla 3.53 se muestran 10 renglones que representan deciles acumulativos basados en el orden de la muestra de acuerdo a las tasas de respuesta predichas de los segmentos finales.

Nodos	Percentil	Percentil: n	Resp: n	Resp: %	Ganancia (%)	Indicador (%)
9;3	10	8,104	163	17 5081	2 019	175 723
3;10	20	16,208	289	31.0419	1.786	155.433
10;1	30	24,312	392	42.1053	1 612	140.357
1	40	32,416	480	51.5575	1 481	128.929
1	50	40,520	568	61.0097	1.402	122.072
1;8	60	48,624	655	70.3545	1.349	117.434
8;6;7	70	56,728	734	78.8400	1.295	112 757
7	80	64,832	800	85.9291	1 234	107 441
7	90	72,936	865	92.9108	1.187	103 307
7	100	81,040	931	100 0000	1 149	100.000

Tabla 3.53: Tabla de ganancia resumen.

En la primera columna, *Nodos*, se listan los que forman cada decil. Cada decil se forma añadiendo los nodos que, tienen la tasa de ganancia o puntaje más alto. Por ejemplo, el primer decil (el 10% de la muestra) lo forman los nodos 9 y 3. De la tabla

de ganancias nodo por nodo (Tabla 3.51) se observa que el nodo 9 es el que tiene la ganancia más alta y el número de casos que lo integran es de 1,979. Cada decil debe tener el 10% de la muestra lo que corresponde a 8,104 casos. Como el número de casos del nodo 9 no es suficiente para alcanzar esta cifra, se añaden casos del siguiente nodo con ganancia más alta, que como se observa en la Tabla 3.51, es el nodo 3. El nodo 3 tiene 6,198 casos; sin embargo, no se puede pasar del tamaño de cada decil por lo que, sólo se aumentan el número de casos faltante para llegar al 10% de la muestra, es decir, 6,125. Las observaciones restantes (73) sirven para formar el siguiente decil y así con los demás deciles.

La columna *Percentil: n* hace referencia al número acumulado de casos del total de la muestra que debe contener cada cuantil. Como se mencionó anteriormente, cada decil contiene exactamente un décimo de la muestra (8,104 casos), por lo que, esta columna se va incrementando en esta cantidad hasta llegar a 81,040.

La siguiente columna *Resp: n* muestra el número de casos que respondieron a la promoción correspondiente a cada decil. Este número está basado en la participación de cada segmento en el cuantil, por lo que, para obtenerlo se calcula la proporción de casos de cada segmento que forman el cuantil. El nodo 9 tiene 46 respuestas y como se ocuparon la totalidad de sus casos para formar el cuantil, se consideran las 46 respuestas para el primer cuantil más las respuestas aportadas por el nodo 3. El nodo 3 reporta 119 respuestas totales pero, como sólo se ocuparon 6,125 casos de 6,198 totales del nodo 3, lo que representa un 98.82%, esta misma proporción se multiplica por el número de respuestas en el nodo 3 (119) obteniéndose un resultado de 117 respuestas. Sumando las respuestas del nodo 9 (46) y las correspondientes al nodo 3 (117) se obtienen 163 respuestas para el primer decil.

La columna, *Resp: %, indica el porcentaje de respuestas del total de cada cuantil.* Para el primer cuantil, esto resulta de dividir el número de respuestas por cuantil entre el total de respuestas, $\frac{163}{931} = 17.50\%$. La *ganancia* se obtiene de sumar las ganancias de cada nodo que forman el cuantil por la proporción del cuantil perteneciente al nodo. El

primer cuantil está conformado en un 24.42% con observaciones del nodo 9 y el restante 75.58% proviene del nodo 3. Multiplicando estos datos por las ganancias de cada nodo y sumando los resultados se obtiene: $(.2442 \times .0232 + .7558 \times .0192) \times 100 = 2.01\%$. La ganancia indica el porcentaje de respuesta promedio por cuantil.

Finalmente, la columna *Indicador* (%), se refiere a la ganancia promedio del cuantil con relación a la ganancia promedio del total de la muestra, para lo cual, se divide la ganancia entre el porcentaje de respuestas promedio del total de la muestra, por ejemplo, para el primer cuantil, la ganancia es de 2.01, y el porcentaje de respuesta del total de la muestra es $(\frac{931}{81040}) \times 100 = 1.15\%$. Entonces, el indicador para el primer cuantil es $(\frac{2.01}{1.15}) \times 100 = 175.72\%$. Lo que refleja que, la respuesta del primer cuantil es 75% mayor que la respuesta promedio de la muestra.

Si las ganancias reflejaran costos o beneficios, este tipo de tablas serían útiles para cuantificar el costo asociado con cada segmento final del análisis y determinar los segmentos a los cuáles enviar por correo futuras promociones. Por ejemplo, si se quiere tomar en cuenta el costo del envío por correo, tanto a los que pagan la suscripción como a los que que no. y la ganancia de los que contestan y pagan, para determinar a qué porcentajes de hogares se debe enviar información promocional y se calcula una tasa de respuesta de punto de equilibrio en la que los costos sean igual a las ganancias o en la que no se gane ni se pierde, y se establece igual a 1.45%. Con base en la tabla de ganancias de la segmentación final, la decisión para obtener un beneficio consistirá en enviar a sólo el 40% de los hogares porque el puntaje para estos (1.48%) excede el nivel de equilibrio. Segmentos abajo de este nivel no ofrecen ningún beneficio. En la siguiente sección se brinda una explicación más detallada de los puntajes de categoría.

3.2.4 Los puntajes de categoría

Utilizando el mismo conjunto de datos de la suscripción a la revista, se realizará un análisis de CHAID para mostrar el uso de los puntajes de categoría en las variables politómicas. Se ocupará la variable politómica *resp3* como variable dependiente en lugar

de la variable dicotómica del análisis anterior. El uso de la variable *resp3* conduce a sopesar los costos relativos o consecuencias de las diferentes categorías de esta variable. Esta información en CHAID es tratada como puntajes de categoría.

La variable politómica *resp3* tiene dos categorías de sujetos que respondieron (los que pagaron la suscripción y los que no) y una categoría de individuos que no respondieron. Por simplicidad se limitará el tamaño de los árboles estableciendo el nivel de profundidad igual a 2.

A la categoría *respondió y pagó,* se le asignará el puntaje de 500. Este valor representa la cantidad de pesos ganada cuando alguien regresa la forma de suscripción y la paga. A la segunda categoría, *respondió y no pagó,* se le asignará el valor de -100 que representa el costo ($100.00) de enviar información a personas que, regresaron la forma de suscripción pero la cancelaron antes de pagarla. La tercera categoría, *no respondió*, tiene un puntaje de -2 Este valor representa el costo de 2 pesos por enviar un correo a gente que no solicitó la suscripción a la revista.

Método nominal

Si la variable dependiente es tratada como nominal, no se necesitan asignar puntajes de categoría (Sección 2.8), porque sólo se utilizan en las tablas de ganancias y no afectan al algoritmo de segmentación.

El árbol resultante del análisis se muestra en la Figura 3.9. Cada nodo presenta el porcentaje de casos que: respondieron y pagaron la suscripción, sí respondieron pero no pagaron y no respondieron, así como el número de observaciones en cada nodo.

Se observan algunas diferencias con respecto al árbol de *resp2*, se utilizó la misma variable predictora para dividir los segmentos (*num_personas*), pero, en lugar de los 4 obtenidos en el análisis pasado, ahora se tienen 5. La categoría de hogares de 2 personas permaneció sola, difiriendo del análisis anterior en el que se unió con los hogares de 3 personas Esta última categoría se unió en este estudio con los hogares de 4 personas.

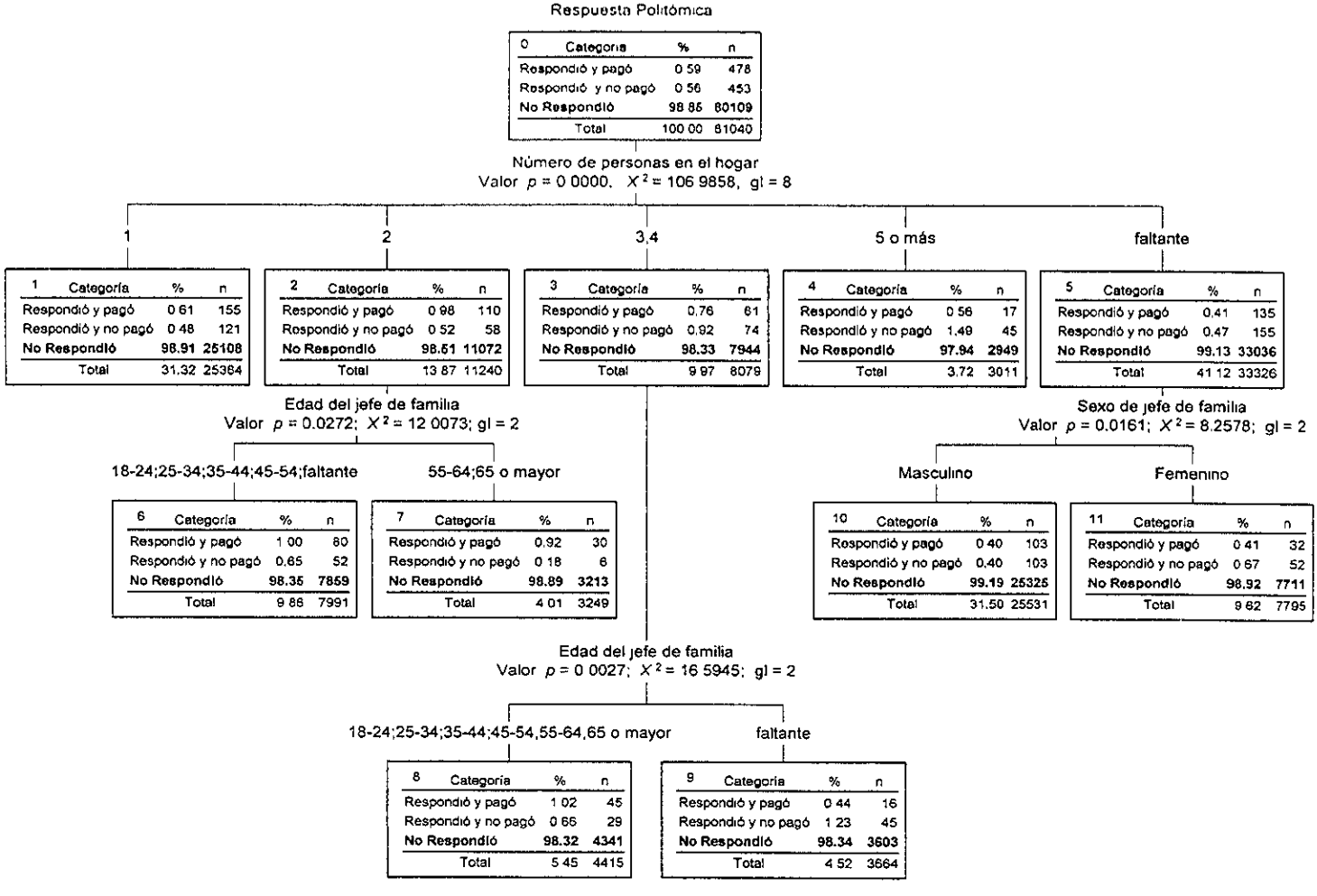

Figura 3.9: Árbol final con la variable dependiente *resp3* considerada nominal.

A partir del segundo nivel, se realizaron divisiones diferentes a las referidas en el análisis de la variable dicotómica *resp2*. El nodo correspondiente a hogares con 2 personas se dividió en aquellos cuyo jefe de familia tiene 55 años o más y aquellos que su edad es desconocida o menor a 55. Los nodos de 3 y 4 personas y aquel con valores faltantes, se dividieron con base en las variables *edad* y *sexo*, respectivamente. Este árbol es el mismo que el que resulta cuando no se asignan puntajes de categoría.

Con base en la categoría de personas que respondieron a la promoción y pagaron la suscripción, se elaboró la tabla de ganancias detallada de este árbol que se presenta en la Tabla 3.54. La columna *Resp: n* indica el número de personas que sí respondieron y pagaron y la *Ganancia* es el porcentaje de individuos que sí respondieron y pagaron la suscripción en cada nodo. Sin embargo, se observa que esta tabla no toma en cuenta los costos y ganancias de cada categoría de la variable dependiente, por lo que, a simple vista, parece difícil determinar cuáles son los segmentos más ventajosos.

	Nodo por nodo						Acumulativas	
Nodo	Nodo: n	Nodo %	Resp: n	Resp. %	Ganancia (%)	Indicador (%)	Ganancia (%)	Indicador (%)
8	4,415	5.45	45	9.41	1 01925	172.8038	1 0193	172.8038
6	7,991	9 86	80	16.74	1.00113	169.7307	1 0076	170.8243
7	3,249	4.01	30	6.28	0.92336	156.5464	0.9901	167 8611
1	25,384	31.32	155	32.43	0.61062	103.5245	0.7554	128 0668
4	3,011	3.72	17	3 56	0 56460	95.7215	0 7423	125.8558
9	3,664	4.52	16	3.35	0.43668	74.0348	0 7189	121.8765
11	7,795	9.62	32	6.69	0.41052	69.5994	0.6756	114 5353
10	25,531	31 50	103	21 55	0.40343	68 3976	0 5898	100.0000
Total	81,040	100.00	478					

Tabla 3.54: Tabla de ganancias detallada.

La Tabla 3.55 facilita esta decisión porque muestra las ganancias de los segmentos finales, particularmente, las ganancias promedio asociadas con mandar promociones a un hogar que pertenezca a un nodo final. Los segmentos están ordenados de mayor a menor basados en la ganancia promedio. Evidentemente, los segmentos más altos son los más promisorios para futuras promociones.

Nodo	Nodo por nodo				Acumulativas			
	Nodo: n	Nodo: %	Ganancia ($)	Indicador (%)	Nodo: n	Nodo: %	Ganancia ($)	Indicador (%)
8	4,415	5.45	2.47	598.54998	4,415	5 45	2 47	598.54998
7	3,249	4 01	2.45	594.03846	7,664	9.46	2 47	596.63741
6	7,991	9.86	2.39	577.97733	15,655	19.32	2.43	587.11248
1	25,384	31.32	0.60	144.78187	41,039	50.64	1.30	313.51614
10	25,531	31.50	-0.37	-89.58845	66,570	82.14	0.66	158 91702
11	7,795	9.62	-0.59	-143.51651	74,365	91.76	0 53	127 21569
4	3,011	3.72	-0 63	-152.57153	77,376	95.48	0 48	116.32809
9	3,664	4.52	-1.01	-244.81498	81,040	100.00	0 41	100.00000
Total	81,040	100.00						

Tabla 3.55: Tabla de ganancias promedio.

Esta tabla contiene las mismas columnas que la tabla de ganancias detallada salvo las columnas *Resp: n* y *Resp: %*. Sin embargo, el cálculo de las ganancias y el indicador es diferente de la tablas detalladas nodo por nodo. El segmento 8 es el que tiene la ganancia promedio más alta ($2.47) y representa los hogares con tres o cuatro personas mayores de 18 años. Le sigue el nodo 7 con una ganancia de $2.45 que está formado por hogares de dos personas y mayores de 55 años. Los puntajes o ganancias negativos indican que, en promedio, se pierde dinero en ese nodo. Este es el caso de los hogares con cinco o más personas, en los que se tiene una pérdida de $0.63. El segmento que mayores pérdidas reportó es el de hogares con 3 o 4 personas y cuya edad del jefe de familia que fue desconocida.

Para los hogares de una persona, la ganancia promedio es de $0.60 por cada hogar. Estas ganancias o pérdidas se obtuvieron de la siguiente manera: del árbol de porcentajes, considere el nodo de los hogares compuestos por una sola persona; para obtener la ganancia promedio de ese nodo, se deben ponderar los porcentajes de cada categoría de la variable dependiente con los puntajes que tienen asignados cada categoría, en este caso,

$$\begin{aligned} \text{Gan. prom. hogares de 1 persona} &= (0.61 \times 500 + 0.48 \times (-100) + 98\,91 \times (-2)) \div 100 \\ &= 0.60 \end{aligned}$$

donde el 0.61 representa el porcentaje de hogares con una persona, que respondieron a la promoción y sí pagaron la suscripción, 0.48% se refiere a los hogares de una persona que sí respondieron a la promoción pero no pagaron la suscripción y, finalmente, 98.91% son aquellos que no respondieron a la promoción. Los valores de 500, -100 y -2 son los puntajes asignados a cada categoría y representan ganancias o pérdidas.

Dado lo anterior, la ganancia del total de la muestra se calcula como:

$$\begin{aligned} \text{Gan. prom total de la muestra} &= (0.59 \times 500 + 0.56 \times (-100) + 98.85 \times (-2)) \div 100 \\ &= 0.41 \end{aligned}$$

que representa un promedio de ganancia de $0.41 por cada hogar en la muestra. Esta cifra se observa al final de la columna *ganancia acumulativa* de la Tabla 3.55. Esta columna representa la ganancia combinada de los segmentos. El indicador resulta entonces de dividir cada ganancia promedio del nodo entre la ganancia promedio de la muestra (0.41), por ejemplo. para el nodo 8 es

$$\left(\frac{2.47}{0.4131}\right) \times 100 = 598.55\%.$$

El análisis de las ganancias promedio puede ayudar a mejorar la elección de los segmentos más beneficiosos. Obsérvese que, aunque la segmentación no ha cambiado, las Tablas 3.54 y 3.55 presentan en diferente orden los nodos finales. Así, por ejemplo, las ganancias en la Tabla 3.55 sugieren no mandar correos a personas en el segmento 4 (con una pérdida de $0.63), a pesar de que en la Tabla 3.54 tiene un porcentaje de respuesta del 0 56%.

Si se quisiera asegurar una ganancia promedio de 1 por hogar, se deberán contemplar los nodos 8,7,6 y 1 para futuras campañas. Finalmente, la tabla de ganancias resumen se presenta en la Tabla 3.56.

Nodos	Percentil	Percentil: n	Ganancia ($)	Indicador (%)
8;7;6	10	8,104	2.46	595.62
6;1	20	16,208	2.36	572 02
1	30	24,312	1.77	429 61
1	40	32,416	1.48	358 40
1	50	40,520	1 30	315.68
1;10	60	48,624	1.04	250.63
10	70	56,728	0.83	202.03
10	80	64,832	0.68	165.58
10;11	90	72,936	0.55	132.52
11;4;9	100	81,040	0.41	100.00

Tabla 3.56: Tabla de ganancias resumen método nominal.

Método ordinal

En esta sección se continuará el análisis CHAID tratando a la variable dependiente (*resp3*) como ordinal para tomar en consideración los costos y beneficios asociados con cada una de las categorías de la variable dependiente en el criterio de segmentación. Los puntajes asignados de 500, -100 y -2, permanecen sin cambios.

El algoritmo tradicional identifica segmentos basados en diferencias en la distribución de la variable dependiente, sin importar si éstas están relacionadas a la ganancia esperada. El nuevo algoritmo identifica segmentos que difieren con respecto al valor o ganancia esperada

Los árboles y las tablas de ganancias resultantes difieren sustancialmente de los obtenidos con el análisis tradicional. El diagrama de árbol para *resp3*, indica que, aunque la variable *num_personas* es todavía utilizada para la primera división, ahora existen 4 categorías unidas en lugar de 5. La Figura 3.10 muestra el diagrama de árbol correspondiente al análisis actual.

En el siguiente nivel del árbol, otra vez, los resultados son diferentes de aquellos suministrados por el método nominal. Mientras que, en este último se sugería que los hogares con una persona eran un segmento homogéneo con una ganancia esperada de 0.60 por hogar (ver Tabla 3.55), el análisis ordinal divide los hogares de una persona, basándose en la variable predictora *ingreso*, en un grupo de bajo ingreso (desde menos

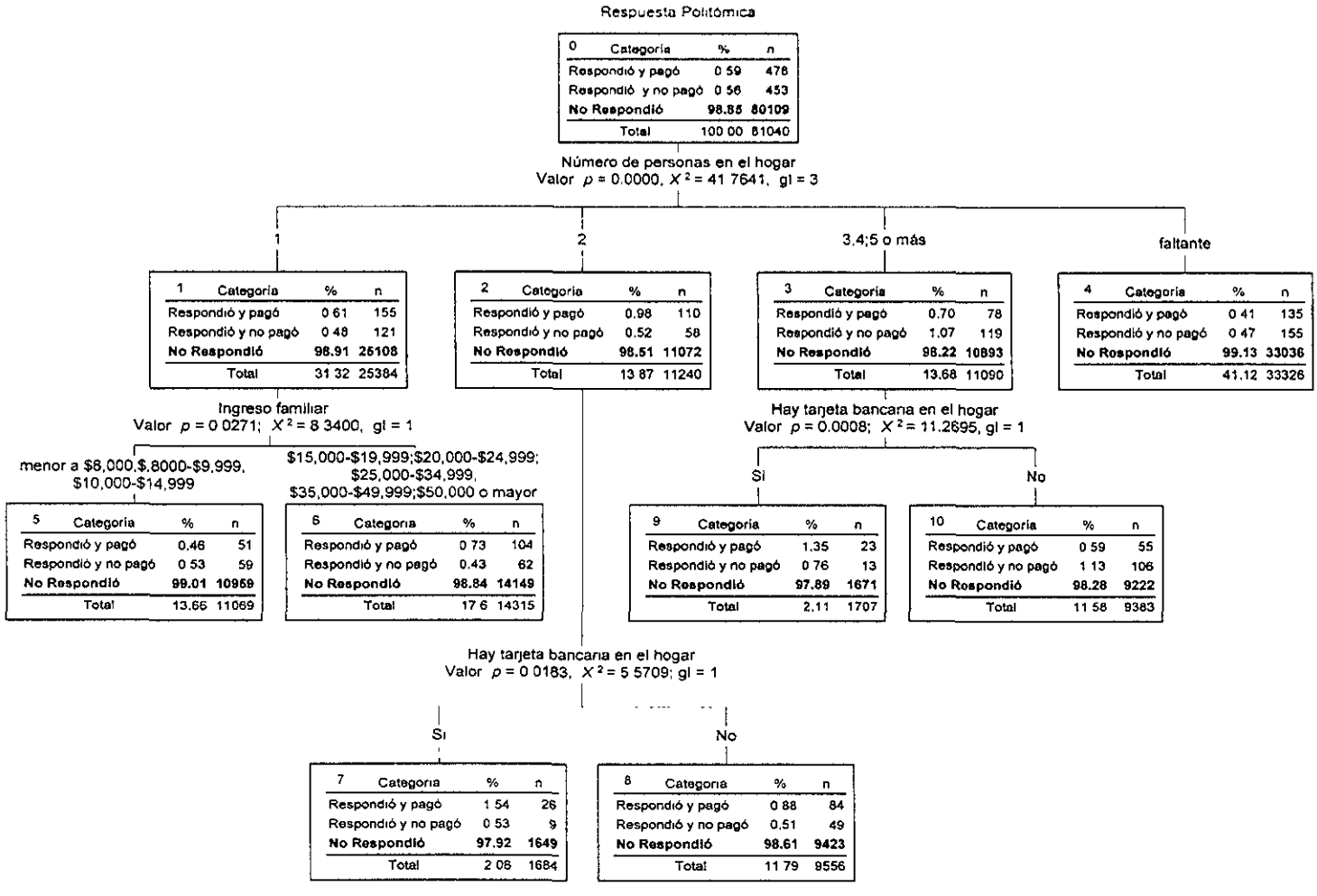

Figura 3.10: Árbol final con la variable dependiente *resp3* considerada ordinal.

de $8,000 hasta $14,999) y en un grupo de alto ingreso (de $15,000 o más). La tabla de ganancias promedio (Tabla 3.57), muestra que los correos a hogares de una persona con altos ingresos proporcionan una ganancia de $17.66, mientras que, los correos a hogares con una persona con bajos ingresos produce una pérdida promedio de $0.21.

	Nodo por nodo				Acumulativa	
Nodo	*Nodo. n*	*Nodo: %*	*Ganancia ($)*	*Indicador (%)*	*Ganancia ($)*	*Indicador (%)*
7	1,684	2 08	5.23	1265.1072	5 23	1265.1072
9	1,707	2 11	4.02	972.4158	4.62	1117.7689
8	9,556	11.79	1 91	462 3490	2.62	634.0126
6	14,315	17 66	1.22	295 9269	1 89	456.4872
10	9,383	11 58	-0.16	-39.8285	1.36	329.4049
5	11,069	13.66	-0 21	-50 6866	1.00	241.2289
4	33,326	41 12	-0.42	-102 2023	0.41	100.0000
Total	81,040	100.00				

Tabla 3.57: Tabla de ganancias detallada.

Los hogares de 2 personas se dividieron de acuerdo a las categorías de la variable *tarjeta*. La ganancia esperada para los hogares con dos personas que poseen una tarjeta bancaria es casi tres veces que el de los hogares con el mismo número de personas pero sin tarjeta bancaria ($5.23 contra $1.91).

Los hogares con tres o más personas también se dividieron con base en la tarjeta bancaria. Los que no poseen ninguna tarjeta bancaria tienen una pérdida esperada de $0.16 por hogar mientras que, los que si tienen aparecen como el segundo subgrupo de mayor ganancia ($4 02).

En general, el diagrama de árbol sugiere, que para futuras promociones a suscripciones de revista, los segmentos más promisorios los representan hogares con dos o más personas y que cuenten con una tarjeta bancaria. Los segmentos que tuvieron el valor más bajo fueron los hogares de tamaño desconocido, por lo que se pueden considerar los menos prometedores.

La tabla de ganancias resumen (Tabla 3.58), indica que el mejor 10% tiene una ganancia promedio de $3.04, comparado con $2.46 del análisis nominal (Tabla 3.56).

Nodos	Percentil	Percentil: n	Ganancia (%)	Indicador (%)
7;9;8	10	8,104	3.04	736.60
8;6	20	16,208	2.34	565.99
6	30	24,312	1.97	475.97
6;10	40	32,416	1.56	377.58
10;5	50	40,520	1 21	293.06
5;4	60	48,624	0.97	234.80
4	70	56,728	0.77	186.66
4	80	64,832	0.62	150.55
4	90	72,936	0.51	122.47
4	100	81,040	0.41	100 00

Figura 3.58: Tabla de ganancias resumen método ordinal.

Para terminar este análisis, es necesario considerar que también es posible transformar la variable politómica *resp3* en una variable dicotómica que compara los casos que responden (que pagan y no pagan la suscripción) con los que no responden asignando el mismo puntaje (100), a cada categoría de los que contestan y 0 a la categoría de los que no responden. El análisis basado en estos puntajes es equivalente al análisis de la variable dicotómica *resp2* en la Sección 3.2.1.

Conclusiones

De lo planteado en los capítulos anteriores, se desprenden tres puntos relevantes: el método de CHAID, los resultados que se obtienen al emplear este método y sus aplicaciones potenciales.

Con respecto al primero, se mostró que esta técnica basa su metodología en la teoría de las tablas de contingencia. Esto permite que un amplio número de usuarios encuentre útil a CHAID, ya que no se requieren conocimientos especializados o vastos para entender el funcionamiento de la técnica.

El algoritmo de CHAID sigue pasos sencillos permitiendo un reducido tiempo de cómputo Comparado con otros buscadores de interacción, CHAID: permite divisiones múltiples minimizando el sesgo de los predictores con más de dos categorías; toma en cuenta valores faltantes o desconocidos que asigna como otra categoría, y se puede aplicar a variables dependientes categóricas y continuas. Todo lo anterior lo hace una técnica más versátil y poderosa con respecto a otras similares.

Gracias a que los resultados se muestran en forma de árbol, se facilita su interpretación, obteniendo un buen uso de ellos sin necesidad de poseer rigurosos conocimientos matemáticos. Asimismo, las extensiones, como las tablas de ganancia han permitido aprovechar los resultados de la segmentación que arroja el análisis. Este es un segundo aspecto que conviene resaltar por su importancia y utilidad.

Para el tercer aspecto, es decir, las aplicaciones potenciales de esta técnica, los ejemplos tratados en esta tesis, han mostrado que CHAID es un buscador efectivo de interacciones entre las variables que permite ahorrar tiempo, dinero y recursos o incrementar las ganancias de las empresas en la búsqueda de perfiles de consumidores o usuarios. Sin embargo, sus beneficios no se reducen al campo de la mercadotecnia. CHAID tiene una amplia, diversa y creciente gama de aplicaciones, en campos como la psicología, las finanzas, la medicina. la demografía, etc. De igual forma. es una herramienta notable

como ayuda a otras técnicas estadísticas para la exploración de datos, facilitando la labor del investigador.

En los últimos cinco años su uso se ha incrementado notablemente; particularmente, en la investigación de mercado, convirtiéndose en una herramienta muy solicitada. Debido a esta popularidad, existen varios paquetes que permiten correr un análisis CHAID; sobresaliendo los CHAID y AnswerTree, ambos de SPSS. Al elegir alguno de estos paquetes, se debe tomar en cuenta que, CHAID de SPSS segmenta, en ocasiones, de acuerdo al valor p no ajustado y, en otras, con respecto al valor p ajustado. AnswerTree siempre toma el valor p ajustado tal y como describe el algoritmo. Estas diferencias conducen a resultados diferentes de segmentación.

Otro campo en el que CHAID puede ofrecer aportaciones valiosas es el de la enseñanza. Dado que emplea conceptos de tablas de contingencia y árboles de manera práctica y evita cálculos complejos, esta técnica de segmentación puede servir para que el estudiante aplique sus conocimientos, reafirme la comprensión de los mismos, realice interpretaciones de datos y/o resuelva problemas prácticos relacionados con la segmentación de poblaciones. Para los docentes e investigadores es una técnica novedosa, eficiente, sencilla y flexible que puede ahorrar tiempo y recursos.

Finalmente, CHAID es una importante plataforma potencial para otros algoritmos que no sólo busquen detectar interaciones en los datos, sino que, que también se interesen en el desarrollo y empleo de otros criterios de segmentación más apropiados para discriminar poblaciones.

Apéndice A

Desigualdad de Bonferroni

Si se tienen k eventos con una probabilidad de $1-\alpha$ para cada uno de ellos, la probabilidad conjunta de todos los eventos no es $1-\alpha$. Para verificar esto, suponga que A_i es el evento i-ésimo y sea $p(A_i) = 1 - \alpha_i$. Si A_i^c denota el evento complemento de A_i, entonces

$$\begin{aligned} p(\bigcap_{i=1}^{k} A_i) &= 1 - p((\bigcap_{i=1}^{k} A_i)^c) \\ &= 1 - p(\bigcup_{i=1}^{k} A_i^c) \\ &\geq 1 - \sum_{i=1}^{k} P(A_i^c) \qquad (A.1) \\ &= 1 - \sum_{i=1}^{k} \alpha_i. \end{aligned}$$

Para el caso en que $\alpha_i = \alpha$, $i = 1, \ldots, k$

$$p(\bigcap_{i=1}^{k} A_i) \geq 1 - k\alpha. \qquad (A.2)$$

Así que, la probabilidad de que pasen todos los eventos no es $1 - \alpha$ sino algo más grande que $1 - k\alpha$. Por ejemplo, si $\alpha = 0.05$ y $k = 10$ entonces $1 - k\alpha = 0.5$. Si k no es muy es grande y α es pequeña, la diferencia en (A.2) no es tan significativa.

Nótese por otro lado que si la dependencia entre los eventos fuera pequeña

$$\begin{aligned} p(\bigcap_{i=1}^{k} A_i) &= p(A_1)p(A_2 \mid A_1)\ldots p(A_k \mid A_1, A_2, \ldots, A_{k-1}) \\ &\cong p(A_1)p(A_2)\ldots p(A_k) \\ &= (1-\alpha_1)(1-\alpha_2)\ldots(1-\alpha_k). \end{aligned}$$

Si se utiliza un nivel de significancia individual de α/k en lugar de α para cada uno de los k eventos, entonces

$$p(\bigcap_{i=1}^{k} A_i) \geq 1 - k(\alpha/k) = 1 - \alpha.$$

Así que la probabilidad conjunta es de al menos $1-\alpha$. La expresión (A.1) es conocida como la desigualdad de Bonferroni.

Apéndice B

Cuestionario de suscripción

Instrucciones: Para cada pregunta marque la opción que identifique su respuesta. Solamente conteste las preguntas que se indican.

1. ¿Cuántas personas habitan en su hogar?

   (a) Una

   (b) Dos o Tres

   (c) Cuatro o más

   (d) Desconocido

   Si la respuesta es (a) o (c), de por terminado el cuestionario.

   Si la respuesta es (b) pase a la pregunta 2.

   Si la respuesta es (d) pase a la pregunta 4.

2. ¿Cuál es la edad del jefe de familia?

   (a) Menor a 65

   (b) 65 o mayor

   (c) Desconocido

3. ¿Alguien en su hogar cuenta con una tarjeta bancaria?

   (a) Si

   (b) No

   De por terminado el cuestionario.

4. ¿El jefe de familia es hombre o mujer?

   (a) Hombre

   (b) Mujer

   De por terminado el cuestionario.

Bibliografía

[1] Agresti, A. (1984). *Analysis of ordinal categorical data.* Jonh Wiley and Sons. New York.

[2] Agresti, A. (1990). *Categorical data analysis.* Jonh Wiley and Sons. New York:

[3] Bishop, Y. M., S. E. Fienberg, y P. W. Holland (1975). *Discrete multivariate analysis.* MIT Press. Cambridge, MA.

[4] Breiman, L., J. H. Friedman, R. A. Olshen, y C. J. Stone (1984). *Classificatión and regression trees.* Wadsworth. Belmont, CA.

[5] Clogg, C. C. y S. R. Eliasin (1987). Some common problems in log-linear analysis. *Sociological Methods and Research*, 16:1, 8-44.

[6] Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, 23, 315-345.

[7] Conover, W. J. (1974). Some reasons for not using the Yates continuity correction on 2×2 contingency tables. *Journal of the American Statistical Association*, 69, 374-376.

[8] Conover, W. J. (1980). *Practical nonparametric Statistics.* Nueva York: John Wiley and Sons.

[9] Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press. Princeton, N. J.

[10] Christensen, R. (1990). *Log-linear models.* Springer -Verlag. New York.

[11] Deming, W. E. y F. F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11: 427-444.

[12] Derrick, F. y J. Magidson. (1992). Using CHAID with the gains chart option. *Proceedings of the 1992 Annual Meeting of the American Statistical Association*, Business and Economics Section.

[13] Dreyfus, S. E. y A. M. Law (1977). *The art and theory of dynamic programming.* Academic Press, New York.

[14] Feller, W. (1968). *An introduction to probability theory and its applications*, Vol I, John Wiley and Sons, New York.

[15] Fienberg, S. E. (1977). *The analysis of cross classified categorical data.* Cambridge, Mass. : MIT Press.

[16] Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53, 759-798.

[17] Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74, 537-552.

[18] Goodman, L. A. (1985). The analysis of cross classified data having ordered and/or unordered categories: association models, correlation models, and asymetry models for contingency tables with or without missing entries. *The Annals of Statistics*, 13:1, 10-69.

[19] Goodman, L. A. (1991). Measures, models and graphics in the analysis of cross classified data. *Journal of the American Statistical Association*, 86:1, 1085-1138.

[20] Grizzle, J. E. (1967). Continuity correction in the χ^2-test for 2×2 tables. *The American Statistican*, 21:4, 28-32.

[21] Haberman, S. (1978). *Analysis of quantitative data.* Vol 1: Introductory topics. Academic Press. New York:

[22] Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29:2, 119-127.

[23] Magidson, J. (1987). Weighted log-linear modeling. *Proceedings of the Social Statistics Division*, American Statistical Association, 171-174.

[24] Magidson, J. (1989). CHAID, LOGIT and loglinear modeling. *Marketing Information Systems Report* 11-130. Datapro Research Corporation. Delran, N. J.

[25] Magidson, J. (1992). Chi squared analysis of a scalable dependent variable. *Proceedings of the 1992 Annual Meeting of the American Statistical Association*, Educational Statistics Section.

[26] Magidson, J. (1993a). The use of the new ordinal algorithm in CHAID to target profitable segments. *Journal of Database Marketing*, 1:1.

[27] Magidson, J., and SPSS Inc. (1993b). *SPSS for Windows CHAID release 6.0.* SPSS Inc.Chicago.

[28] McCullagh, P.(1978) A class of parametric models for the analysis of square contingency tables with ordered categories. *Biometrika*, 65, 413-418.

[29] Montgomery, D. C. (1982). *Introduction to linear regression analysis.* John Wiley and Sons, New York.

[30] Mood, A., F. A. Graybill y D. C. Boes. (1983). *Introduction to the theory of statistics.* McGraw-Hill. Tokio.

[31] Messenger, R. C. y L. M Mandel (1972). A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American Statistical Association*, 67, 768-772.

[32] Myers, J. H. (1996). *Segmentation and Positioning for strategic marketing decisions.* American Marketing Association. Chicago.

[33] Morgan, J. N. y Messenger, R. C. (1973). *THAID- a sequential analysis program for the analysis of nominal scale dependent variables.* Survey Research Center, Institute for Social Research, University of Michigan.

[34] Morgan. J. A. y J. N. Sonquist (1963). Problems in the analysis of survey data: and a proporsoal. *Journal of the American Statistical Association*, 58, 415-434.

[35] Pearson, E. S. (1947). The choice of statistical test illustraded on the interpretation of data classed in a 2×2 tables. *Biometrika*, 51, 327-338.

[36] Serré, J. P. (1982). *Trees.* Springer-Verlag. New York.

[37] Siegel. S y N. J. Castellan (1995). *Estadística no paramétrica.* Trillas, México.

[38] Smith. W. (1956). Product differentiation and market segmentation as alternative marketing strategic. *Journal of Marketing*, 21, 3-8.

[39] Yates. F. (1934). Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society Supplement*, 1, 217-235.