UNIVERSIDAD DE SALAMANCA

Departamento de Estadística

Máster en Análisis Avanzado de Datos Multivariantes

Trabajo Fin de Máster

TAID versus CHAID

Búsqueda de perfiles de mujeres trabajadoras en el servicio doméstico

Autor:

Sergio Ramos Hernández

Tutoras:

María Purificación Galindo Villardón

María del Carmen Patino Alonso

2015

Dpto. de Estadística
Universidad de Salamanca

DRA. Mª PURIFICACIÓN GALINDO VILLARDÓN

Profesora Titular del Departamento de Estadística de la Universidad de Salamanca

DRA. Mª CARMEN PATINO ALONSO

Profesora Ayudante Doctora del Departamento de Estadística de la Universidad de Salamanca

---

CERTIFICAN que **D. Sergio Ramos Hernández** ha realizado en la Universidad de Salamanca, bajo su dirección, el trabajo que para optar título de Máster en Análisis Avanzado de Datos Multivariantes, presenta con el título **TAID versus CHAID. Búsqueda de perfiles de mujeres trabajadoras en el servicio doméstico**, autorizando expresamente su lectura y defensa.

Y para que conste, firman el presente certificado en Salamanca a 25 de junio de 2015.

Mª Purificación Galindo Villardón

Mª Carmen Patino Alonso

TAID versus CHAID.
Búsqueda de perfiles de mujeres trabajadoras en el servicio doméstico

Dpto. de Estadística
Universidad de Salamanca

Trabajo para optar al título de Máster en

Análisis Avanzado de Datos Multivariantes

por la Universidad de Salamanca.

Presenta:

Sergio Ramos Hernández

Salamanca

2015

Prefacio

Años antes de entrar en la carrera yo ya tenía más que claro lo que quería estudiar. Las matemáticas siempre me motivaron y me gustaron, por ello, quería estudiar Ciencias Exactas. En mi llegada a la carrera todo fue decepcionante, no era como me esperaba y decidí decantar la balanza por la estadística, aquella parte de los temarios de la E.S.O. que siempre quedábamos sin explicar pese a que inicialmente esta Ciencia tampoco me convencía. Ante esta decisión cursé en mi carrera las prácticas externas de cuarto curso en el Departamento de Estadística de la Universidad de Salamanca. Fue entonces cuando la Dra. Purificación Galindo me mostró una parte de la estadística que desconocía y me enseñó a amarla. No fue la única, allí conocí a mentes maravillosas que presentaban una Ciencia muy distinta a la que yo me esperaba y así fue como decidí dedicar mi vida a tal tarea.

Cuando mi directora me sugirió involucrarme en este tema pensé que me iba a costar más adaptarme al mismo ya que no era algo que dominase en exceso, pero nunca me arrepentiré de su sugerencia porque tras documentarme bien descubrí un tema apasionante en los métodos de segmentación, un tema que pretendo seguir investigando en el futuro. Ha sido un camino difícil, con ciertos obstáculos, pero, pese a todo, conseguí llegar al final pudiendo así presentar mi trabajo fin de máster para, con ello, cerrar este pequeño fragmento de mi vida y empezar con el siguiente.

Agradecimientos

A mi directora, la Dra. Purificación Galindo, quien me enseñó la belleza de la estadística, que las matemáticas podían tener un lado bonito, me aconsejó el tema en el que me he visto inmerso estos últimos meses y quien ha guiado mi camino.

A mi codirectora, la Dra. Carmen Patino, quien me ha sufrido en cada momento, me ha acompañado, ayudado y animado y a quien le debo haber conseguido llegar a este punto.

Al resto de docentes del máster, a algunos con más cariño, por haber contribuido a mi formación como estadista. Me quedo con vuestras clases, vuestras enseñanzas, vuestros consejos, vuestro apoyo…

A mi familia, a quienes les debo todo lo que soy. A mi hermano Pablo por haber estado siempre a mi lado y ser la persona que más me aguanta, por sus consejos y su sabiduría. A mi madre por su apoyo incondicional, sus constantes preguntas para informarse de cómo iba el proceso y por estar ahí. A mis pequeñas, Elena y Laura, porque no se imaginan de la fuerza que pueden llegar a dar con lo pequeñas que son y porque son lo más importante de mi vida.

A mis amigos, en especial a Mercedes y a Maite por haberme apoyado en todo en la vida.

A mis compañeros del máster, por haberme ayudado a construir este camino porque su apoyo se ha notado en más de una ocasión.

Venid hasta el borde, les dijo.

Tenemos miedo, podríamos caer.

Venid hasta el borde, les dijo.

Ellos fueron. Los empujó… y volaron.

(Guillaume Apollinaire)

Índice

Resumen 1

Introducción y Objetivos 2

Algoritmos para análisis de segmentación 6

1.1. Métodos AID 7

1.2. Algoritmo Chaid 8

1.2.1. Tipos de predictores 9

1.2.2. Criterios de Parada 10

1.2.3. Problemática del Algoritmo 11

1.2.4. Contribución al Algoritmo Chaid 13

1.2.5. Aplicaciones del Algoritmo Chaid 14

1.3. Algoritmos DÁVILA 17

1.3.1. Algoritmo 1 (DÁVILA 1) 18

1.3.2. Algoritmo 2 (DÁVILA 2) 20

1.4. Algoritmos DORADO 21

1.4.1. Algoritmos ADORADO 21

1.4.2. Algoritmos DDORADO 22

1.5. Algoritmo Taid 23

1.5.1. Análisis de clases latentes 23

1.5.2. Coeficiente de predictividad e índice de Catanova 28

1.5.3. Análisis no simétrico de correspondencias 32

1.5.4. Algoritmo TAID 35

Aplicación a datos reales 38

2.1. El sector de trabajadoras del servicio doméstico 39

2.2 Objetivos 42

2.3 Material y Métodos 43

2.4 Descripción de la muestra 43

2.5. Búsqueda de los perfiles de las mujeres empleadas de hogar, basada en información multivariante mediante el algoritmo CHAID 49

2.6. Búsqueda de los perfiles de las mujeres empleadas de hogar, basada en información multivariante mediante el algoritmo Taid 58

2.7. Estudio comparativo de los algoritmos Chaid y Taid 76

Conclusiones 80

Bibliografía 82

Anexo I 89

Resumen

El análisis de Cluster y la agrupación de individuos es un tema en auge en el campo de la estadística. En las últimas dos décadas, se han vuelto populares como alternativas a la regresión, al análisis discriminante y a otros procedimientos basados en los modelos algebraicos, los análisis de datos en árboles de clasificación. Una buena manera de hacer una agrupación según ciertas características de los datos son los métodos de segmentación. Por ello, el interés de estudiar en este trabajo ciertos modelos que consten de una o más variables respuestas. En esta investigación se ha llevado a cabo una revisión de los métodos de segmentación, centrándonos especialmente en el algoritmo Chaid, propuesto por Kass (1980) como una modificación a AID (Automatic Interaction Detection) para las variables definidas como dependientes e independientes y el Taid (Tau Automatic Interaction Detection), algoritmo de partición recursiva que genera un modelo de árbol ternario para la clasificación o segmentación de las observaciones, basado en el análisis de correspondencias no simétrico sobre tablas de contingencia.

Posteriormente se han aplicado sendos algoritmos a un conjunto de datos reales realizando un estudio comparativo entre ellos. La principal diferencia entre ambos reside en el hecho de que el Chaid cuenta sólo con una única variable respuesta mientras que el Taid presenta un conjunto de ellas.

Como consecuencia de la aprobación del Real Decreto-ley 29/2012, de 28 de diciembre, de mejora de gestión y protección social en el Sistema Especial para Empleados de Hogar y otras medidas de carácter económico y social (BOE de 31 de diciembre de 2012), se han introducido una serie de modificaciones en la configuración jurídica del Sistema Especial para Empleados de Hogar, con lo que se han convertido en un colectivo de especial interés en nuestro país. Por todo ello, nos ha parecido interesante estudiar los perfiles de estas trabajadoras mediante estas técnicas de análisis multivariante, las cuáles son idóneas para este tipo de análisis dada la naturaleza de los datos. El estudio se ha llevado a cabo en una muestra de 1170 trabajadoras en el sector del servicio doméstico pertenecientes a distintas provincias españolas.

Palabras clave: métodos segmentación, Chaid, Taid, árbol ternario, empleadas hogar.

Introducción y Objetivos

Introducción

A lo largo de la historia la estadística siempre ha intentado hacer agrupaciones óptimas de los individuos de determinadas muestras con el objetivo de obtener mejores conclusiones. Estas agrupaciones pueden hacerse de distintas maneras, una de ellas es creando árboles mediante análisis de segmentación.

El análisis de segmentación es una técnica que fragmenta las muestras utilizando un proceso secuencial descendente, que delimita grupos homogéneos según los criterios de una variable respuesta mediante combinaciones jerárquicas de una selección de otras variables propuestas.

Los métodos de segmentación presentan propuestas importantes para formar los grupos de individuos realizando clasificaciones en función de diversas características, como son las socio-demográficas y las de opinión o intereses. En la mayoría de las propuestas, la segmentación se realiza mediante una única variable respuesta y varias variables independientes a las que llamaremos predictores.

Estos métodos están presentes en muchas áreas de conocimiento como las Ciencias Sociales (consumidores, votantes…), Ciencias de las Salud (antecedentes médicos, tipos de alimentación…), Ciencias de la Comunicación (lectores de periódicos, televidentes de cadenas o programas…) y la Economía (segmentación de mercados…).

Podemos destacar varios beneficios del análisis de segmentación como que se puedan construir perfiles más precisos de los individuos o que nos permita hacer grupos para estudiar mejor un subgrupo poblacional o conseguir mejores pronósticos sobre el comportamiento de grandes grupos de datos.

En los últimos años se han hecho grandes avances de estos métodos que tienen como idea principal proponer un esquema de segmentación cuyo resultado provea lo que podemos considerar como la mejor segmentación.

La estructura de nuestro trabajo será la siguiente:

- En el capítulo I se hace una revisión bibliográfica exhaustiva de los métodos de análisis de segmentación. En primer lugar se analiza el algoritmo Chaid, por ser el algoritmo más utilizado y frecuente de los

métodos de segmentación. En su revisión se estudian también sus problemáticas: trabaja con tablas colapsadas, ante este problema se revisan nuevos métodos capaces de mejorar el algoritmo. Se procede a la revisión de los métodos Dávila y Dorado, capaces de resolver las problemáticas presentadas en el algoritmo Chaid mediante el estudio de las respectivas hipótesis de independencia condicionadas de las respectivas tablas trifactoriales. Estos algoritmos ya resuelven los problemas del Chaid pero no son válidos para conjuntos de datos que presenten un conjunto de varias variables de tipo respuesta, por ello la necesidad de la revisión de métodos que sí los consideren. Se introduce ahora el algoritmo Taid, algoritmo capaz de capturar la asimetría de las tablas de contingencia y, además, de trabajas con varias variables respuestas. Para la revisión de este último algoritmo fue requerido un estudio de la técnica de análisis de clases latentes y Bootstrap, el coeficiente de predictividad, el índice de Catanova (o C-statistic) y el análisis no simétrico de correspondencias.

- En el capítulo II, se hace una aplicación práctica de las técnicas CHAID y TAID, comparando ambos métodos. La muestra objeto de estudio son mujeres que trabajan en el servicio doméstico, empleadas de manera regular o irregular.

La aplicación práctica de este trabajo se ha realizado con varios programas estadísticos, informáticos o de lenguaje simbólico:

- IBM SPSS Statistics 22
- Microsoft Excel 2010
- Winmira 2001
- R i386 3.1.2

con un procesador AMD E1-2100 APU 1,00 GHz.

Objetivos

Búsqueda bibliográfica estudiando los artículos cuidadosamente para llevar a cabo una revisión y un estudio comparativo de los métodos de segmentación:

- Redacción del estado del arte en los métodos de segmentación.
- Estudio comparativo del algoritmo de Castro *vs* algoritmo CHAID.

Aplicación práctica de las alternativas estudiadas a datos reales (datos de mujeres trabajadoras a raíz de la reforma laboral de 2012) con el fin de la obtención de los perfiles de las empleadas del servicio doméstico.

Capítulo 1

Algoritmos para análisis de segmentación

1.1. Métodos AID

El análisis de segmentación se basa en técnicas estadísticas donde hay una o más variables respuesta y un conjunto de predictores cuyo objetivo es segmentar la población en grupos homogéneos que puedan describir las variables respuesta. Es una técnica de dependencia entre variables, el cual debe ser utilizado principalmente con una finalidad exploratoria, ya que su mecanismo consiste en la búsqueda de las mejores asociaciones de las variables independientes con las dependientes (Escobar, 1998)

Los predictores son variables de tipo cualitativo mientras que las variables respuesta pueden ser de tipo cualitativo o cuantitativo. El análisis de segmentación está basado en los métodos de detección automática de la interacción.

Los métodos “Automatic Interaction Detection” (AID), propuestos por Morgan y Sonquist (1963), tienen como objetivo detectar la interacción en un modelo predictivo mediante una segmentación de los individuos en grupos homogéneos, excluyentes entre sí, de modo que se pueda explicar la relación entre los predictores y la variable respuesta (en este caso solo una) sin quedar esta enmascarada por la interacción.

Estos métodos tratan con datos tipo regresión, una variable dependiente y un conjunto de predictores que son de tipo cualitativo. Tienen como característica particular el que su aplicación está prácticamente libre de los supuestos requeridos por los métodos basados en el modelo lineal. El objetivo básico de los métodos AID es agrupar los predictores con perfil similar para la variable dependiente.

El procedimiento general que utilizan los métodos AID es de tipo iterativo. Los grupos finales quedan formados después de un proceso por etapas, en cada una de las cuales se selecciona el mejor predictor. Debido a esta estrategia de particiones sucesivas de los datos, se conoce a estos métodos con el nombre de Análisis de Segmentación.

En función del tipo de variable respuesta del estudio hay diferentes tipos de métodos AID. Aquí describiremos brevemente el propuesto por Morgan y Sonquist (1963). En este método, sus autores propusieron un conjunto de datos con una variable dependiente cuantitativa y varios predictores de tipo dicotómico. Se basaron en divisiones binarias sucesivas utilizando como criterio la reducción de cuadrados de la suma no explicada.

El algoritmo es el siguiente: se realizan todas las divisiones binarias y se busca la que produzca una mayor reducción en la suma de cuadrados del residual. Si este

valor es al menos el 1% de la suma de cuadrados residual del total se particiona en dos segmentos, uno por categoría. Se repite el proceso en cada nodo hasta que la suma de cuadrados del residual del nodo sea menor al 2% de la suma total que entonces se detiene.

Del resto de propuestas sobre los métodos AID la más conocida es la propuesta por Kass (1980) denominada, Chaid.

1.2. Algoritmo Chaid

El algoritmo Chaid es considerado el método general de segmentación, por ello el resto de propuestas de los métodos AID podrían ser consideradas como casos particulares de esta.

En el algoritmo Chaid (Kass, 1980) se propone para una variable respuesta de tipo cualitativo y predictores cualitativos y se basa en el test chi-cuadrado para contrastar las independencias. El algoritmo Chaid es un algoritmo secuencial y su esquema es el siguiente:

Figura 1.1: Esquema del algoritmo Chaid

Este algoritmo tiene cuatro fases que procederemos a detallar:

Fase I

En esta etapa Kass propone realizar el test chi-cuadrado, cruzando la variable respuesta con cada predictor, y ver qué categorías tienen un perfil similar con respecto a la variable respuesta (que no son significativas). En ese caso, dichas categorías se agruparán.

Propone cruzar cada par de categorías y fusionar el par con mayor p-valor no significativo. Este proceso se repetiría hasta que no pudiesen agruparse más categorías.

Fase II

Finalizada la fase anterior, hay que seleccionar el mejor predictor. De entre los que se tiene se elige para segmentar aquel con un menor p-valor tras realizar el test chi-cuadrado. Entonces, el mejor predictor será aquel que discrimine mejor a los individuos según la variable respuesta.

Fase III

Fijado el p-valor si el predictor seleccionado es significativo segmentamos en tantas ramificaciones como categorías tenga este.

Fase IV

Finalizada la segmentación se realiza nuevamente el proceso completo en cada una de la ramificaciones hasta que los nodos sean terminales, es decir, que no haya predictores significativos para dicho nodo.

En este proceso hemos hablado de agrupar categorías y hemos fijado como criterio de parada la no presencia de predictores significativos pero esto conlleva problemas. En primer lugar no se pueden agrupar categorías sin seguir ciertas pautas, dependerá de los tipos de predictores que tengamos. En segundo lugar, fijar sólo la ausencia de predictores significativos como criterio de parada nos lleva a tener árboles muy poco ocupados, es decir, muchos nodos terminales con muy poco individuos en cada uno, por tanto, se han de introducir más criterios de parada.

1.2.1. Tipos de predictores

Como ya hemos dicho hay que prestar atención a los predictores antes de agrupar sus categorías. Veamos los tipos que hay.

Predictor Monótono

Es de tipo ordinal, es decir, las categorías siguen un orden establecido, de modo que sólo se podrán agrupar dos categorías contiguas. Como por ejemplo el nivel de estudios. Si esta variable tuviera como valores: “primarios”, “secundarios” y “universitarios”, el procedimiento permitiría la fusión de las categorías primera y segunda o segunda y tercera, y descartaría la posibilidad de formar un grupo compuesto por sujetos con estudios primarios y universitarios

Predictor Libre

Es aquel que sea de tipo nominal, o sea, no se puede establecer un criterio de orden en sus categorías. En este caso, sí tendrá sentido agrupar dos categorías aleatorias, sin necesidad de que sean contiguas, es decir, cualquier par de categorías puede ser agrupada. Una variable de este tipo es por ejemplo la variable situación laboral con los valores: “activo”, “parado”, e “inactivo”. La categoría “activo” podría formar grupo con “parados” y/o “inactivo”.

Predictor Flotante

Es aquel que tiene todas sus categorías, es una escala ordinal menos una que no se conoce su lugar en la escala ordinal y se conoce como categoría flotante. En este caso al igual que con un predictor monótono, solo se pueden agrupar dos categorías contiguas con excepción de la flotante que podrá quedar sola o unirse a cualquiera de los grupos ya formados. Por ejemplo consideraremos que la variable nivel de estudios, tuviera el valor “Ns/Nc”. generalmente el “no sabe, no contesta”, puede agregarse libremente a cualquiera de las categorías establecidas.

1.2.2. Criterios de Parada

Hemos comentado anteriormente la necesidad de incluir otros criterios de parada para finalizar el algoritmo, veamos aquí esos criterios, también llamados filtros de proceso:

Significación de Categoría

En el primer paso del algoritmo se agrupan las categorías con un perfil similar, para ello, se realiza el test chi-cuadrado y se compara el p-valor obtenido con esta significación, que queda prefijada antes de empezar el algoritmo.

Significación de Predictor

Esta significación es con la que comparamos el p-valor obtenido tras realizar el test en la fase de selección del mejor predictor. Al igual que en el caso anterior, también queja fijada previamente a comenzar el algoritmo.

Filtros de Asociación

Es el encargado de medir el nivel de asociación existente entre el predictor elegido con la variable dependiente, para ello, se calcula un coeficiente de asociación

entre la variable respuesta y el predictor elegido como mejor predictor, en el caso de que este coeficiente sea menor que el filtro prefijado se considera el predictor como el mejor para segmentar, de lo contrario se elimina este predictor en esta segmentación y se considera el siguiente. La mayor ventaja que tienen estos filtros frente a los de significación es que no presentan sensibilidad respecto al número de individuos del estudio.

Tamaño antes

Se establece este tamaño como el número necesario de individuos para poder seguir segmentando. De esta manera, si un nodo tiene menos individuos de los establecidos, se considera como nodo terminal.

Tamaño después

Este es el tamaño mínimo considerado para que un grupo pueda ser formado, de tal manera que si al segmentar uno de los subgrupos tienen un tamaño muy pequeño de individuos esta segmentación es descartada.

Estos dos últimos filtros (tamaños antes y después) tienen como objetivo evitar la formación de grupos muy pequeños o de grupos balanceados ya que de este modo se generalizaría una conclusión que no sería demasiado fiable.

Filtros de Nivel

Se determina el número mínimo y máximo de niveles en un paso previo a comenzar el algoritmo, evitando así que la ramificación presente solo un nivel, lo cual sería muy sencillo de interpretar pero demasiado simple y evitando también todo lo contrario, que la ramificación presente un amplio número de niveles ya que puede resultar muy compleja presentando grandes dificultades de interpretación.

1.2.3. Problemática del Algoritmo

Este algoritmo trabaja con contrastes de hipótesis sobre distribuciones marginales, esto lleva a un problema ya que nos podemos encontrar con casos en los que dos variables sean independientes de manera marginal pero no lo sean cuando se estudien junto a otras. Esto se conoce como Paradoja de Simpson.

Se denomina Paradoja de Simpson al cambio en el sentido de una asociación entre dos variables cuando se controla el efecto de una tercera variable. Describe la

desaparición de una asociación o comparación significativa de dos variables cuando los datos son desagregados por grupos. Recibe el nombre en honor de Edward Simpson, quien la describió en 1951, aunque fue previamente descrita por el estadístico británico G. Udny Yule a inicios de 1900.

Ejemplo: Se dispone de una muestra de 376 individuos de los que se sabe si tienen preocupaciones políticas o no y si votan o no distribuidos de la siguiente manera:

	Votante	No votante
Político	110	78
Apolítico	78	110

Tabla 1.1: Tabla de contingencia completa del ejemplo

Esta tabla de contingencia tiene un valor experimental $\chi^2 = 10{,}89$ y $p = 0{,}001$. En este caso concluiríamos que ambas variables son dependientes. Pero vamos a separarlos por edad. Por un lado los jubilados y por otro los que forman parte de la población activa, obteniendo los siguientes resultados:

Población activa	Votante	No votante
Político	10	70
Apolítico	8	100

Tabla 1.2: Tabla de contingencia para la población activa

Jubilados	Votante	No votante
Político	100	8
Apolítico	70	10

Tabla 1.3: Tabla de contingencia para los jubilados

En ambos casos tiene como valor experimental $\chi^2 = 1{,}38$ y $p = 0{,}24$, concluyendo entonces que sendas variables son independientes en ambos casos, es decir, se presenta aquí la paradoja de Simpson.

Otros ejemplos de la paradoja de Simpson son las siguientes:

- En el deporte, concretamente en la NBA, se estudia la presencia de esta paradoja en los playoffs (Ma & Ma, 2011).
- En psicología se da la paradoja en casos de neurociencia cognitiva y genética del comportamiento,a lo cual se ofrecen soluciones psicométricas (Kievit et al, 2013).

Para resolver este problema surgieron métodos capaces de estudiar las tablas trifactoriales (multifactoriales) sin necesidad de reducirlas a tablas bifactoriales y que revisaremos posteriormente.

1.2.4. Contribución al Algoritmo Chaid

A lo largo de estos puntos se ha desarrollado el algoritmo Chaid. Como hemos visto, para la selección de la mejor variable se calcula el p-valor referente al chi-cuadrado sobre la tabla de contraste. No obstante, en el caso de haberse realizado una agrupación de las categorías, la significación debería verse alterada de modo que se valore esa agrupación. En este punto revisaremos una manera de resolver este problema, propuesta por Ramírez (1995), que consistirá en una aproximación de Bonferroni.

Algunos procedimientos de Bonferroni, como los que vamos a ver a continuación, son de tipo secuencial o paso a paso y se basan en el ordenamiento de los p-valores $p_1 \geq p_2 \geq \cdots \geq p_n$ para ordenar también las respectivas hipótesis: $H_1 \geq H_2 \geq \cdots \geq H_n$. En este caso vamos a ver procedimientos ascendentes y descendentes. Denotaremos aquí a_i a los niveles críticos, que variarán en función del número de hipótesis de independencia que tengamos.

Procedimientos Descendentes

Se estudia en primer lugar el p-valor más pequeño. Si $p_n > a_1$ se aceptan todas las hipótesis y el proceso finaliza. De no ser así, se rechaza la hipótesis H_n y continúa el proceso eligiendo siempre el p-valor más pequeño hasta encontrar aquel que nos permita aceptar la hipótesis. La secuencia de valores críticos será:

$$a_1 = \frac{\alpha}{n} \leq a_2 = \frac{\alpha}{n-1} \leq \cdots \leq a_i = \frac{\alpha}{n-1+i} \leq \cdots \leq a_n = \alpha$$

Además, Holm (1979) demostró que este proceso tiene una significación α y mayor potencia que el procedimiento de Bonferroni.

Procedimiento Ascendente

Este procedimiento es inverso al otro. Empieza por el p-valor más grande p_1 y comprueba si $p_1 \leq a_1$ entonces se rechazan todas las hipótesis. De no ser así se pasa al siguiente y se repite el proceso que se ejecuta hasta que se rechace alguna hipótesis de independencia o hasta que se demuestre que se aceptan todas. En este caso, la secuencia de niveles críticos será:

$$a_1 = \alpha \geq a_2 = \frac{\alpha}{2} \geq \cdots \geq a_i = \frac{\alpha}{i} \geq \cdots \geq a_n = \frac{\alpha}{n}$$

Este procedimiento fue propuesto por Hochberg (1988) donde además demuestra que este y el de Holm merecen un control fuerte de la tasa de error global ya que la probabilidad de rechazar la hipótesis H_i es menor o igual que α, al margen del número de hipótesis que haya y de cuáles sean ciertas.

1.2.5. Aplicaciones del Algoritmo Chaid

El algoritmo Chaid es aplicado en numerosos campos de la ciencia. Así en el ámbito de las Ciencias Sociales, Economía, Turismo y Educación podemos destacar entre otros los siguientes:

Legohérel, Hsu y Daucé (2015) investigaron el uso de la variedad en la búsqueda de la segmentación de los viajeros internacionales, utilizando principalmente los criterios tradicionales de segmentación, como la nacionalidad, país de origen, otras características de disparo de los viajeros internacionales, y los comportamientos de los consumidores. Esto identificó varios segmentos con base en el comportamiento de búsqueda de variedad.

Por otro lado, Zübeyir (2015) identificó los niveles de burnout de los enfermeros que les impiden hacer su negocio más eficiente y examinan los efectos de diversas variables demográficas sobre el agotamiento. Su utilizo el Chaid determinando que las enfermeras experimentan bajo agotamiento.

Suh y Alhaery (2015) intentaron predecir la propensión de jugar a máquinas tragaperras y juegos de mesa, mediante la aplicación de un algoritmo de minería de datos a los clientes de casinos. Se utilizó el algoritmo Chaid para predecir la propensión cruzada juego. Los resultados de este estudio permitirán a los administradores del casino estimar los ingresos del juego incrementales y a gastar dinero en marketing de la manera más eficiente para aumentar al máximo los ingresos.

La clase trabajadora industrial ya no puede ser considerada como la hegemónica. Existe una nueva clase trabajadora de servicios que se ocupa principalmente en aquellos sectores intensivos en mano de obra, principalmente en el sector servicios de las economías desarrolladas. García (2014) trató de mostrar para España la existencia de esta nueva clase obrera.

La relación entre consumo turístico y desigualdad social en el marco de la sociedad española, actualizando la perspectiva sociológica sobre este tema a partir de una encuesta realizada en origen a población adulta fue estudiada por Rodríguez y Turégano (2014). Los resultados del estudio muestran que el consumo turístico en España se encuentra altamente polarizado, con un grupo relativamente reducido de personas que realizan muchos viajes (20%), aproximadamente un tercio de la población con una participación media en el turismo y más de un 40% de la población excluida del consumo turístico, ya sea por motivos económicos (28%) o por otros motivos (14%).

La contaminación microbiana de las aguas subterráneas utilizadas para el agua potable puede afectar a la salud pública y es de gran preocupación para las autoridades locales de agua y los proveedores de agua. Por ello Bichler, Neumaier y Hofmann (2014) propusieron una técnica de minería de datos no paramétricos para explorar la presencia de coliformes totales en una extracción de agua subterránea. El algoritmo Chaid reveló relaciones estadísticamente significativas entre la precipitación y la presencia de coliformes totales, tanto en un pozo de producción como en una zona de control.

También podemos ver un método para mejorar el proceso de evaluación del desempeño económico que se aplica para el caso de los 28 países de la Unión Europea utilizando los indicadores de crecimiento económico, la productividad laboral, la tasa de desempleo y los ingresos netos reales para el año 2013, propuesto por Popescu, Andreica y Micu (2014). Se clasificó a los países y en base a esta clasificación, se utilizó el algoritmo Chaid para el análisis.

Las decisiones de otorgamiento de crédito son cruciales en la administración de riesgos. Las instituciones financieras han desarrollado y usado modelos de *credit scoring* para estandarizar y automatizar las decisiones de crédito. En este trabajo se presenta una metodología general para construir un modelo sencillo de *credit scoring* enfocado justamente a esa población, la cual ha venido tomando una mayor importancia en el sector crediticio latinoamericano (Soldic-Aleksic, 2012).

Los resultados de la aplicación combinada de dos modelos de minería de datos, mapa de Kohonen y el modelo de árbol de decisión CHAID, para el problema de la agrupación en el ámbito de la comercialización fueron presentados por Soldic-Aleksic (2012). Se encontró que el modelo Chaid puede ser un buen intérprete visual de los resultados de clustering del mapa de Kohonen representando el mecanismo "estadístico" del proceso de agrupación.

La carne bovina es un alimento importante para la nutrición del ser humano y el buen funcionamiento del organismo. La Zona Metropolitana del Valle de México es el principal centro de comercialización y consumo de este alimento en el país. El objetivo del presente estudio fue realizar una caracterización del consumidor de carne bovina en la Zona Metropolitana del Valle de México para conocer el tipo de productos que demanda, asociando variables como nivel de ingresos, nivel de consumo y servicios integrados, entre otras. La metodología empleada fue el algoritmo Chaid entre otras concluyendo que los consumidores con ingresos medios y bajos (37.2 %) demandan cortes populares (bistec, molida y retazo) que los compran principalmente en los mercados públicos (Tellez, Mora, Martínez, García, & García, 2012).

Alvarado, Luyano y Téllez (2012) presentaron los resultados de una caracterización del consumidor de carne de pollo en el área metropolitana de Monterrey, efectuada mediante el uso del algoritmo Chaid, que permite segmentar las variables por estudiar y realizar pruebas de asociación entre ellas aplicando la distribución Chi-cuadrado. La característica primordial de los consumidores es que sus ingresos son menores de veinte mil pesos, y que demandan carne de seis a quince veces por mes.

El desempeño en la educación superior de la escuela secundaria en la India es un punto de inflexión en la vida académica de los estudiantes. En la presente investigación, se adoptó una metodología experimental para generar una base de datos y se construye a partir de una fuente primaria y secundaria. Se obtuvieron 772

expedientes que se clasificaron con el algoritmo Chaid. (Ramaswami & Bhaskaran, 2010).

1.3. Algoritmos DÁVILA

En una tabla bifactorial es clara cuál es la hipótesis de independencia, pero en una tabla trifactorial no solo hay una, sino que son posibles varias hipótesis de independencia. Para ello, llamaremos i, j y k a nuestras variables.

En primer lugar, tenemos la independencia completa, **i $\amalg$ j $\amalg$ k,** donde las tres variables son independientes entre sí.

En segundo lugar, tenemos la independencia múltiple, **i $\amalg$ (j, k), j $\amalg$ (i, k) o k $\amalg$ (i, j),** donde una de las variables es independiente de las otras dos pero sin embargo esas dos pueden estar relacionadas.

En tercer lugar, tenemos la independencia condicionada, **i $\amalg$ j / k, i $\amalg$ k / j o k $\amalg$ j / i** donde las dos primeras variables de cada caso son independientes entre sí para cada nivel de la tercera pero cualquiera de ella puede estar relacionada con la tercera e, incluso, estarlo las dos.

Denotamos ahora V como un conjunto de vértices con las variables dadas. Matemáticamente, un grafo G es un conjunto de vértices V y de bordes B que son pares de elementos tomados de V.

Denotaremos **V \ {i}** el conjunto que resta si al conjunto de vértices de V le quitamos el vértice i. Podemos introducir entonces las hipótesis de independencia condicionada como **i $\amalg$ j / V \ {ij}.**

Introduciremos ahora el término colapsabilidad. Una tabla trifactorial *i, j, k* será colapsable sobre el factor *i* si los odds ratio en la tabla marginal $p_{.jk}$ son idénticos a los odds ratio para cada fila de la tabla trifactorial de partida.

En términos de odds ratio diremos que una tabla trifactorial es colapsable sobre *i,* si para todo *i, j, j', k, k'* se verifica:

$$\frac{p_{.jk}p_{.j'k'}}{p_{.j'k}p_{.jk'}} = \frac{p_{ijk}p_{ij'k'}}{p_{ij'k}p_{ijk'}}$$

Si se cumple esta igualdad podemos colapsar sobre *i*, podremos analizar la tabla bifactorial *j, k* obteniendo conclusiones óptimas.

Ávila (1996) propone algoritmos basados en contrastes de hipótesis condicionadas con la forma: **i** $\amalg$ **j / V \ {j},** en las que es posible colapsar en *j* según las definiciones dadas anteriormente. Conforme a esto, los algoritmos que propone son descendentes, es decir, se parte de todas las variables (del árbol completo) y se van eliminando ramificaciones. El esquema de estos algoritmos es el siguiente:

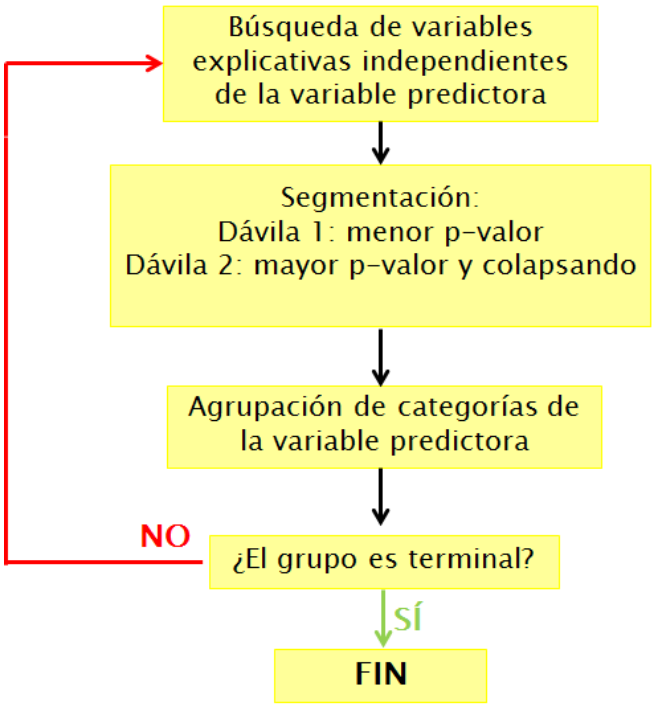

Figura 1.2: Esquema de los algoritmos Dávila

Veamos ahora estos algoritmos.

1.3.1. Algoritmo 1 (DÁVILA 1)

Se parte del árbol completo como ya hemos dicho y se sigue el siguiente procedimiento:

Fase I

En este paso se buscan aquellos predictores que sean independientes de la variable respuesta. Para ello, se buscan todas las variables $j \in V$ tales que **i** $\amalg$ **j / V **

{j}. A continuación se colapsan o eliminan las variables sin información significativa, que serán aquellas para las cuales la hipótesis citada se acepte.

Fase II

En esta fase se segmenta la población. El objetivo es colapsarla en segmentos. Se realizará un contraste de hipótesis condicionada, como ya hemos visto, y se seguirá un proceso similar al de Kass para seleccionar el mejor predictor. Se escoge aquel que de un menor p-valor en la hipótesis de independencia condicionada contrastada sobre la tabla colapsada.

Si el mejor predictor es la variable *j,* por ejemplo, entonces obtendremos *J* ramas, que se identifican con los subgrupos del predictor. Estas ramas serán analizadas por separado.

Fase III

En este paso se agrupan las categorías del predictor. La idea es simplificar el problema o árbol uniendo las ramas que tengan un perfil de respuesta similar.

Para realizar este proceso, se contrasta la hipótesis de independencia condicionada siguiente: **i** $\coprod$ **j (**j_1, j_2**) / V \ {j (**j_1, j_2**)}** donde la notación utilizada significa que se restringe la variable *j* a sus categorías (j_1, j_2) y se contrasta la hipótesis para cada pareja de categorías de la variable.

La fase finaliza agrupando aquellas categorías para las cuales el contraste sea no significativo. Hay que tener en cuenta que la agrupación de categorías no puede hacerse aleatoriamente sino siguiendo unas indicaciones que ya hemos visto anteriormente en este trabajo.

Fase IV

Este paso es el de la interacción. Se repite el algoritmo propuesto en cada una de las ramificaciones obtenidas hasta que todos los nodos sean terminales y no se pueda seguir colapsando y no haya variables para separar.

Con este paso finalizaría este algoritmo, veamos ahora el otro que Ávila propuso en su tesis.

1.3.2. Algoritmo 2 (DÁVILA 2)

Este algoritmo es muy similar al anterior. Se parte del árbol completo y tiene todos los pasos menos uno prácticamente iguales. Los únicos cambios con respecto a DÁVILA 1 son los siguientes:

Fase II

En este caso se realiza el contraste de hipótesis condicionada pero, contrariamente al algoritmo anterior y al de Kass, se escogerá, de entre los significativos, aquel con mayor p-valor de todos.

El siguiente paso será buscar los subgrupos en los que se pueda colapsar. Se realiza el contraste de hipótesis **i** $\coprod$ **j / (k, V \ {j, k})** con $k = 1, \cdots, K \; \forall \; \boldsymbol{k} \in \boldsymbol{V} \backslash \{\boldsymbol{j}\}$ y se colapsa en aquellos grupos en los que el contraste sea no significativo.

De entre todas las posibles se selecciona aquella con menor p-valor. Las categorías de esta variable serán las que nos den las distintas ramificaciones en las que segmentaremos a la población.

En el caso de no poder colapsar dentro de las categorías de ninguna de las variables se consideran las categorías de cada par de variables concatenadas y se vuelve a repetir el proceso.

Por otro lado, si la variable que se ha seleccionado (mayor p-valor) no tiene subgrupos colapsables se descarta, escogiendo para realizar el proceso la que tenga un p-valor inmediatamente menor a la seleccionada.

Fase III

Esta fase es igual a la del algoritmo anterior exceptuando que el proceso no se realiza únicamente en la variable elegida para segmentar si no en todos y cada uno de los predictores.

El algoritmo finaliza en el mismo momento que el anterior, cuando todos los nodos sean terminales por no poder seguir colapsando y por no tener variables para separar.

Basados también en contraste de hipótesis condicionada tenemos los algoritmos DORADO, propuestos por Dorado (Dorado, 1998) y son los que veremos a continuación.

1.4. Algoritmos DORADO

Los algoritmos Dorado (1998) se basan en criterios de entropía, por ello, nos vemos obligados a introducir este término y situarlo en el contexto.

Se entiende por entropía a la medida de desorden de un sistema o a la medida de incertidumbre que existe ante un conjunto de mensajes.

El *Índice de Shannon* define la **entropía** o incertidumbre media enfrentada como la ganancia media de información:

$$H(P_1, \cdots, P_n) = -\sum_i P_i \cdot \log P_i$$

Los algoritmos propuestos por Dorado son: algoritmos ascendente y descendente basados en criterios de entropía.

1.4.1. Algoritmos ADORADO

Llamado también Algoritmo Ascendente Basado en Criterios de Entropía.

Fase I

Para valorar el desorden existente en el conjunto calculamos la entropía de la variable independiente, denotada por *i*.

$$H(i) = -\sum_i P_i \cdot \log P_i$$

Podemos segmentar la población con los segmentos más homogéneos en relación a la respuesta.

$$H(i/j) = -\frac{1}{J}\sum_i \sum_j P(i = i/j) \cdot \log P(i = i/j)$$

$$I(i/j) = \widehat{H}(i) - \widehat{H}(i/j)$$

Teniendo en cuenta que sigue una distribución chi-cuadrado podemos ver si el cambio es significativo. Elegiremos para segmentar aquella variable que produzca un mayor descenso en la entropía significativo y segmentaremos en tantas categorías como tenga la variable escogida.

Fase II

En este paso se repite el proceso del primer paso en cada segmento hasta no encontrar ninguna variable predictora significativa que produzca un descenso en la entropía.

1.4.2. Algoritmos DDORADO

Este algoritmo, también llamado algoritmo descendente basado en contrastes de independencia condicionada y criterios de entropía, parte del árbol completo.

Fase I

En este paso se pretende averiguar si es posible eliminar variables del árbol, variables que no aporten información a la variable respuesta. Para realizar este paso se calculará la entropía de cada variable (como ya hemos visto en el método anterior) y posteriormente se eliminarán aquellas que no produzcan un incremento significativo de la entropía. En el caso de que todas las variables fuesen significativas, solo se eliminaría aquella con mayor p-valor y posteriormente se procedería a repetir el proceso.

Fase II

En el paso anterior se ha dicho que se eliminan todas la variables que sean significativas pero ahora miramos cuál causa la significación, es decir, se estudia variable a variable la diferencia de la entropía en cada una de sus categorías ya que el hecho de que el incremento de esta sea significativo puede deberse solo a una de ellas, es decir, si en una categoría el incremento de la entropía es significativo, este nos resultará significativo en toda la variable cuando realmente solo debemos eliminar la categoría de la que hablamos.

Hasta este punto hemos tratado los algoritmos Chaid y una contribución a este propuesta por Ramírez, con su correspondiente problemática, ya que como se ha estudiado, en el algoritmo Chaid se da la paradoja de Simpson. Para resolver este problema, se plantearon los algoritmos Dávila y Dorado que usando hipótesis de independencia condicionada conseguían estudiar las tablas trifactoriales sin necesidad de reducirlas a tablas bifactoriales Estos últimos corrigen la problemática de los anteriores pero siguen quedando escasos ya que rara vez se tiene un conjunto de datos en el cual sólo se considere una única variable respuesta, sino que es más frecuente observar un conjunto de datos con varias variables respuesta. Para resolver

este problema Castro propuso el algoritmo Taid, por esto se ha visto la necesidad de estudiarlo.

1.5. Algoritmo Taid

Este algoritmo es una de las partes centrales del trabajo. Fue propuesto por Castro (2005).

El algoritmo sugiere cambios con respecto a los anteriores, entre ellos que para segmentar se utiliza un análisis no simétrico de correspondencias y se basa en las ideas de Siciliano y Mola (1997) (que veremos posteriormente) con el fin de conseguir árboles ternarios. Pero la aportación fundamental es, sin duda, el hecho de poder considerar varias variables respuestas en lugar de tener sólo una como en lo algoritmos estudiados hasta entonces.

Para introducir este algoritmo es necesario conocer previamente ciertos análisis estadísticos. En las siguientes páginas revisaremos el análisis de clases latentes, formularemos el coeficiente de predictividad y el índice de Catanova y estudiaremos el análisis de correspondencia no simétrico. Tras esto procederemos a exponer el algoritmo de Castro (2005).

1.5.1. Análisis de clases latentes

El LCA se basa en el concepto de probabilidad y recurre a los datos observados para estimar los parámetros del modelo: la probabilidad de cada clase latente, cuya suma debe ser igual a 1 (tamaño); y las probabilidades de respuesta condicional, lo cual representa la probabilidad de una respuesta particular en una variable observada condicionada por la pertenencia a una clase latente determinada.

En un modelo de clases latentes se parte de una matriz que contiene los resultados observados de *p* variables categóricas que llamaremos variables manifiestas y serán denotadas como X_j, conformando un vector columna de *p* componentes $X' = (X_1, \dots, X_p)$ sobre una muestra total de *n* individuos:

$$\mathbb{X} = \begin{pmatrix} x_{11}\, x_{12} \cdots x_{1p} \\ x_{21}\, x_{22} \cdots x_{2p} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ x_{n1}\, x_{n2} \cdots x_{np} \end{pmatrix}$$

Denotamos x_{ij} como la observación de un individuo sobre una variable y uniendo todas las observaciones se forma el vector x_i llamado también patrón de

respuesta. La variable latente se representará como *Y* con *T* categorías latentes. Las *p* variables manifiestas se consideran como indicadoras de la variable *Y*.

El vector $x' = (x_1, \dots, x_p)$ denota un determinado patrón de respuesta en el cual cada una de las x_j toman diferentes valores dependiendo de las categorías de la correspondiente variable manifiesta. Estas variables conforman una tabla de contingencia múltiple con $\prod_{j=1}^{p} I_j$ patrones de respuesta, tal que cada X_j contiene categorías I_j.

La variable latente cumple el principio de independencia local, es decir, la relación entre las variables indicadoras queda explicada por la asociación de cada individuo con una clase latente concreta, teniendo presente que cada una de ellas tendrá probabilidades condicionadas de respuesta a las variables manifiestas, diferentes a las probabilidades condicionadas asociadas a otra clase latente distinta y que los individuos que pertenecen a la misma clase latente tendrán la misma probabilidad de responder a las variables manifiestas en cualquier combinación de categorías de las mismas. Este hecho sirve para diferenciar a los individuos pertenecientes a diferentes grupos y poder caracterizar tanto la variable latente como las clases latentes.

Así, a partir de este principio la densidad condicionada de que un individuo con patrón de respuesta *x* pertenezca a una clase latente se formula como:

$$\pi_{\boldsymbol{X}/Y(c)}(\boldsymbol{x}) = \prod_{j=1}^{p} \pi_{x_j/Y(c)}(x_j)$$

dónde $\pi_{x_j/Y(c)}(x_j) = P(X_j = x_j/Y = c)$, $x_j = 1, \dots, I_j$ y $c = 1, \dots, T$

La distribución conjunta de X e Y está dada por:

$$\pi_{\boldsymbol{X},Y}(\boldsymbol{x}, c) = \prod_{j=1}^{p} \pi_Y(c)\pi_{\boldsymbol{X}/Y(c)}(\boldsymbol{x})$$

dónde $\pi_Y(c) = P(Y = c)$ representa la proporción de elementos que se encuentran en la clase latente *c*, es decir, la probabilidad *a* priori.

Utilizando las expresiones anteriores, el modelo de clases latentes se expresa como:

$$\pi_{\boldsymbol{X}}(\boldsymbol{x}) = \sum_{c=1}^{T} \pi_Y(c) \prod_{j=1}^{p} \pi_{x_j/Y(c)}(x_j)$$

dónde $\pi_{x_j/Y(c)}(x_j)$ representa la probabilidad de respuesta condicionada de cada una de las variables manifiestas dentro de la clase latente c, para $c = 1, \dots, T$, $x_j = 1, \dots, I_j$ y $j = 1, \dots, p$.

Las probabilidades de clase $\pi_Y(c)$ y las probabilidades condicionadas $\prod_{j=1}^{p} \pi_{x_j/Y(c)}(x_j)$ están sujetas a las siguientes restricciones:

$$\sum_{c=1}^{T} \pi_Y(c) = 1$$

$$\sum_{x_j=1}^{I_j} \pi_{x_j/Y(c)}(x_j) = 1$$

El número de clases latentes de la variable equivale al número de tipologías definidas por el modelo para los valores de las variables observadas en la tabla de contingencia. El tamaño relativo de cada una de ellas proporciona información significativa para la interpretación de las probabilidades de estas, indicando el tipo de distribución de la población de elementos en las diferentes clases.

Consideramos que las variables X_j son de tipo dicotómico para simplificar el estudio. El principio de independencia local supone que las variables manifiestas son estadísticamente independientes para los individuos que tengan la misma posición en la variable latente. La probabilidad condicionada de observar un patrón de respuesta *x* podrá expresarse como el producto de las probabilidades de respuesta condicionadas para cada una de las diferentes variables manifiestas, esto es:

$$\pi_{\boldsymbol{X}_j/Y(c)}(\boldsymbol{x}) = \pi_{jx}^{x_j}(1 - \pi_{jc})^{1-x_j}$$

dónde π_{jc} es la probabilidad condicional de obtener una respuesta positiva en la variable X_j para un individuo de la clase latente c. Así, el modelo de clases latentes lo podemos reescribir como:

$$\pi_{\boldsymbol{X}}(\boldsymbol{x}) = \sum_{c=1}^{T} \pi_Y(c) \prod_{j=1}^{p} \pi_{jx}^{x_j}(1 - \pi_{jc})^{1-x_j}$$

Para calcular las probabilidades de clase, conjuntas y condicionales, se utilizan procedimientos iterativos basados en estimaciones de máxima verosimilitud. Los más conocidos son el algoritmo de Newton-Raphson (Haberman, 1979) y el algoritmo Esperanza-Maximización (Demspter, Laird, & Rubin, 1977).

El análisis de clases latentes es una técnica muy utilizada y se pueden encontrar desarrollada completamente en numerosos artículos y tesis (Sepúlveda, 2003).

1.5.1.1. Clases latentes con Bootstrap

Las tablas poco ocupadas incorporan problemas al modelo de clases latentes. Para superar estos inconvenientes se puede recurrir a realizar un remuestreo Boostrap. El método Bootstrap, propuesto por Efron (1979), trata de aproximar la distribución desconocida F_θ de un estadístico de contraste θ, donde $\hat{\theta}$ es una función de los datos observados $(x_1, \dots, x_n)$ y de un vector de parámetros φ de modo que $\hat{\theta} = (x_1, \dots, x_n, \varphi)$.

En el caso del análisis de clases latentes se considera θ como un estadístico de bondad de ajuste. La distribución empírica $F_{\hat{\theta}^*}$ de este estadístico es aproximada simulando datos adicionales $(x_{1b}^*, \dots, x_{nb}^*)$ y un conjunto de resultados independientes $\hat{\theta}_b^*$ para $b = 1, \dots, B$ donde B es el número de remuestras Bootstrap y x_{1b}^* es un patrón de respuesta. Existen dos tipos de remuestreos Bootstrap: el paramétrico y el no paramétrico (Araya, 2010).

Bootstrap No Paramétrico:

Consideramos $(x_1, \dots, x_n)$ una muestra aleatoria de tamaño n de patrones de respuesta. Cada patrón x_h está formado por las respuestas obtenidas de las variables manifiestas binarias.

Consideremos el estadístico θ con distribución de probabilidad desconocida estimado por $\hat{\theta}$. El método Bootstrap no paramétrico remuestrea la población definida por $F_{\hat{\theta}}$ para estimar la distribución muestral de $\hat{\theta}$. Además a cada patrón $(x_1, \dots, x_n)$ se le asigna una probabilidad de $1/n$. Así, se consigue una remuestra $\{x_1, \dots, x_n\}$ tomada de la muestra original con reemplazamiento.

El estadístico $\hat{\theta}_b^*$ se calcula como $\hat{\theta}$ pero con la b-ésima remuestra Bootstrap. Dada una colección de B remuestras Bootstrap independiente se obtienen los estadísticos $(\hat{\theta}_1^*, \dots, \hat{\theta}_B^*)$ y se define $\hat{\theta}$ construyendo un histograma considerando las B réplicas.

Bootstrap Paramétrico

Este método también conocido como Monte Carlo, trata de simular un conjunto de datos adicionales mediante un modelo de clases latentes hipotetizado como el apropiado para los datos de la muestra. Se calculan los parámetros de la muestra, se genera un nuevo conjunto $(x_{1b}^*, \dots, x_{nb}^*)$ para encontrar la distribución empírica del estadístico de bondad de ajuste. Este proceso se repite *B* veces o hasta conseguir un criterio de precisión para que se cumpla la aproximación. El método de Monte Carlo se puede dividir en seis fases:

Fase I

Ajustar el modelo de clases latentes consiguiendo las estimaciones de los parámetros y de los estadísticos de bondad de ajuste.

Fase II

Utilizando los resultados obtenidos en la primera etapa, calcular el número de elementos de cada clase latente.

Fase III

Generar los datos para las clases latentes y los patrones de respuesta de cada una. Este proceso se realizará usando por orden los valores de la primera replica del Bootstrap simulado.

Fase IV

Repetir la tercera fase con las *B* réplicas del Bootstrap paramétrico.

Fase V

Ajustar el modelo hipotético a las *B* muestras. Los valores replicados de los estadísticos obtenidos en el proceso Bootstrap estiaman una proporción de la verdadera distribución del estadístico.

Fase VI

Comparar los estadísticos de bondad de ajuste de la muestra original para el $(1-\alpha)B$-ésimo percentil con los del ajuste Bootstrap ordenadas de mayor a menor de modo que si el estadístico de bondad de ajuste es mayor que el $(1-\alpha)B$-ésimo percentil rechazar el modelo y en caso contrario no hacerlo.

Pese a su reciente incorporación al análisis de clases latentes, el análisis de clases latentes con remuestreo Bootstrap ya está en auge en la comunidad científica y podemos verlo aplicado en diferentes casos:

- Salud adolescente y reacciones a diferentes axiomas clínicos y/o educacionales (Lanza & Rhoades, 2013).
- Comparación del método de Monte Carlo con otros (Oberski, van Kollenburg, & Vermunt, 2013).

1.5.2. Coeficiente de predictividad e índice de Catanova

El coeficiente de predictividad τ fue introducido originariamente por Goodman y Kruskal (1954) para una matriz de probabilidad para medir el incremento relativo de la probabilidad de predecir correctamente la variable fila conociendo el nivel de la variable columna. Después de ellos se han realizado varias versiones de este coeficiente, sin embargo nosotros vamos a centrarnos en una.

Light y Margolin (1971), lo utilizaron para una muestra con el fin de analizar la heterogeneidad o la variabilidad de datos categóricos. Desarrollaron un procedimiento de análisis de varianza de dos vías para tablas de contingencia al que llamaron índice de Catanova. Esta técnica se extendió para permitir un análisis de tablas de contingencia multidimensionales por Anderson y Landis (1980). Estos autores mostraron que la medida de la variación debida a Gini (1912) podría ser utilizado para desarrollar una medida de la "variación explicada" para una variable respuesta Y que es atribuible a un solo factor X. Propusieron también una medida R^2 de asociación basado en un análisis de varianza para los datos categóricos (D'Ambra, Beh, & Amenta, 2005).

Se considera una tabla de contingencia de dos vías IxJ donde I y J son el número de categorías de las variables Y y X respectivamente. Se denota por f_{ij} a la

frecuencia observada (es decir, número de los sujetos que pertenecen conjuntamente a la categoría i-ésima de la variable I y a la categoría j-ésima de la variable X).

Por otra parte, se considera $p_{ij} = \frac{f_{ij}}{n}$ la probabilidad relativa de cada casilla. En este caso *n* es el tamaño muestral. Además, tenemos los siguientes totales:

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

dónde $p_{i.}$ son las probabilidades totales por filas y $f_{i.}$ el total marginal de cada fila. De modo análogo tenemos los valores columnas:

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

Usando ahora la notación del análisis de la varianza (ANOVA) la suma total de cuadrados TSS (Total Sum of Squares) es equivalente a lo siguiente:

$$TSS = \frac{n}{2} - \frac{1}{2n}\sum_{i=1}^{I} f_{i.}^2 = \frac{n}{2}(TSS^*)$$

La segunda parte de la igualdad es simplemente una manera de denotarlo.

Además, la suma de cuadrados total sabemos que podemos descomponerla como suma de cuadrados, la suma de cuadrado dentro (WSS) y la suma de cuadrados entre grupos (BSS), que presentan las siguientes expresiones:

$$WSS = \sum_{j=1}^{J}\left(\frac{f_{.j}}{2} - \frac{1}{2f_{.j}}\sum_{i=1}^{I} f_{ij}^2\right) = \frac{n}{2} - \frac{1}{2}\sum_{j=1}^{J}\frac{1}{f_{.j}}\sum_{i=1}^{I} f_{ij}^2$$

y

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

respectivamente.

Además, la descomposición de suma de cuadrados totales en varios componentes nos permite conocer la variabilidad explicada de variable respuesta que

se puede atribuir a un solo factor de variación en un marco categórico. Esta medida viene dada por R^2 de la siguiente manera:

$$R^2 = \frac{BSS}{TSS} = \frac{\sum_{j=1}^{J} \frac{1}{f_{.j}} \sum_{i=1}^{I} f_{ij}^2 - \frac{1}{n} \sum_{i=1}^{I} f_{i.}^2}{n - \frac{1}{n} \sum_{i=1}^{I} f_{i.}^2}$$

En este caso, la medida R^2 no solo sirve para interpretar la variabilidad explicada de la variable respuesta al estimador de máxima verosimilitud de la medida de asociación τ de Goodman y Kruskal, es decir:

$$\tau = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{p_{ij}^2}{p_{.j}} - \sum_{i=1}^{I} p_{i.}^2}{1 - \sum_{i=1}^{I} p_{i.}^2} = \frac{BSS^*}{TSS^*}$$

dónde el denominador corresponde con el coeficiente de heterogeneidad de GINI y explica la medida de heterogeneidad de las categorías de la variable respuesta. Por otro lado, el numerador se corresponde con la heterogeneidad explicada por el poder predictivo. Este coeficiente varía entre 0 (sin poder predictivo) y 1 (máximo poder predictivo):

- Si $\tau = 0$ existe independencia para cada celda, es decir: $\frac{p_{ij}}{p_{.j}} = p_{i.}$
- Si $\tau = 1$ para cada *j* existe una *i* tal que $p_{ij} = p_{.j}$

Light y Margolin (1974) descubrieron que el método Catanova aproxima mejor el tamaño del test que el método de chi-cuadrado. El test chi-cuadrado requiere una restricción en la frecuencia esperada que en el caso del método de Catanova no existe. El enfoque del test de Catanova fue extendido al análisis de categorías ordinales por Anderson y Landis con una variable respuesta ordinal. Mediante el uso de un sistema de puntuación natural para ordenar categorías, Anderson y Landis (1982) propusieron, en un marco de ANOVA, la medida global de la proporción de la variabilidad de una variable Y explicada por otra X. Para una tabla de contingencia de dos vías de esta medida es:

$$R^2 = \frac{\sum_{j=1}^{J} \frac{\bar{a}_j f_{.j}}{n} - \mu_I^2}{\sum_{i=1}^{I} \frac{s_I(i)^2 f_{i.}}{n} - \mu_I^2}$$

dónde $s_I(i)$ es el valor asignado a la categoría i-ésima de la variable X.

$\mu_I = \frac{\sum_{i=1}^{I} s_I(i) f_{i.}}{n}$ es la puntuación media de la distribución marginal de la variable Y, y para el grupo j-ésimo $\bar{a}_j = \frac{\sum_{i=1}^{I} s_I(i) f_{ij}}{f_{.j}}$. Además, Anderson y Landis demostraron que nR^2 sigue una distribución chi-cuadrado con (J-1) grados de libertad.

Sea $\pi_{ij} = \frac{p_{ij}}{p_{.j}} - p_{i.}$ es fácil demostrar bajo estas circunstancias que la suma de cuadrados entre grupos (BSS), definida anteriormente, puede escribirse como:

$$BSS^* = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{.j} \left(\frac{p_{ij}}{p_{.j}} - p_{i.} \right)^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{.j} \, \pi_{ij}^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{1}{p_{.j}} \left(p_{ij} - p_{i.} p_{.j} \right)^2.$$

La hipótesis de independencia más frecuente es que toda categoría j-ésima tiene la misma estructura de probabilidad multinomial. Al imponer la hipótesis de independencia $H_0: p_{ij} = p_{i.}$ para todas las categorías *I* y *J,* Light y Margolin demostraron que el índice de Catanova se podía calcular como:

$$C = (n-1)(I-1)\tau = (n-1)(I-1)\frac{BSS^*}{TSS^*} \sim \chi^2_{(I-1)(J-1)}$$

que sigue una distribución chi-cuadrado con $(I-1)(J-1)$ grados de libertad.

En el ANOVA clásico, las sumas de cuadrados dentro y entre son independientes y por tanto la suma de cuadrados entre y la total no lo son. En este caso, Light y Margolin afirman que ocurre lo contrario. Bajo la hipótesis de independencia:

- $\frac{TSS}{n-1}$ se puede considerar como una constante que verifica que: $\frac{TSS}{n-1} = \frac{1}{2}\left(1 - \sum_p p_i^2\right)$
- TSS y BSS pueden ser considerados asintóticamente independientes a lo largo de $f_{.j}$

Esta técnica estadística se ha utilizado con frecuencia en la actualidad:

- Para obtener modelos de asociación parcial correspondientes a la relación de hipótesis condicional de un par de variables teniendo en cuenta el resto de variables del modelo (Olmus & Erbas, 2014).
- Evaluación del desempeño innovador de las empresas manufactureras en campaña (D'Ambra & Crisci, 2014)
- Estimación de la edad de los restos esqueléticos (Mays, 2014)

- Evaluación de las diferencias de los salarios del profesorado de distintas instituciones de educación superior según su género y su cargo (Bell, Meier, & Guyot, 2013)
- Evaluación de la tendencia migratoria vigente para Mombasa (Kenya) según su medio ambiente, entorno socio-económico… (Kiranga, 2013)
- Evaluar las emisiones de óxidos de nitrógeno mediante un análisis de la conducción de un vehículo (Camminatiello, D'Ambra, & Ragione, 2011).
- Estudio de las categorías que llevan al análisis no simétrico de correspondencias en una tabla de contingencia (A. D'Ambra & Crisci, 2014).
- Comparación del índice de Catanova con otros índices (D'Ambra, D'Ambra, & Pasquale, 2010).
- Propuesta de un nuevo análisis de correspondencias no simétrico basado en el índice de Catanova (Takane & Jung, 2009),
- El índice de Catanova en el modelo Anova de medidas repetidas (Singh, 2004).
- Discusión de las asimetrías de las especies (Chessel & Gimaret, 1999).

1.5.3. Análisis no simétrico de correspondencias

El análisis de correspondencias se aplica a varias variables categóricas (no necesariamente dicotómicas) y es utilizado para detectar la relación entre ellas. Generalmente estas técnicas utilizan como gráficos mapas factoriales simétricos y asimétricos con coordenadas estándar y coordenadas principales. Sin embargo, en los estudios más actuales se ha recomendado utilizar como gráfico el Biplot, que en los últimos años ha cobrado fuerza en estos estudios.

En el análisis de correspondencias cuando se estudian dos variables categóricas sin necesidad de que quede impuesto cuál de ellas es la variable dependiente y cuál es la independiente decimos que estamos ante un análisis de correspondencias simétrico.

En el caso contrario, cuando existe una variable dependiente y la otra independiente (hay una relación de dependencia entre las variables), entonces decimos que estamos ante un análisis de correspondencias no simétrico o eso es lo más conveniente según Beh (2008).

Esta segunda técnica será la que utilizaremos nosotros por estar en el caso de tener una variable latente como variable respuesta y varios predictores como variables independientes.

El análisis de correspondencias no simétrico fue propuesto por Lauro y D'Ambra (1984) como una variación del análisis de correspondencias simétrico propuesto por Benzecri (1973).

El análisis de correspondencias simétrico utiliza el índice de asociación Φ^2, que está basado en el test chi-cuadrado y supone una relación simétrica de las variables. Lauro y D'Ambra propusieron el uso de un nuevo estadístico, el coeficiente de predictividad, coeficiente que hemos estudiado en el apartado anterior.

Características del análisis de correspondencias no simétrico:

1. En este análisis la variable respuesta se expone por filas mientras que el predictor se coloca en columnas. La variable respuesta es considerada la variable dependiente y el predictor como variable independiente.

2. Como hemos dicho, el estadístico usado no es el chi-cuadrado, sino que se basa en la τ de Goodman-Kruskal vista anteriormente. Además podemos afirmar lo siguiente:

   - Si se verifica que $\tau = 0$, entonces la variable respuesta no se presenta alterada por el predictor o la variable independiente.
   - Si se verifica que $\tau = 1$, podemos decir que la variable independiente del estudio explica completamente a la variable dependiente.

3. En el análisis no simétrico de correspondencias si alguna categoría de la variable respuesta se representa cerca de una categoría de la variable independiente podemos decir que presentan una fuerte relación. De modo análogo, si sendas categorías no se representan de manera cercana podemos decir que la casilla de estudio tiene un porcentaje pequeño de individuos sobre el total.

La distancia de un punto j al origen será:

$$d^2(j, O) = \sum_{i=1}^{I} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{i.}}{f_{..}} \right)^2$$

Diremos entonces que un punto cualquiera j estará más alejado del origen cuanto mayor sea la desviación a la hipótesis de independencia. Recordemos la hipótesis de independencia del análisis simétrico o no simétrico (es la misma): $H_0: \frac{p_{ij}}{p_{.j}} - p_{i.} = 0$.

De modo análogo tenemos la distancia de un punto *i* al origen:

$$d^2(i, O) = \sum_{j=1}^{J} \left(\frac{f_{.j}}{f_{..}}\right)\left(\frac{f_{ij}}{f_{.j}} - \frac{f_{i.}}{f_{..}}\right)^2$$

Como vemos, esta distancia se ve afectada por el peso de las columnas, por ello, las columnas con mayor peso son las que hacen que los puntos fila se alejen más del origen.

4. Finalmente, destacamos sobre ventaja del análisis de correspondencias no simétrico con respecto al simétrico que la inercia total en el simétrico es sensible a proporciones marginales pequeñas de la variable respuesta. Contrariamente, esto no ocurre en el caso del análisis no simétrico.

Podemos ahora estudiar la dispersión total en torno al origen, que será lo mismo que la inercia de la nube de puntos de los elementos *j*. Esta inercia viene dada por:

$$In(J) = \sum_i \sum_j \frac{f_{.j}}{f_{..}}\left(\frac{f_{ij}}{f_{.j}} - \frac{f_{i.}}{f_{..}}\right)^2$$

En la inercia influye el peso de las columnas. Las columnas con mayor peso serán las que más contribuyan a la inercia y aquellas con pesos pequeños podrían eliminarse.

Por otro lado, podemos ver la nube de puntos de los elementos *i.*

$$In(I) = \sum_i \frac{f_{i.}}{f_{..}} d^2(i, O) = In(J)$$

Como vemos, la diferencia entre el análisis de correspondencias simétrico y no simétrico reside en la métrica utilizada. El no simétrico se basa en la métrica euclídea mientras que el simétrico se basa en la métrica chi-cuadrado (Salvo, 2002).

Hasta este punto hemos estado revisando las técnicas estadísticas que utiliza el algoritmo Taid. Ahora podemos proceder a introducir el algoritmo.

1.5.4. Algoritmo TAID

El algoritmo sigue el siguiente esquema:

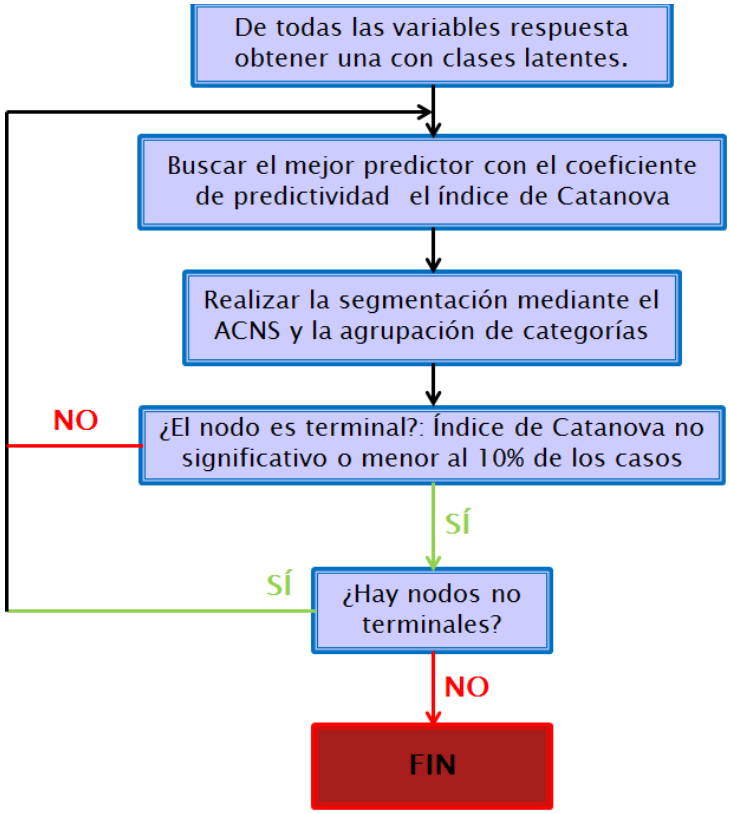

Figura 1.3: Esquema del algoritmo Taid

Fases del proceso del algoritmo:

Fase I

Como ya hemos dicho, este algoritmo presenta varias variables respuestas. En este paso se definirá una variable latente capaz de recoger el carácter multivariante del conjunto de variables que tenemos, es decir, se ha de realizar un modelo de clases latentes con todas y cada una de las variables respuestas a fin de encontrar una única variable respuesta con varias clases latentes que contenga la información relativa a un alto porcentaje de las que se han utilizado para su construcción.

Realizado este paso ya se puede trabajar con el conjunto de datos de manera normal, puesto que la variable latente que hemos creado actúa como variable respuesta y sus clases como categorías y estaríamos nuevamente en el caso de tener una variable respuesta univariante.

Fase II

En esta fase hemos de buscar el mejor predictor para segmentar. Se podría realizar el proceso para segmentarlo estudiando la tabla de contingencia, usando el estadístico chi-cuadrado pero tendríamos el mismo problema que en el Chaid que, como ya hemos dicho, trata a sendas variables de manera simétrica. Este algoritmo utiliza para segmentar el cálculo del coeficiente de predictividad (visto en puntos anteriores) y el cálculo del índice de Catanova.

En primer lugar se calcula el coeficiente de predictividad, a continuación se halla el índice de Catanova y finalmente se estudia si este último es significativo. De entro todos los predictores significativos se escogerá para segmentar aquel con un mayor coeficiente de predictividad.

Fase III

En los algoritmos más clásicos hemos visto que se segmenta conforme al número de categorías que tenía el predictor elegido, una ramificación por categoría. Pero este algoritmo se basa en la idea de Siciliano y Mola (1997) de construir árboles ternarios. Para hacer esta construcción se basó en un análisis no simétrico de correspondencias entre la variable respuesta y el predictor seleccionado ya que el poder predictivo puede ser representado en el plano principal del análisis de correspondencias no simétrico. Se clasificarán las categorías siguiendo las ideas de Siciliano y Mola como sigue:

$$\begin{cases} \varphi_{j1} \geq 1 \text{ categorías fuerte por la derecha} \\ |\varphi_{j1}| < 1 \text{ categorías débiles} \\ \varphi_{j1} \leq -1 \text{ categorías fuerte por la izquierda} \end{cases}$$

dónde φ_{j1} es la coordenada sobre el primer eje factorial del análisis de correspondencias no simétrico. Como vemos se crearán tres segmentos: uno con alto poder predictivo positivo, otro con alto poder predictivo negativo y otro con las que tienen un bajo poder predictivo y que necesitarán de otros predictores para explicar la variable respuesta al no poder hacerlo por sí solos.

Fase IV

En este paso se recopilan los resultados obtenidos anteriormente, el mejor predictor y la clasificación de las categorías y se realiza la segmentación.

Fase V

En esta fase se buscan los segmentos terminales. Para ello hay que imponer unos criterios de parada. En este caso se consideró como nodo terminal aquel que no tuviera más predictores significativos, es decir, que el p-valor asociado al índice de Catanova de los predictores restantes fuesen mayores que 0,05 o aquel nodo cuyo tamaño fuese inferior al 10% de la muestra total.

La idea de esta fase es, pues, repetir las fases de la dos a la cuatro en todos los nodos que no sean terminales.

Capítulo 2

Aplicación a datos reales

2.1. El sector de trabajadoras del servicio doméstico

En España, a principios del siglo XX, el menor nivel de industrialización y de desarrollo económico, en comparación con el entorno occidental, influyó en la participación laboral de la mujer. Dentro del sector terciario, el servicio doméstico constituía una de las actividades laborales tradicionalmente más desarrolladas por las mujeres, e indudablemente era la más importante desde un punto de vista cuantitativo, pues representaba el 72,6% del total de trabajadoras del sector servicios, sobrepasando el conjunto de empleos de la industria (González, 1982). No será hasta comienzos del siglo XX cuando se cambie la concepción de este servicio y comience a verse como una actividad laboral, sobre todo esto se realizará a través de políticas llevadas a cabo en las distintas épocas de democracia de nuestro país. Fue reconocido como actividad laboral por vez primera a través de la Ley de Contratos de Trabajo, aprobada en 1931 durante la Segunda República. Con la llegada del Franquismo y la creación del Montepío Nacional del Servicio Doméstico en 1959 esta actividad será incluida en él a través de disposiciones sociales. Aun así no es hasta 1985 cuando se establece esta situación como laboral a todos los efectos pues con anterioridad solo se encontraba contemplada en el Código Civil demostrando así las características atípicas que el legislador encontraba en ella.

El cambio económico sufrido en el siglo XX, que estableció una economía de mercado desplazando a la economía agraria dominante hasta el momento, no logró que las mujeres dispusieran de puestos concretos para ellas provocando una emigración de estas a las ciudades en busca de un puesto de trabajo y será el trabajo doméstico en el que estas logren una mejor y mayor incorporación, logrando un sector mayoritariamente copado por mujeres.

Los trabajadores comprendidos dentro del campo de aplicación del Régimen Especial de Empleados de Hogar, son los que se dediquen a servicios exclusivamente domésticos para uno o varios cabezas de familia, siempre que estos servicios sean prestados en la casa que habite el cabeza de familia y que perciban un sueldo o remuneración de cualquier clase. Están incluidos los trabajos de guardería, jardinería, conducción de vehículos y otros análogos en los supuestos en que se desarrollen formando parte del conjunto de tareas domésticas.

Como consecuencia de la aprobación del Real Decreto-ley 29/2012, de 28 de diciembre, de mejora de gestión y protección social en el Sistema Especial para Empleados de Hogar y otras medidas de carácter económico y social (BOE de 31 de

diciembre de 2012), se han introducido una serie de modificaciones en la configuración jurídica del Sistema Especial para Empleados de Hogar. Por otra parte, la Ley 36/2014, de 26 de diciembre, de Presupuestos Generales del Estado para el año 2015 (BOE de 30 de diciembre de 2014), ha variado el tipo de cotización por contingencias comunes y su distribución entre empleador y empleado.

Desde el 1 de abril de 2013, los trabajadores incluidos en el Sistema Especial para Empleados de Hogar establecido en el Régimen General de la Seguridad Social que presten sus servicios durante menos de 60 horas mensuales por empleador deberán formular directamente su afiliación, altas, bajas y variaciones de datos cuando así lo acuerden con tales empleadores. En todo caso, las solicitudes de alta, baja y variaciones de datos presentadas por los empleados de hogar deberán ir firmadas por sus empleadores.

El tipo de cotización por contingencias comunes será el 24,70%; siendo el 20,60% a cargo del empleador y el 4,10% a cargo del empleado. Para la cotización por contingencias profesionales se aplicará el 1,10%; a cargo exclusivo del empleador.

Con efectos desde el 1 de abril de 2013, el empleado de hogar que preste sus servicios durante menos de 60 horas mensuales por empleador, y que hubiera acordado con este último la asunción de las obligaciones en materia de encuadramiento en el Sistema Especial para Empleados de Hogar, será el sujeto responsable de la obligación de cotizar. Deberá, por tanto, ingresar la aportación propia y la correspondiente al empleador (o, en su caso, empleadores) con el que mantenga tal acuerdo, por contingencias comunes y profesionales.

Es muy reciente pues la proyección jurídica por la cual la legislación ha buscado equiparar el sector del servicio doméstico al resto de sectores de producción. Conociendo que este sector se encuentra claramente feminizado y que son las mujeres, mayoritariamente, las que se ven afectadas por la situación jurídica irregular en la que este se encuentra nos centramos en ellas durante el desarrollo del presente trabajo. No existen datos oficiales ofrecidos por los organismos nacionales de referencia, por lo que la única opción a la hora de estudiar esta situación es mediante un procedimiento basado en entrevistas realizadas a las mujeres implicadas, para ello seleccionamos una muestra de mujeres empleadas en el servicio doméstico.

En este trabajo hemos incluido, además de las tareas y mantenimiento del hogar, otras como cuidado de menores, personas ancianas, personas enfermas y

discapacitados. Además hemos incluido a las mujeres que trabajan en limpieza de oficinas, dada la similitud e irregularidad laboral en ambas tareas.

Numerosos trabajos han estudiado la problemática de la situación laboral de las mujeres en el servicio doméstico. Entre algunos de ellos destacan:

Patino, Vicente y Galindo (2011) en un estudio realizado en la provincia de Salamanca, comparan las características socioeconómicas de las trabajadoras en el servicio doméstico en situación laboral regular e irregular, e identifican cuatro conglomerados bien diferenciados, dos constituidos por mujeres de nacionalidad española en situación laboral no regular: casadas y solteras. Otro, constituido por mujeres inmigrantes que buscan un complemento monetario para sus gastos personales. Y un último grupo formado por las mujeres empleadas de hogar regularizadas.

En la comunidad de Madrid, la situación específica de la mujer inmigrante en el mercado laboral exige medidas adicionales encaminadas a eliminar la discriminación institucional respecto a los trabajadores en función de su nacionalidad. Cabe destacar la eliminación de los permisos de trabajo que están circunscritos únicamente a un sector de actividad y que frenan la movilidad laboral. Al mismo tiempo, puesto que muchas inmigrantes rumanas disponen de un cierto nivel educativo, debería avanzarse en el reconocimiento de sus cualificaciones específicas (Marcu, 2009).

El servicio doméstico es la actividad donde se concentra la mayor irregularidad laboral femenina, un 36% del empleo irregular total, seguido de la hostelería con un 25%, y el pequeño comercio (Cabo, González, Roces, & Muñoz, 2002).

El estudio llevado a cabo por Escrivá (2000) afirma que las inmigrantes entrevistadas aparecen desencantadas al poco tiempo de llegar por las condiciones tan adversas para ellas, como son la escasez de trabajos bien pagados y el alto costo de la vida como de los alquileres, desarrollando su actividad laboral básicamente como empleadas de hogar. Sin embargo, a diferencia de su trabajo que encuentra cuatro situaciones dentro de las mujeres de nacionalidad extranjera ya asentadas.

El estudio realizado por Perrons, Plomien y Kikey (2010) concluyen que dado que el trabajo doméstico masculino puede comandar salarios superiores a la media, entonces los operarios migrantes no están necesariamente situados en la parte inferior de la jerarquía social. A pesar de que la Unión Europea tiene políticas comunes, y los padres en el Reino Unido y Polonia tienen aspiraciones similares con respecto a los

roles dentro de la familia, las prácticas se forman en gran medida por las divisiones económicas y sociales. Además, las políticas sociales existentes para la promoción de la igualdad de género no reconocen o corregir las normas de género profundamente arraigadas.

En países como Grecia se ha producido también desde 1990 la llegada de mujeres procedentes de los Balcanes y países ex socialistas de Europa central y oriental dedicándose en su mayoría a la actividad del servicio doméstico, cuidado de niños y personas mayores, siendo muy precarias tanto las condiciones laborales como las cotizaciones con respecto a la Seguridad Social (Vassilikou, 2007).

En países como Suecia la situación durante la década de 1990 es un recordatorio de la década de 1930. El aumento de las diferencias de ingresos, así como familias económicamente fuertes se pueden encontrar en ambos períodos. La presencia de la clase media las mujeres en el mercado laboral también parece ser una razón recurrente para la alta demanda para los servicios domésticos. El argumento que se utiliza en ambos periodos fue similar: era considera que es un beneficio para la sociedad si las mujeres bien educadas podrían tener un profesional carrera, mientras que las mujeres menos instruidas se encargó del trabajo de rutina en los hogares (Platzer, 2006).

En este trabajo tratamos de estudiar los perfiles de las empleadas de hogar en las Comunidades de Castilla y León, Castilla la Mancha, Extremadura, Andalucia y Madrid dado el interés que tiene este colectivo en el contexto del trabajo irregular.

2.2 Objetivos

- Conocer las características personales, familiares, profesionales y laborales de las mujeres empleadas de hogar.
- Estimar el porcentaje de trabajadoras en situación laboral irregular en el sector de actividad del servicio doméstico.
- Conocer e identificar los perfiles socioeconómicos y laborales de las mujeres que se dedican a dicha actividad.
- Comparar los métodos de segmentación CHAID y TAID en la búsqueda de perfiles de las trabajadoras empleadas de hogar.

2.3 Material y Métodos

La unidad estadística objeto de investigación son mujeres que se dedican a la actividad del servicio doméstico, con edad comprendida entre 16 y 65 años, empleadas de manera regular o irregular.

Para configurar las dimensiones y los ítems que conforman el cuestionario se realizaron entrevistas en profundidad y grupos de discusión, con el fin de obtener una visión global del fenómeno objeto de estudio desde distintas perspectivas.

El Cuestionario consta de 50 preguntas que recogen Características Sociodemográficas, Formación y Cualificación, Situación laboral y trabajo, Consecuencias de la situación de irregularidad sobre la vida laboral y personal, Demandas y Medidas.

La muestra se obtuvo en dos etapas: la primera tuvo lugar en el año 2012, tras la modificación de la ley que regulaba el empleo doméstico y cuenta con 742 trabajadoras. La segunda etapa ha sido realizada a principio del año 2015 y forman parte de ella 428 trabajadoras. En sendos casos y dadas las características de este estudio se ha realizado un muestreo en bola de nieve.

Cabe destacar que en la muestra obtenida casi el 60% son trabajadoras de la provincia de Salamanca. Por otro lado, del 40% restante la mayoría de trabajadoras son de Castilla y León y Extremadura (lugares de procedencia de dichos alumnos) con algunas excepciones que dejan presentes ciudades de Castilla la Mancha (como Toledo o Cuenca), ciudades andaluzas (como Almería o Huelva) y la ciudad Madrid.

2.4 Descripción de la muestra

La muestra objeto de nuestro estudio está compuesta por 1170 trabajadoras empleadas en el servicio doméstico, de las cuales la mayoría (78%) tienen entre 30 y 55 años, de ellas el 40% son mayores de 46 años y el 38% menores. Por otro lado, menos del 1% de las trabajadoras tienen una edad inferior a 20 años (ver figura 2.1). El 13% están entre 20 y 29 años y el 9% son mayores de 55 años.

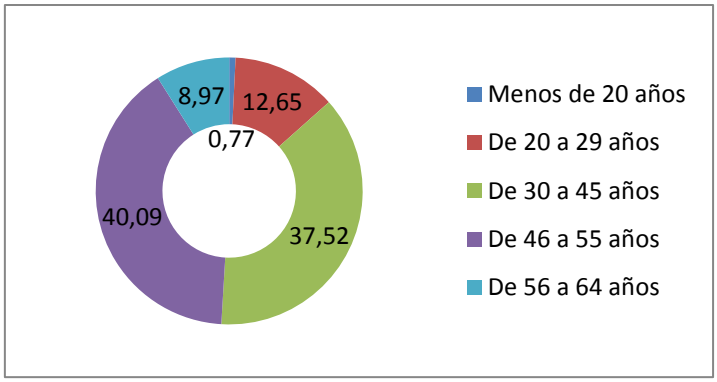

Figura 2.1: Porcentajes de la edad de las encuestadas

Atendiendo ahora a su estado civil, podemos decir que el 57% (la mayoría) son casadas y un 21% están solteras. De todas las trabajadoras el 13% están separadas o divorciadas y el 9% restante son viudas o tienen una pareja de hecho repartidas casi al 50% (tabla 2.1).

	Frecuencia	Porcentaje
Solteras	244	20,9%
Casadas	688	57,1%
Separadas/Divorciadas	146	12,5%
Viudas	61	5,2%
Pareja de hecho	51	4,4%

Tabla 2.1: Frecuencias y porcentajes del estado civil de las encuestadas

En cuanto a su nivel de estudios, aproximadamente el 50% tienen un nivel primario, el 9% no tiene ningún tipo de estudios y el resto (41%) tienen estudios secundarios, universitarios o estudios no reglados. De este 41%, lel 23% tienen estudios secundarios y solo un 1% tienen estudios no reglados. El 12% de ellas tienen un título de formación profesional y solo el 5% estudios superiores: diplomadas, licenciadas o estudios de postgrado. De ambas categorías, vemos que la más frecuente con aproximadamente un 3% son las diplomadas, seguidas por un 1% de licenciadas y finalmente, el colectivo menos usual son aquellas que tienen estudios de postgrado que en nuestro caso sólo tenemos 4 trabajadoras que lo cumple (un 0,3%) (ver figura 2.2).

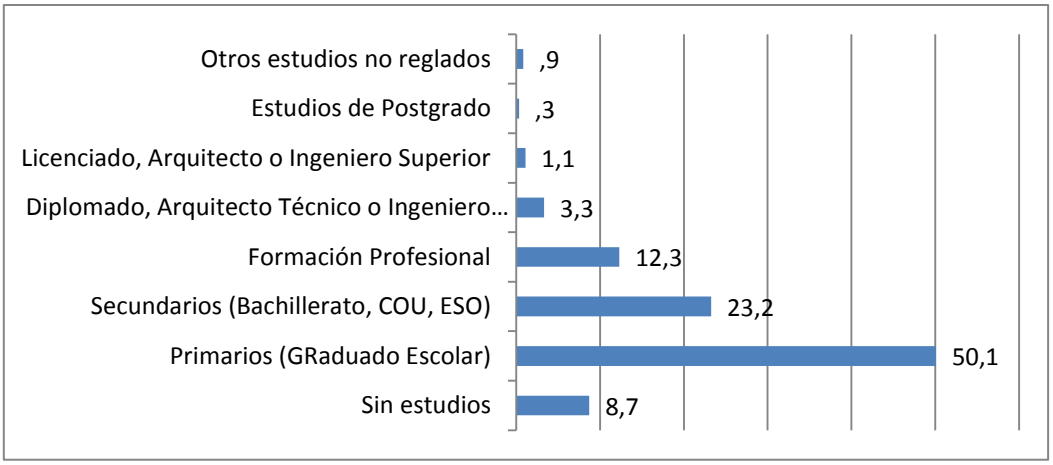

Figura 2.2: Porcentajes del nivel de estudios de las encuestadas

De todas las trabajadoras de la muestra un 59% tienen contrato de trabajo, el 21% tienen un acuerdo verbal con su empleador y el 20% no tienen contrato de trabajo, ni acuerdo verbal por lo que están en una situación de empleo irregular (ver figura 2.3).

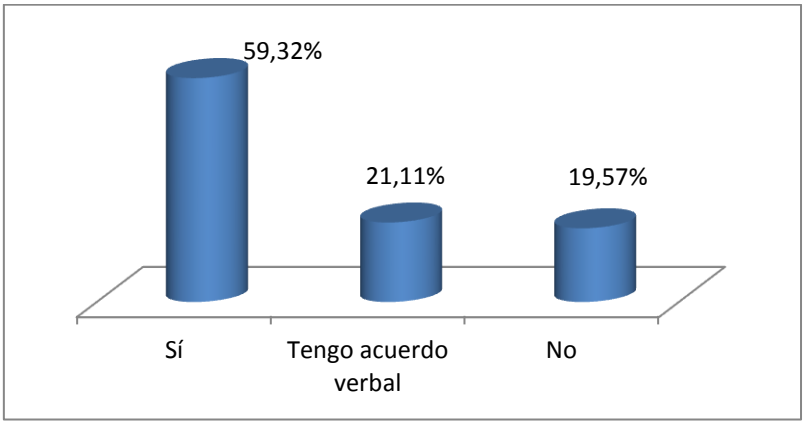

Figura 2.3: Porcentajes del contrato de las encuestadas

A pesar de esto, el 95% de las empleadas reciben remuneración por su trabajo mientras que 54 de las empleadas trabajan sin percibir ninguna remuneración (ver tabla 2.2).

	Frecuencia	Porcentaje
Sí	1116	95,4%
No	54	4,6%

Tabla 2.2: Frecuencias y porcentajes de la remuneración de las encuestadas

De todas las trabajadoras de la muestra el 59% están en una situación regular, bien sea porque cotizan por todas las horas que trabajan (52%) o por estar exentas en su cotización (7%), sin embargo, el 31% restante presentan una irregularidad en su situación laboral, bien porque coticen por menos horas de las que deban (15%) o no coticen aun teniendo que hacerlo (26%) (ver figura 2.4).

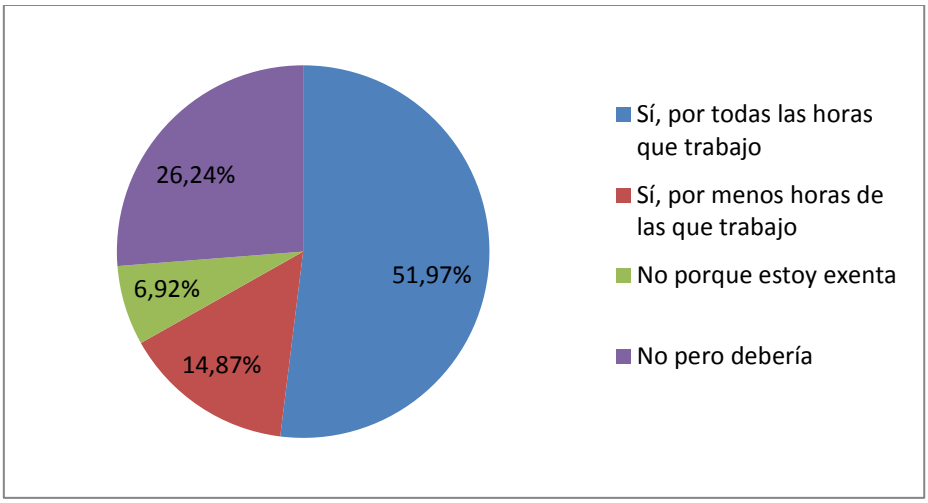

Figura 2.4: Porcentajes de la cotización de las encuestadas

Por otro lado, encontramos que aquellas que están es una situación irregular no tienen tarjeta de la Seguridad Social como titulares, si analizamos este punto detenidamente observamos que el 73% tiene una tarjeta de la Seguridad Social como titulares mientras que el 27% restante tendrá una tarjeta de la Seguridad Social como beneficiaria, o incluso carecerá de dicha tarjeta (ver tabla 2.3).

	Frecuencia	Porcentaje
Titulares	859	73,4%
Otros	311	26,6%

Tabla 2.3: Frecuencias y porcentajes del tipo de tarjeta de las encuestadas

En cuanto a las trabajadoras que cotizan a la Seguridad Social (67%), es importante señalar que la cotización no ha de ser llevada a cabo exclusivamente por las propias trabajadoras, según la legislación vigente se puede cotizar de tres maneras: cotizando la propia trabajado, realizándolo el empleador o a partes iguales entre la empleada y el empleador. Si miramos nuestra muestra detenidamente, observamos que de las trabajadoras que sí cotizan a la Seguridad Social tan sólo el 28% lo hacen ellas, otro 32% lo hacen a partes iguales el empleador y la empleada mientras que el 40% restante, es decir, en el grupo mayoritario, la cotización recae sobre el empleador (ver figura 2.5).

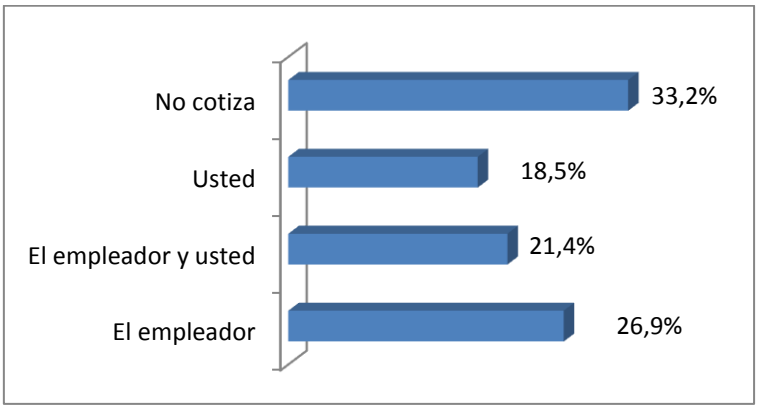

Figura 2.5: Porcentajes de el/la cotizador/a en la Seguridad Social

Por otro lado, el 33% de las empleadas no cotizan en la Seguridad Social, lo cual puede ser debido a diferentes motivos. De las trabajadoras que no cotizan de nuestra muestra, el 22% lo hace porque no tienen obligación (7% del total que estaban exentas), el 36% porque no le interesa o no le compensa, mientras tanto el 23% no es consciente ni siquiera de si tiene que hacerlo y el 19% restante da otro tipo de explicación a este hecho (ver tabla 2.4).

	Frecuencia	Porcentaje
No tiene obligación	87	7,4%
No le interesa o compensa	138	11,8%
No sabe si tiene que cotizar	89	7,6%
Otra explicación	74	5,3%
Sí cotiza	782	66,8%

Tabla 2.4: Frecuencias y porcentajes del motivo de no cotización de las encuestadas

Las trabajadoras del servicio doméstico pueden trabajar en distintas modalidades de empleo. Pueden ser trabajadoras internas, residentes en el domicilio o trabajadoras externas. En la muestra que en este caso nos ocupa, el 15% son trabajadoras internas y el 85% restante son trabajadoras externas. De este segundo grupo diferenciaremos un 34% que son externas pero presentan un trabajo fijo y el otro 51% que serán externas pero trabajando por horas (ver figura 2.6).

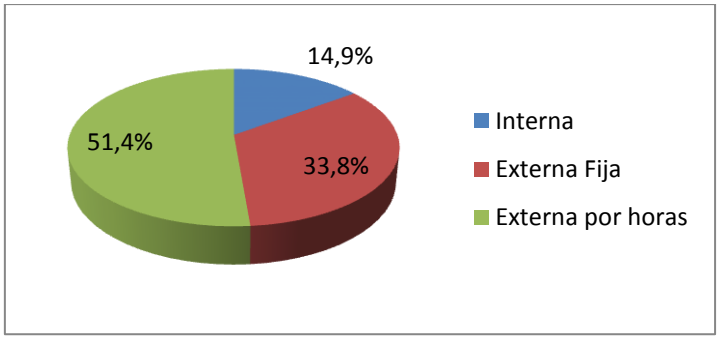

Figura 2.6: Porcentajes del tipo de empleo de las encuestadas

Podemos también observar el tiempo que llevan estas trabajadoras dedicándose al servicio doméstico. La mayoría (68%) lleven ya más de tres años en este sector. Por otro lado, el 19% llevan entre uno y tres años en el servicio doméstico mientras que el 13% restante llevan menos de un año, de ellas, un 8% más de seis meses, un 4% entre tres y seis meses y solo un 1% acaba de empezar y lleva un tiempo inferior a tres meses en este sector (ver figura 2.7).

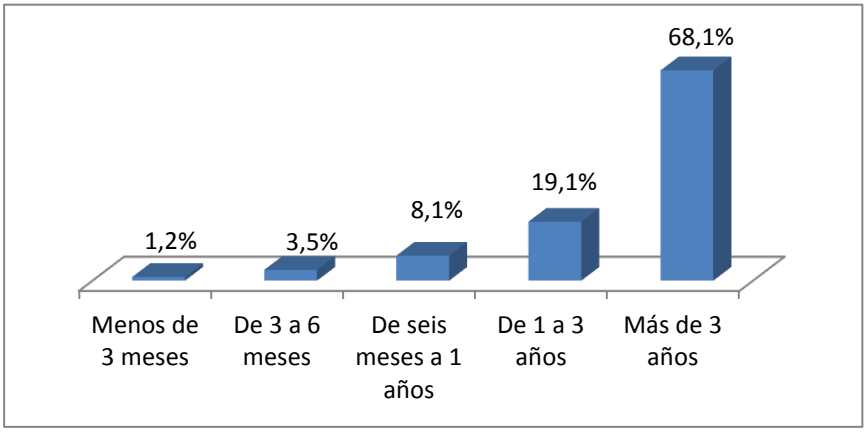

Figura 2.7: Porcentajes del tiempo en el servicio doméstico de las encuestadas

2.5. Búsqueda de los perfiles de las mujeres empleadas de hogar, basada en información multivariante mediante el algoritmo CHAID

Partimos de cinco predictores (edad, estado civil, nivel de estudios, modalidad de empleo y tiempo de dedicación al servicio doméstico) y seis variables respuestas (recibir remuneración por su trabajo, tipo de contrato, cotización o no a la Seguridad Social, quien cotiza o los motivos por los que no se cotiza y posesión de la tarjeta sanitaria de la Seguridad Social). En este caso, las variables respuestas no pueden ser usadas de manera conjunta de modo que tendremos que realizar tantos Chaids como variables respuestas tenemos (uno por cada una) para explicarlos y poder encontrar las diferencias con el método Taid.

- En primer lugar, procedemos a la realización del algoritmo CHAID cogiendo como variable respuesta **si las trabajadoras, tienen firmado o no su contrato de trabajo** y los cinco predictores considerados en el análisis: edad, estado civil, nivel de estudios, modalidad de empleo y tiempo de dedicación al servicio doméstico

Estudiamos si existe algún predictor que no sea válido para segmentar. Para ello, cruzamos la variable respuesta con cada predictor y hallamos el valor del chi-cuadrado, el p-valor y el coeficiente de contingencia.

Este proceso se realiza con todos y cada uno de los predictores. Analizando los resultados uno a uno obtendremos los siguientes valores:

Variable	Categorías	Chi	G.L.	p-valor	CC
Estado Civil	5	7,126	8	0,523	0,078
Nivel Estudios	8	29,124	14	0,01	0,156
Edad	5	31,287	8	0,000	0,161
Tiempo dedicación	5	13,397	8	0,026	0,121
Modalidad Empleo	3	125,355	4	0,000	0,311

Tabla 2.5: Estadísticos de cada variable para la segmentación

En esta tabla observamos que hay una variable que no presenta asociación estadísticamente significativa para segmentar, el estado civil, puesto que no es significativa. Por otro lado, las otras cuatro sí son significativas. Procedemos ahora a intentar agrupar las categorías según tengan o no un perfil similar con respecto a la variable respuesta. Para ello vamos a ver primero que tipo de variables tenemos y cuáles se pueden agrupar.

Predictor	Tipo
Nivel Estudios	Flotante
Edad	Monótono
Tiempo dedicación	Monótono
Modalidad Empleo	Libre

Tabla 2.6: Tipos de predictores

Es decir, en la edad y el tiempo de dedicación se pueden sólo agrupar categorías contiguas, en la modalidad de empleo se pueden agrupar sin importar el orden y en el nivel de estudios se podrán agrupar de manera contiguas las categorías ordinales (Sin estudios, Primaria, Secundaria) seguida por las nominales (Formación Profesional, Diplomada o Licenciada) y nuevamente por la ordinal (Máster-Doctorado) con mucho cuidado en las categorías que se agrupan.

Procedemos a la agrupación de las categorías empezando por la variable edad:

Parejas	Chi-cuadrado	p-valor
1-2	23,047	0,001
2-3	22,128	0,001
3-4	26,317	0,000
4-5	30,075	0,000

Tabla 2.7: Posibles agrupaciones de categorías

Como vemos todas las parejas posibles son significativas, por tanto, no podemos hacer ninguna agrupación. Este proceso habría que realizarlo con las otras dos variables que nos faltan. Después de esto habría que realizar nuevamente las tablas de contingencia para ver cuál es el mejor predictor. En este caso, la tabla de contingencia se crea con las agrupaciones hechas. Por ejemplo, la clasificación de la edad no varía ya que no hemos agrupado ninguna pero en la modalidad de empleo sí, por ello esa fila de la tabla antes creada cambiaría. Una vez elegido el mejor predictor se segmenta en tantas ramificaciones como categorías tenga el predictor. Posteriormente se realiza todo el proceso en cada nodo no terminal. Realizado todo el algoritmo obtenemos el árbol final.

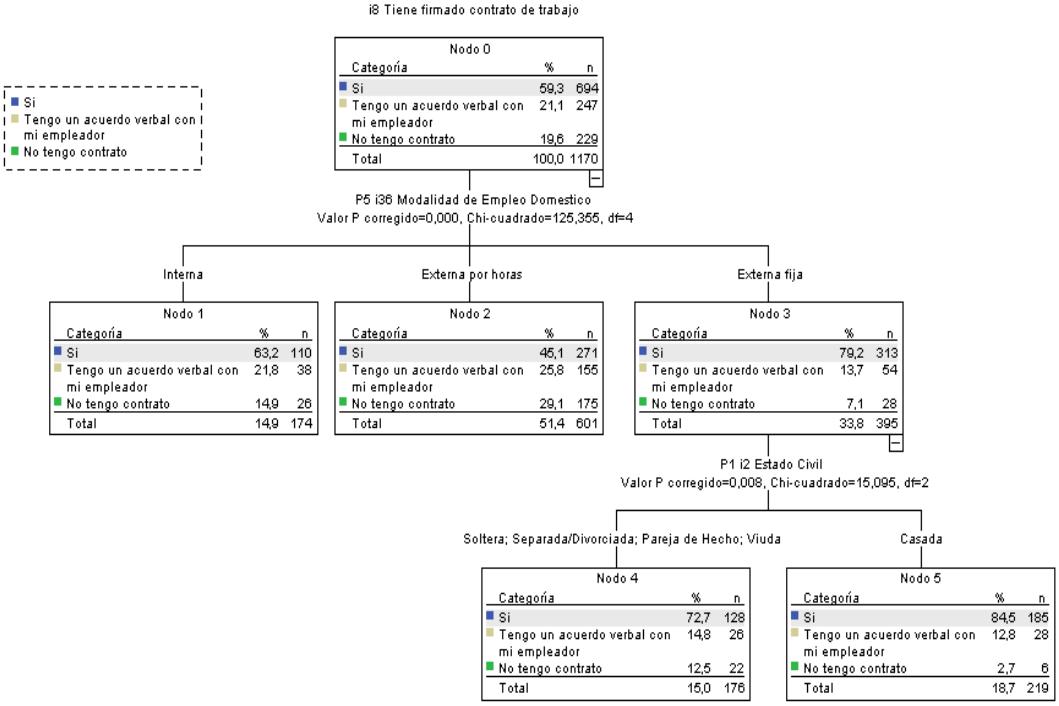

Figura 2.8: Árbol final para el tipo de contrato

En este caso tenemos 6 nodos de los cuales 4 son nodos terminales, que son los que caracterizan a la variable respuesta, por lo que encontramos cuatro perfiles de las trabajadoras en cuanto al tipo de contrato que tienen las trabajadoras.

En primer lugar, el 59,3% de las trabajadoras tienen firmado un contrato de trabajo, el 21,1 tienen un acuerdo verbal y el 19,6% no tienen contrato. A continuación, se selecciona el mejor predictor en el grupo de predictores que en este caso vemos que es la variable modalidad de empleo para las mujeres entrevistadas según el contrato firmado. También observamos que esta variable no presenta una agrupación

óptima de sus categorías. Ahora veremos si el predictor es significativo o no para cada una de sus categorías para realizar una segunda segmentación. Vemos que tanto para la modalidad de trabajadoras internas como las externas por horas son categorías significativas. En sendos grupos observamos que es más frecuente tener un contrato de trabajo en las internas con un 63% y en las externas por horas con un 45%. Para la categoría de trabajadoras fijas el algoritmo realiza otro cruce y para ello escoge como mejor predictor la variable estado civil. En este caso, vemos que el algoritmo sí realiza una agrupación, junta todas las categorías menos a las casadas, que quedan separadas del resto. En este caso, los dos nodos obtenidos son significativos y, por tanto, terminales.

Podemos concluir que el 84% de las trabajadoras tienen firmado un contrato de trabajo, están casadas y trabajan como externas fijas. mientras que del resto de externas fijas (tengan el estado civil que tengan) solo tienen un contrato el 72%. En ambos grupos hay aproximadamente el mismo porcentaje de acuerdos verbales pero en el segundo que hemos explicado hay un 10% más de mujeres sin contrato.

- A continuación procedemos a la realización del algoritmo CHAID cogiendo como variable respuesta **si las trabajadoras, reciben remuneración por su trabajo.**

El procedimiento sería el mismo que en el caso anterior. El árbol resultante sería el siguiente:

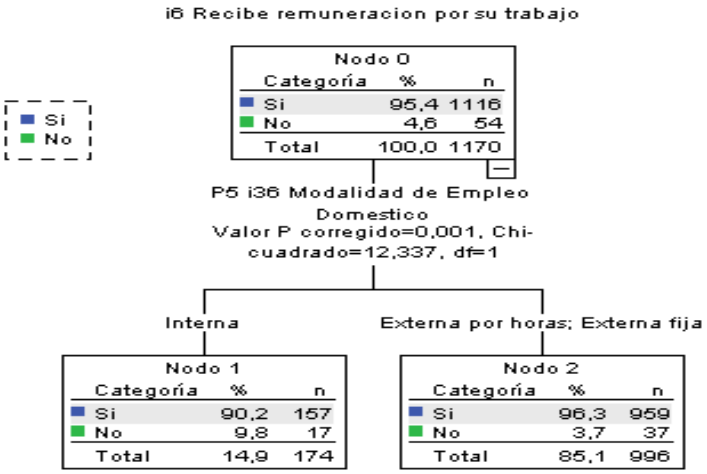

Figura 2.9: Árbol definitivo de la recepción de remuneración

Este árbol presenta solo tres nodos de los cuales dos son terminales. Se selecciona como mejor predictor la variable modalidad de empleo para las mujeres

entrevistadas la remuneración que reciben por su trabajo. En este caso vemos que esta variable presenta una agrupación óptima de sus categorías, agrupa las categorías externa fija y externa por horas. Observamos también que sendos grupos son significativos. En los dos grupos hay un altísimo porcentaje de trabajadoras que reciben remuneración por su trabajo, mayor al 90%, pero la diferencia principal se encuentra en que en las internas hay casi el triple de trabajadoras que no reciben remuneración de las que hay entre las externas.

- A continuación se realiza el algoritmo CHAID cogiendo como variable respuesta **si las trabajadoras, cotizan a la Seguridad Social.**

El gráfico resultante del algoritmo es el siguiente:

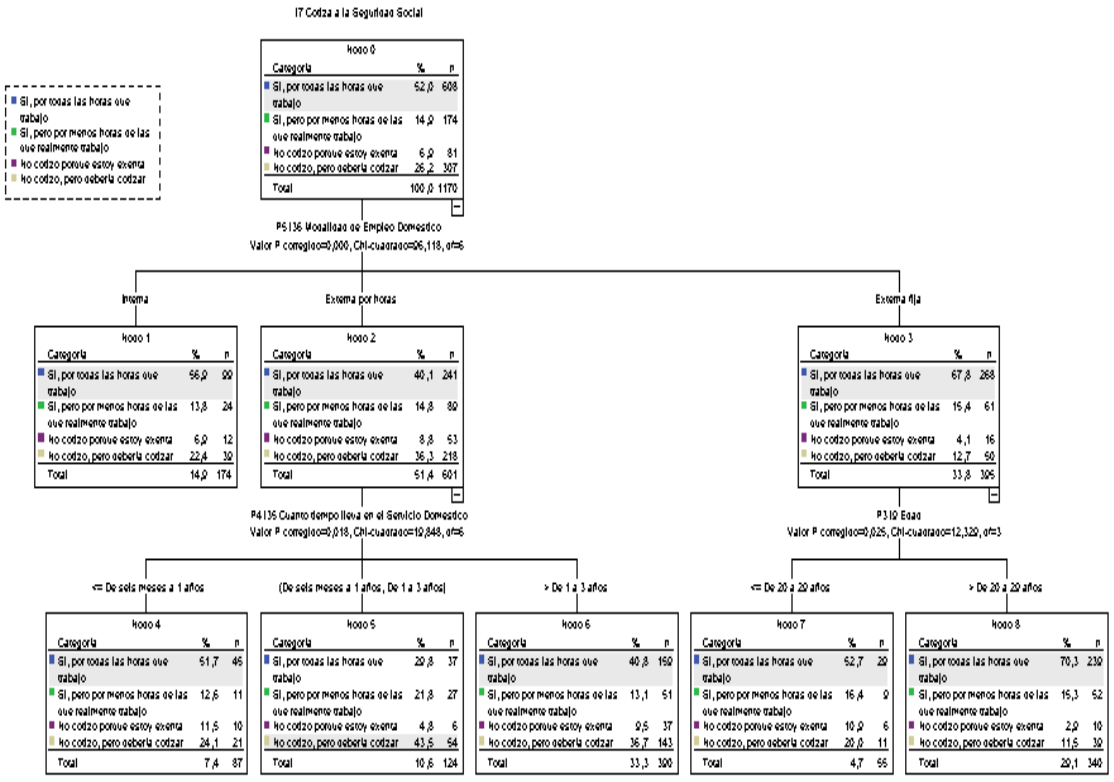

Figura 2.10: Árbol definitivo para la cotización en la Seguridad Social

En este caso obtenemos 9 nodos de los cuales 6 son nodos terminales. En primer lugar, se selecciona el mejor predictor en el grupo de predictores que en este caso vemos que es la variable modalidad de empleo para las mujeres entrevistadas según su cotización en la Seguridad Social. También observamos que esta variable no

presenta una agrupación óptima de sus categorías. Ahora vemos si el predictor es significativo o no para cada una de sus categorías para realizar una segunda segmentación. Vemos que la modalidad de trabajadoras internas es una categoría significativa. El 57% de las trabajadoras sí cotizan a la Seguridad Social. Para en la categoría de trabajadoras fijas, el algoritmo realiza otro cruce y para ello escoge como mejor predictor la variable edad. Como vemos, la variable sí presenta una agrupación óptima de sus categorías, haciendo una distinción entre las mayores y menores de 30 años. En este caso, los dos nodos obtenidos son significativos y por tanto, terminales.

Concluimos por tanto que el 53% de las externas fijas menores de 30 años sí cotizan a la Seguridad Social mientras que las mayores de 30 años cotizan a la Seguridad Social por todas las horas que trabajan el 70%. Vemos también que hay aproximadamente el mismo porcentaje de mujeres que cotizan a la Seguridad Social por menos horas de las que trabajan en sendos nodos. Por otro lado, para la categoría de trabajadoras por horas existe también un predictor significativo para seguir segmentando que es el tiempo de dedicación al servicio doméstico. En este caso, también vemos una agrupación en las categorías de dicha variable: por un lado tenemos las trabajadoras que llevan menos de un año dedicándose al servicio doméstico, por otro las que llevan de 1 a 3 años y en una última rama las que llevan más de 3 años. En este caso, los dos nodos obtenidos son significativos y, por tanto, terminales. El 52% de las trabajadoras externas por horas que llevan menos de un año trabajando cotizan y por todas las horas que trabajan. Sin embargo, de las trabajadoras externas por horas que llevan entre 1 y 3 años trabajando el 43% no cotiza pero debería cotizar. Para finalizar, de las trabajadoras externas por horas que llevan más de 3 años trabajando cotizan por todas la horas que trabajan, el 41% seguidas de cerca con un 37% por aquellas que no cotizan pero sí deberían hacerlo.

- A continuación se realiza el algoritmo CHAID cogiendo como variable respuesta **si las trabajadoras, tienen tarjeta sanitaria como titulares de la Seguridad Social.**

El árbol resultante del algoritmo es el siguiente:

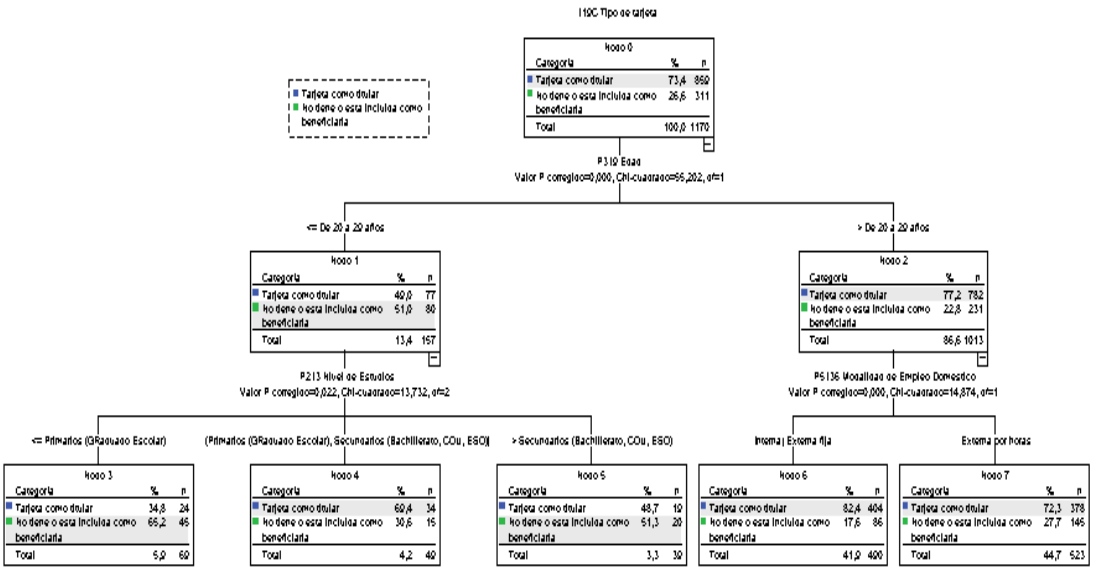

Figura 2.11: Árbol definitivo para el tipo de tarjeta

En este caso se obtienen 8 nodos de los cuales 5 son nodos terminales. Primero se selecciona el mejor predictor que, como vemos, es la variable edad para las mujeres entrevistadas según su tipo de tarjeta de la Seguridad Social. En este caso la variable edad presenta una agrupación óptima de sus categorías, habían quedado dos grupos; uno con mujeres menores de 30 años y otro con las mayores. Ahora vemos si el predictor es significativo o no para cada una de sus categorías para realizar una segunda segmentación. Vemos que en sendos grupos se ha realizado una segunda segmentación. Para la categoría de trabajadoras menores de 30 años se muestra la existencia de un nodo significativo que es el nivel de estudios. En este caso, vemos que la variable presenta una agrupación óptima de sus categorías en tres grupos: las que tienen estudios primarios, las que tienen estudios secundarios y el resto de trabajadoras con otros estudios. En este caso, los dos nodos obtenidos son significativos y, por tanto, terminales. Como vemos el 65% de las mujeres menores de 30 años con estudios primarios no tienen tarjeta de la Seguridad Social como titular. Por otra parte, el 69% de las mujeres menores de 30 años y con estudios secundarios tienen tarjeta propia como titulares de la Seguridad Social. Finalmente, de las mujeres menores de 30 años con estudios superiores a los secundarios aproximadamente la mitad tienen tarjeta de la Seguridad Social como titulares y la otra mitad no. Estudiando las mujeres mayores de 30 años, vemos que en esa ramificación también ha existido otro predictor significativo con el que seguir segmentando que es la variable modalidad de empleo. Además, vemos que dicha variable presenta una agrupación de sus categorías, agrupando por un lado las internas con las que trabajan externas por horas y dejando en el otro lado solo a las que trabajan externas pero fijas.

De las trabajadoras mayores de 30 años y con un trabajo interno o externo por horas, el 82% tienen tarjeta de la Seguridad Social como titulares mientras que en el caso de las trabajadoras que son externas pero están fijas el 72% tienen tarjeta de la Seguridad Social, un 10% menos que las anteriores.

- A continuación, se realiza el algoritmo CHAID cogiendo como variable respuesta **quién cotiza a la Seguridad Social (el empleador, la trabajadora y el empleador, sólo la trabajadora o no cotiza).**

El gráfico resultante es el siguiente:

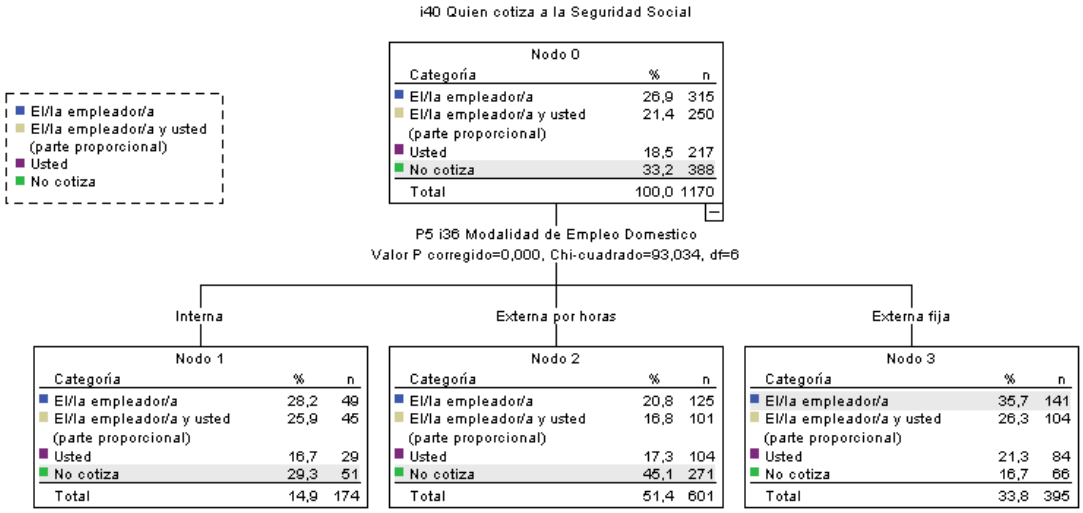

Figura 2.12: Árbol definitivo para la persona que cotiza

En este árbol se observan cuatro nodos, de los que terminales son tres. En primer lugar, se selecciona el mejor predictor en el grupo de predictores que en este caso vemos que es la variable modalidad de empleo para las mujeres entrevistadas según quién cotiza a la Seguridad Social. Esta variable no presenta una agrupación óptima de sus categorías y además que todas ellas son significativas. Observamos que es más frecuente que las trabajadoras no coticen, esto es lógico porque el 33% de las respuestas de esta variable están en la categoría de no cotizan. En el caso de las trabajadoras internas en el 28% quien cotiza es el empleador seguido por casi un 26% que cotizan a medias el empleador y la empleada. En el caso de las externas por horas vemos que los porcentajes son muy similares pero hay un mayor porcentaje de trabajadoras de que quien cotice sea el empleador. Por otro lado, en el caso de las externas fijas quien cotiza es el empleador en un 35%. Vemos también en los casos

de externas que no hay gran diferencia entre que cotice el empleador y la empleada a la mitad y que cotice solo la empleada.

- Y por último, se realiza el algoritmo CHAID cogiendo como variable respuesta **por qué las trabajadoras no cotizan a la Seguridad Social (No tienen obligación, no le interesa o no le compensa, no sabe si tiene que cotizar, otra explicación o sí cotiza)**

El gráfico resultante del algoritmo es el siguiente:

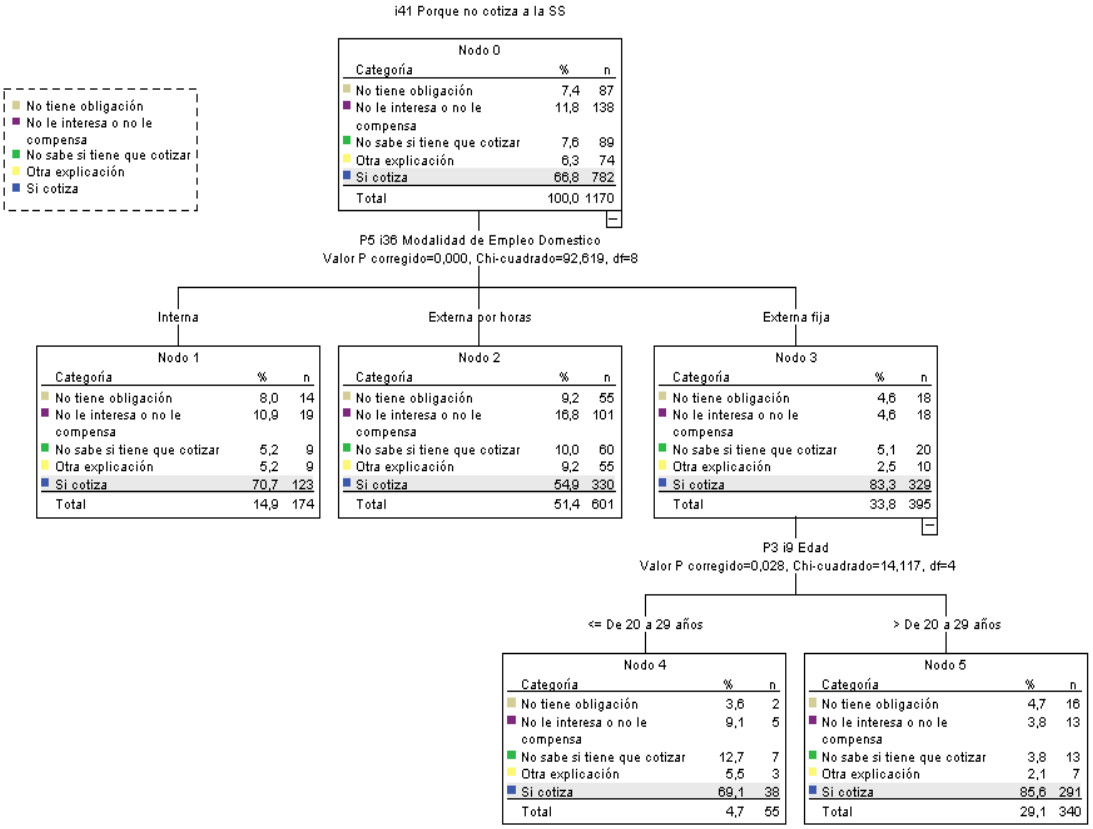

Figura 2.13: Árbol definitivo para el objeto de no cotización

En este caso, el árbol se compone de 6 nodos y 4 de ellos son terminales. Como vemos, nuevamente el mejor predictor es la variable modalidad de empleo para las mujeres entrevistadas según el motivo por el cual no cotizan a la Seguridad Social. La variable nuevamente no presenta una agrupación óptima de sus categorías. Ahora vemos si el predictor es significativo o no para cada una de sus categorías para realizar una segunda segmentación. Vemos que tanto para la modalidad de trabajadoras internas como las externas por horas son categorías significativas. En sendos grupos observamos que es más frecuente que las trabajadoras coticen, esto es lógico porque el 66% de las respuestas de esta variable están en la categoría de sí

cotizan. En ambos casos se ve que es mayor el desinterés o que no cotizan porque no les compensa aunque en el caso de las externas por horas con un 6% más que en el caso de las internas, seguido de cerca porque no tienen obligación. Para la categoría de trabajadoras fijas el algoritmo realiza otro cruce y para ello escoge como mejor predictor la variable edad. En este caso, vemos que la variable sí presenta una agrupación óptima de sus categorías, agrupando por un lado las menores de 30 años y por otro las mayores. Ahora sí los dos nodos obtenidos son significativos y, por tanto, terminales. En el caso de las trabajadoras externas fijas menores de 30 años se ve que no cotizan porque no saben si lo tienen que hacer seguido de cerca por las que no les interesa o no les compensa. Por otro lado, en el caso de las trabajadoras externas fijas mayores de 30 años predominan las que no tienen obligación de cotizar a la Seguridad Social, seguidas por las que no saben si lo tienen que hacer o no les interesa o compensa.

Finalizada la aplicación del algoritmo Chaid procedemos a exponer el algoritmo Taid que, como sabemos, aparte de ser un algoritmo más moderno, corrige la problemática que presentaba el Chaid y su resultado puede ser más efectivo.

2.6. Búsqueda de los perfiles de las mujeres empleadas de hogar, basada en información multivariante mediante el algoritmo Taid

Como ya hemos dicho, la ventaja de este algoritmo es que podremos utilizar los cinco predictores y las seis variables respuestas a la vez y obtener un único árbol.

A continuación, desarrollamos el algoritmo en cada una de sus fases.

FASE I: Búsqueda de clases latentes

Esta fase consiste en un análisis de clases latentes con todas las variables respuestas, como ya hemos dicho con el fin de obtener una única variable respuesta con varias categorías (clase latente).

Nos ha parecido conveniente realizar un análisis de clases latentes con Bootstrap, siguiendo la propuesta de Araya, dada la naturaleza de nuestros datos, dado que tendríamos un número de patrones de respuesta muy elevado y, por tanto es conveniente remuestrear para obtener una mayor representatividad.

Además, para realizar el análisis de clases latentes utilizaremos el programa Winmira (Von Davier, 2001) en el cual tendremos que realizar el análisis de clases

usando dos clases, luego tres y así sucesivamente hasta encontrar un número de clases que nos resulte apropiado..

Realizamos un primer análisis con dos clases latentes. En primer lugar mostramos los resultados del modelo general, sin el remuestreo Bootstrap. Para que el modelo sea considerado adecuado necesitamos un p-valor superior a 0,9:

	Valor experimental	Grados de libertad	p-valor
Cressie Read	487,69	910	1,0000
Chi-cuadrado Pearson	787,07	910	0,9980

Tabla 2.8: p-valores del modelo general para análisis de clases latentes con dos clases

Dado que el p-valor=0,9980, consideramos el modelo adecuado, y por tanto observamos el p-valor en el remuestreo. Para considerar el modelo adecuado los p-valores resultantes deben ser no significativos.

	p-valor
Cressie Read	0,025
Chi-cuadrado Pearson	0,000

Tabla 2.9: p-valores del remuestreo Bootstrap para el análisis de clases latentes con dos clases

En este caso el modelo no podría ser seleccionado (p-valor<0.05), por tanto nos indica que necesitamos coger más de dos clases latentes.

El proceso se repetirá hasta encontrar el mejor número de clases latentes. Mostramos a continuación una tabla con todos los p-valores obtenidos:

p-valores	Modelo General		Remuestreo Bootstrap	
	Cressie Read	Coeficiente Pearson	Cressie Read	Coeficiente Pearson
2 clases	1,0000	0,9980	0,025	0,000
3 clases	1,0000	1,0000	0,050	0,075
4 clases	1,0000	1,0000	0,050	0,075
5 clases	1,0000	1,0000	0,400	0,400

Tabla 2.10: p-valores del modelo general y remuestreo Bootstrap para análisis de clases latentes de dos a cinco clases

Según el modelo general podríamos coger cualquier opción pero centrándonos en el remuestreo Bootstrap solo podemos coger de tres clases en adelante. Por otro lado, coger tres clases latentes o cuatro nos resultaría igual de efectivo en cuanto a los resultados obtenidos, no obstante es mucho más efectivo seleccionar tres clases que cuatro para el análisis puesto que seleccionar cuatro podría ser más complejo. Finalmente seleccionar cinco clases sería mucho más eficaz en cuanto a los p-valores pero mucho más complejo en el análisis.

Por tanto, como resultado del análisis obtenemos tres clases latentes:

Clase 1: Representa el 60% de la muestra. Esta clase está compuesta únicamente por mujeres que cotizan a la Seguridad Social (aunque en la mayoría de los casos la cotización quede a cargo del empleador), de las cuales un 86,6% de ellas lo hace por todas las horas que trabaja. De las mujeres de este grupo, el 93,2% tienen un contrato de trabajo y, además, el 91,3% tienen tarjeta sanitaria de la Seguridad Social como titular. Podemos caracterizarla entonces como una clase compuesta generalmente por mujeres que cotizan a la Seguridad Social por todas su horas, con contrato de trabajo y tarjeta de la Seguridad Social como titulares.

El gráfico siguiente nos muestra las probabilidades de cada categoría de las variables predictoras en esta clase latente:

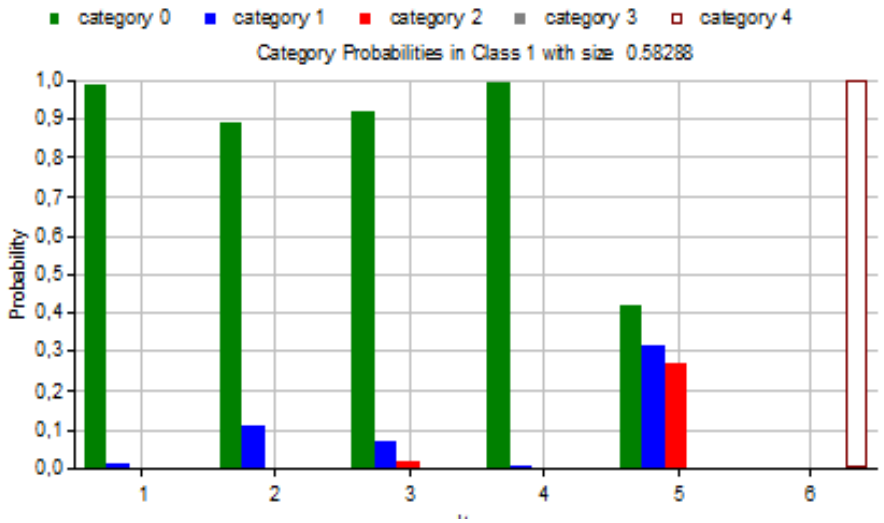

Figura 2.14: Probabilidades de cada categoría de las variables en la clase 1
1: Recibe remuneración por su trabajo: ***category 0:*** Si; **category 1:**No.
2: Cotiza a la Seguridad Social: **category 0**: Por todas las horas que trabajan; **category 1**: Cotizan pero no por todas las horas; **category 2**: Están exentas; **category 3**: No cotizan pero deberían.
3: Tipo de contrato: **category 0**: Si; **category 1**: Acuerdo verbal: **category 2:** No.
4: Tarjeta Seguridad Social: **category 0**: Si; **category 1**: No.
5: Quien cotiza a la Seguridad Social: **category 0**: El empleador; **category 1**: El empleador y la trabajadora; **category 2**: La trabajadora; **category 3**: No cotizan.
6:Porque no cotiza a la Seguridad Social: **category 0**: No tiene obligación; **category 1**: No le interesa o no le compensa; **category 2:** No sabe si tiene que cotizar; **category 3**: Otra explicación/Si cotizan

Clase 2: En este caso se representa al 33% de la muestra. Esta clase está compuestas por todas las mujeres que no cotizan a la Seguridad Social (en su mayoría porque no les interesa), de ellas un 79,1% no lo hace porque están exentas. De las mujeres de este grupo, el 55,7% no tienen contrato de trabajo y el 53,1% no tienen tarjeta de la Seguridad Social o están incluidas en alguna como beneficiarias. Entonces podemos decir que esta clase queda caracterizada por mujeres que no cotizan a la Seguridad Social por estar exentas, sin contrato de trabajo ni tarjeta de la Seguridad Social como titulares.

El gráfico siguiente nos muestra las probabilidades de cada categoría de las variables predictoras:

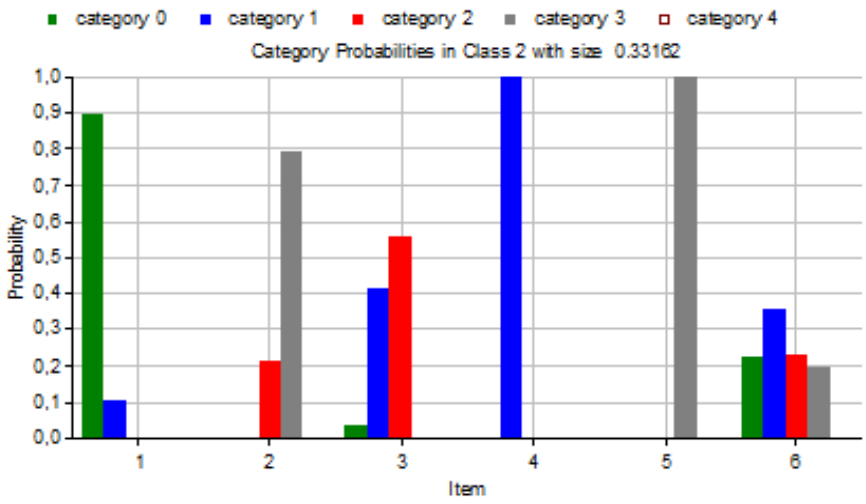

Figura 2.15: Probabilidades de cada categoría de las variables en la clase 2
1: Recibe remuneración por su trabajo: ***category 0:*** Si; **category 1:**No.
2: Cotiza a la Seguridad Social: **category 0**: Por todas las horas que trabajan; **category 1**: Cotizan pero no por todas las horas; **category 2**: Están exentas; **category 3**: No cotizan pero deberían.
3: Tipo de contrato: **category 0**: Si; **category 1**: Acuerdo verbal: **category 2:** No.
4: Tarjeta Seguridad Social: **category 0**: Si; **category 1**: No.
5: Quien cotiza a la Seguridad Social: **category 0**: El empleador; **category 1**: El empleador y la trabajadora; **category 2**: La trabajadora; **category 3**: No cotizan.
6:Porque no cotiza a la Seguridad Social: **category 0**: No tiene obligación; **category 1**: No le interesa o no le compensa; **category 2:** No sabe si tiene que cotizar; **category 3**: Otra explicación/Si cotizan.

Clase 3: Este grupo representa solo al 7% de la muestra. Está formado por mujeres que cotizan a la Seguridad Social (la mayoría cotizan ellas mismas o a medias con el empleador) y todas por menos horas de las que trabajan. De ellas, el 60% tienen un acuerdo verbal con su empleador y, además, el 55% no tienen tarjeta de la Seguridad Social como titulares. La tercera clase latente se puede caracterizar entonces como una clase compuesta por mujeres que cotizan a la Seguridad Social

pero por menos horas de las que trabajan, con un acuerdo verbal con sus empleadores y sin tarjeta de la Seguridad Social como titulares.

El gráfico siguiente nos muestra las probabilidades de cada categoría de las variables predictoras:

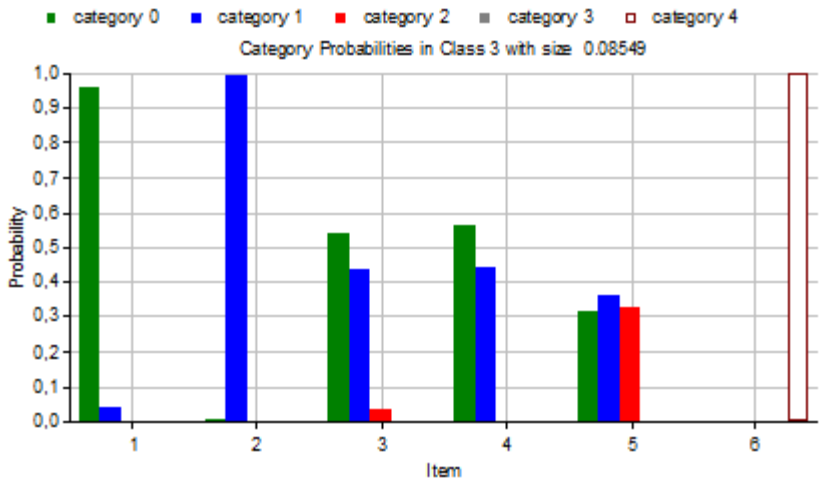

Figura 2.16: Probabilidades de cada categoría de las variables en la clase 3

1: Recibe remuneración por su trabajo: ***category 0:*** Si; **category 1:**No.
2: Cotiza a la Seguridad Social: **category 0**: Por todas las horas que trabajan; **category 1**: Cotizan pero no por todas las horas; **category 2**: Están exentas; **category 3**: No cotizan pero deberían.
3: Tipo de contrato: **category 0**: Si; **category 1**: Acuerdo verbal: **category 2:** No.
4: Tarjeta Seguridad Social: **category 0**: Si; **category 1**: No.
5: Quien cotiza a la Seguridad Social: **category 0**: El empleador; **category 1**: El empleador y la trabajadora; **category 2**: La trabajadora; **category 3**: No cotizan.
6:Porque no cotiza a la Seguridad Social: **category 0**: No tiene obligación; **category 1**: No le interesa o no le compensa; **category 2:** No sabe si tiene que cotizar; **category 3**: Otra explicación/Si cotizan.

A la vista de estos resultados observamos que los rasgos que más separan a las clases es que en la primera y la tercera son mujeres que cotizan a la Seguridad Social la primera por todas sus horas y con contratos y la tercera por menos horas y con acuerdos verbales y en la segunda ninguna de ellas cotiza a la Seguridad Social. Quedando así constituida la variable con sus tres categorías.

FASE II: Búsqueda del mejor predictor

En la primera etapa hemos conseguido una variable latente con tres clases que es la que en el estudio consideraremos como variable respuesta con tres categorías. A parte de esto tenemos las cinco variables que ya hemos comentado en apartados anteriores como variables predictoras.

Para encontrar el mejor predictor realizamos tablas cruzadas con la variable latente y cada una de las predictoras. En cada una de estas tablas calculamos el coeficiente de predictividad, el índice de Catanova y el p-valor asociado a este índice.

Por ejemplo, mostramos la tabla que cruza la variable latente con la variable modalidad de empleo doméstico, que es la siguiente:

	Internas	Externas fijas	Externas por horas
Clase 1	111	309	282
Clase 2	51	66	271
Clase 3	12	20	48

Tabla 2.11: tabla de contingencia que cruza la variable respuesta y el predictor modalidad de empleo

Ahora calculamos su coeficiente de predictividad. Se calcula aplicando la fórmula vista en el apartado de metodología:

$$\tau = \frac{\sum_i \sum_j [\left(\frac{p_{ij}}{p_{.j}}\right) - p_{i.}]^2}{1 - \sum_i p_{i.}^2} = 0{,}0705458$$

Podemos calcular con este valor el índice de Catanova:

$$C = (n-1)(I-1)\tau = (1170-1)\cdot(3-1)\cdot 0{,}0705458 = 164{,}9361$$

Este índice sigue una distribución chi-cuadrado con $(I-1)\cdot(J-1) = 2\cdot 2 = 4$ grados de libertad. A partir de esto, con un valor experimental y sus grados de libertad podemos obtener el p-valor para ver si es significativo. En este caso, por ejemplo, es $1{,}28\cdot 10^{-34}$ que es claramente significativo.

Este proceso se puede repetir con todas y cada una de las variables predictoras obteniendo la siguiente tabla:

Variable	τ	C	g. l.	p-valor
Estado Civil	0,00516556	12,0770793	8	0,14779779
Estudios	0,0214458	50,1402804	14	5,7846E-06
Edad	0,0150925	35,286265	8	2,3711E-05
Tiempo dedicacion	0,00454107	10,6170217	8	0,2243574
Modalidad empleo	**0,0705458**	**164,93608**	**4**	**1,2767E-34**

Tabla 2.12: Coeficientes de predictividad e Índices de Catanova de todos los predictores

Como vemos en la tabla solo tres de los predictores podrían ser escogidos como el mejor por ser significativos: estudios, edad y modalidad de empleo.

De todos ellos se selecciona como mejor predictor aquel con un mayor índice de predictividad. Si nos fijamos en los coeficientes de predictividad, el más alto es el de la modalidad de empleo. Seleccionaremos pues esta variable para segmentar. Esta segmentación se realiza en la fase tercera que veremos ahora.

FASE III: Análisis no simétrico de correspondencias

Para la segmentar tenemos que hacer, previamente, un análisis no simétrico de correspondencia entre la variable respuesta y el predictor seleccionado y clasificar las categorías del predictor en fuertes por la derecha y por la izquierda y débiles según las ideas de Siciliano y Mola que ya hemos visto en la metodología:

$$\begin{cases} \varphi_{j1} \geq 1 \text{ categorías fuerte por la derecha} \\ |\varphi_{j1}| < 1 \text{ categorías débiles} \\ \varphi_{j1} \leq -1 \text{ categorías fuerte por la izquierda} \end{cases}$$

Al hacer el análisis no simétrico de correspondencias entre la variable respuesta y nuestro predictor seleccionado, la modalidad de empleo, vemos que el primer eje tiene una inercia explicada del 99,94%. Nos interesa mirar esta inercia puesto que los valores que observamos para hacer la clasificación son los del primer eje.

Veamos ahora las coordenadas estándar de cada categoría:

Categoría	Etiqueta	Eje 1	Eje 2	Eje 3
Interna	Int	-0.2761	2.3765	-1
Externa-Fija	ExtF	-1.2797	-0.5695	-1
Externa por horas	EpH	0.9210	-0.3138	-1

Tabla 2.13: Puntuaciones de cada categoría del predictor seleccionado

Además de esto, el análisis no simétrico nos proporciona también un biplot que podemos analizar brevemente:

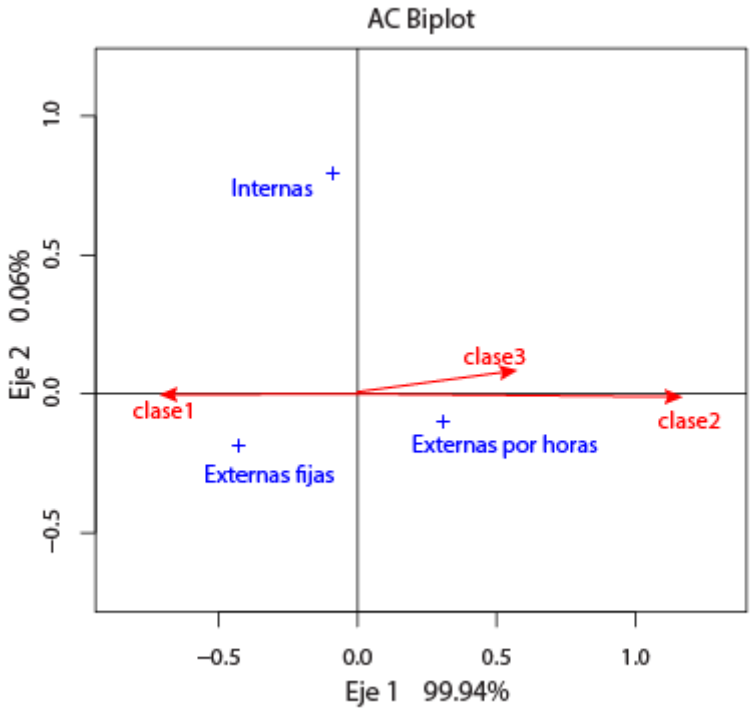

Figura 2.17: Representación gráfica del análisis de correspondencias no simétrico entre la variable respuesta y el predictor modalidad de empleo

Vemos que las empleadas de hogar externas fijas están muy presentes en la primera clase latente y muy poco presentes en las otras dos. Sin embargo las internas se ven representadas en las clases 1 y 3 pero muy vagamente. En cuanto a las externas por horas se ven mejor representadas en la clase 2, seguida por la clase 3 y no se presentan en la clase 1.

También apreciamos que las externas son las que más relación tienen, ya sean fijas o no, y que las internas son las más aisladas del grupo.

Por otro lado, vemos que las clases 2 y 3 están altamente correlacionadas mientras que la 1 presenta una correlación negativa con las anteriores. Las clases 1 y 2 son claramente de eje 1 aunque la clase 3 también lo es pero con menor proporción.

Tras haber analizado brevemente nuestro biplot, podemos seguir con el algoritmo. Ahora procedemos a clasificar las categorías según la clasificación propuesta anteriormente:

Categoría	φ_{j1}	Tipo de categoría
Interna	-0.2761	Débil
Externa-Fija	-1.2797	Fuerte por la izquierda
Externa por horas	0.9210	Débil

Tabla 2.14: Clasificación de las categorías del predictor

FASE IV: Segmentacion

Recopilando los resultados obtenidos anteriormente podemos hacer ya nuestra primera segmentación. Como hemos visto tenemos dos categorías débiles y una fuerte por la izquierda, por tanto, tendremos dos segmentos. Mostramos en la siguiente figura el primer nivel de nuestro árbol:

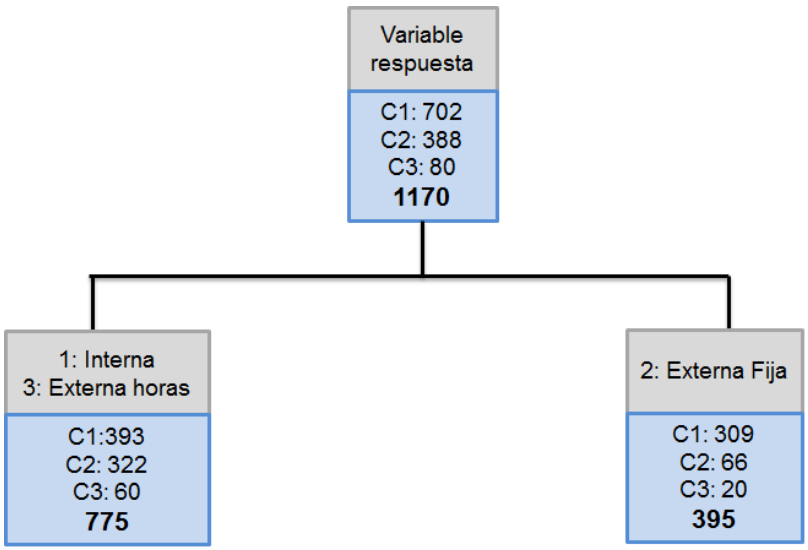

Figura 2.18: Primer nivel del árbol ternario de segmentación

Como vemos en el primer nodo tenemos las 1170 mujeres a las que se ha encuestado, de ellas 702 pertenecen a la primera clase latente, 388 a la segunda y 80 a la tercera.

En la primera segmentación, como ya hemos dicho, tenemos por un lado las débiles y por otro la fuerte por la izquierda, en este caso no había ninguna fuerte por la derecha.

Las categorías débiles (categorías interna y externa por horas de la variable modalidad de empleo) tienen un total de 775 mujeres, es decir, aproximadamente el 66% del total de las cuales poco más de la mitad pertenecen a la primera clase, más del 40% pertenecen a la segunda clase y finalmente menos del 10% pertenecen a la tercera clase.

La categoría fuerte por la izquierda (categoría externa fija de la variable predictora) cuenta con 395 encuestadas, es decir, casi el 34% del total. De estas 395, más del 78%, pertenece a la primera clase, casi el 17% a la segunda y el 5% a la tercera.

Estos resultados son lógicos puesto que ya sabemos que el 60% de las encuestadas pertenecen a la primera clase, un 33% a la segunda y solo el 7% del total pertenecen a la tercera clase.

Como ya sabíamos, las categorías fuertes tienen menor porcentajes de encuestadas ya que son aquellas que presentan capacidades predictivas muy altas lo que quiere decir que la probabilidad de conocer la respuesta conocidas las categorías de los bloques fuertes es muy alta. Esto no suele ser muy habitual, por ello se espera que el colectivo sea más pequeño.

Podemos ahora continuar con el algoritmo de segmentación para obtener las siguientes ramificaciones.

FASE V: REPETIR FASES II-IV.

En esta fase se repite en cada nodo las fases de la segunda a la cuarta (selección del mejor predictor, agrupación de categoría de modo ternario y segmentación) salvo que el nodo tenga menos del 10% de las encuestadas, que no es nuestro caso aún. Declararemos pues que estamos ante un nodo terminal si se cumple, como ya hemos dicho, una poca densidad de encuesta o si no es posible seguir segmentando puesto que no encontremos ningún predictor significativo.

Como aún no podemos declarar ninguno de los dos nodos que hemos obtenidos como terminales podemos seguir trabajando, nos vamos a centrar por ejemplo en el nodo formado por las categorías débiles con 775 mujeres.

Fase II

Realizamos las tablas cruzadas entre nuestra variable respuesta y cada uno de los predictores como hemos dicho antes solo que en este caso no entrarán en estudio las mujeres que trabajan externas fijas, calculamos en cada una su coeficiente de predictividad, su índice de Catanova, sus grados de libertad y su p-valor como ya sabemos para obtener el mejor predictor si lo hubiera. Recopilando todos los cálculos obtenemos la siguiente tabla:

Variable	τ	C	g. l.	p-valor
Estado Civil	0,00358728	5,55310423	8	0,69714966
Estudios	**0,02316145**	**35,8539214**	**14**	**0,00109729**
Edad	0,01307937	20,246871	8	0,00944134
Tiempo dedicación	0,00614506	9,512559	8	0,30091629
Modalidad empleo	0,01650496	25,5496764	2	2,8311E-06

Tabla 2.15: Coeficientes de predictividad e índices de Catanova de todos los predictores

En este caso tenemos tres predictores significativos. Como vemos hemos vuelto a incluir la modalidad de empleo aunque ya hemos segmentado por este predictor, no obstante en algún momento puede ser que lo obtengamos nuevamente como mejor predictor al cambiar las circunstancias en las que nos encontramos.

Vemos que el mejor predictor (el más alto) es el referente a la variable estudios así que será la que usaremos para segmentar.

Fase III

Realizamos nuevamente un análisis no simétrico de correspondencias entre la variable respuesta y nuestro predictor: el nivel de estudios. En este caso, vemos que el primer eje tiene una inercia explicada del 79,64% que ya es menor que en el caso anterior. Mirando las coordenadas estándares del primer eje obtenemos la siguiente clasificación:

Categoría	φ_{j1}	Tipo de categoría
Sin estudios	1.7245	Fuerte por la derecha
Primaria	-0.0241	Débil
Secundaria	0.1085	Débil
Formación profesional	-1.3071	Fuerte por la izquierda
Diplomado	-2.4367	Fuerte por la izquierda
Licenciado	3.7006	Fuerte por la derecha
Máster y doctorado	-5.2128	Fuerte por la izquierda
Otros	3.4093	Fuerte por la derecha

Tabla 2.16: Clasificación de las categorías del predictor

Al realizar este análisis, obtenemos también igual que antes un biplot, cuya interpretación es similar a la anterior.

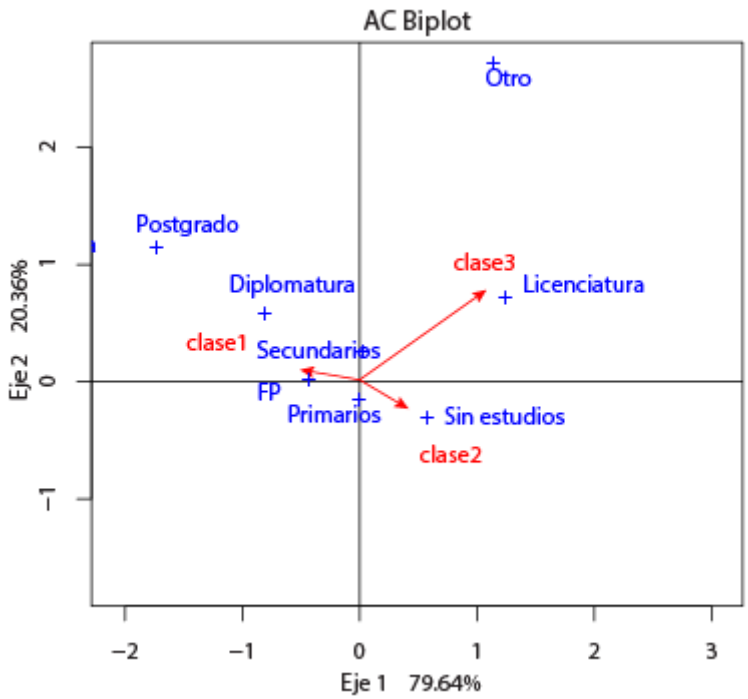

Figura 2.19: Representación gráfica del análisis de correspondencias no simétrico entre la variable respuesta y el predictor nivel de estudios

Vemos que las encuestadas cuyo nivel de estudio es de máster o doctorado, las diplomadas y las de formación profesional (categorías fuertes por la izquierda) son de clase 1. Las de primaria y secundaria (débiles) son las peores representadas. Las que no tienen estudios son de clase dos y finalmente las licenciadas o las que tienen otro nivel de estudios son de clase 3.

La clase 1 es claramente de eje uno. Por otro lado, las clases 2 y 3 no están correlacionadas mientras que la 1 está correlacionada negativamente con la 2 y la 3.

Finalmente podemos ver que las que menor nivel de estudios tienen (sin estudios, primaria, secundaria y formación profesional) son las que más relacionadas están.

Podemos ahora pasar al siguiente paso y hacer la segmentación.

Fase IV

Podemos ahora crear el segundo nivel de una de las ramas que teníamos que eran las débiles. Tenemos ahora según lo obtenido anteriormente tres ramificaciones para cada uno de los tipos de categorías (en este caso sí se nos han presentado los tres). El árbol en este punto tiene la siguiente forma:

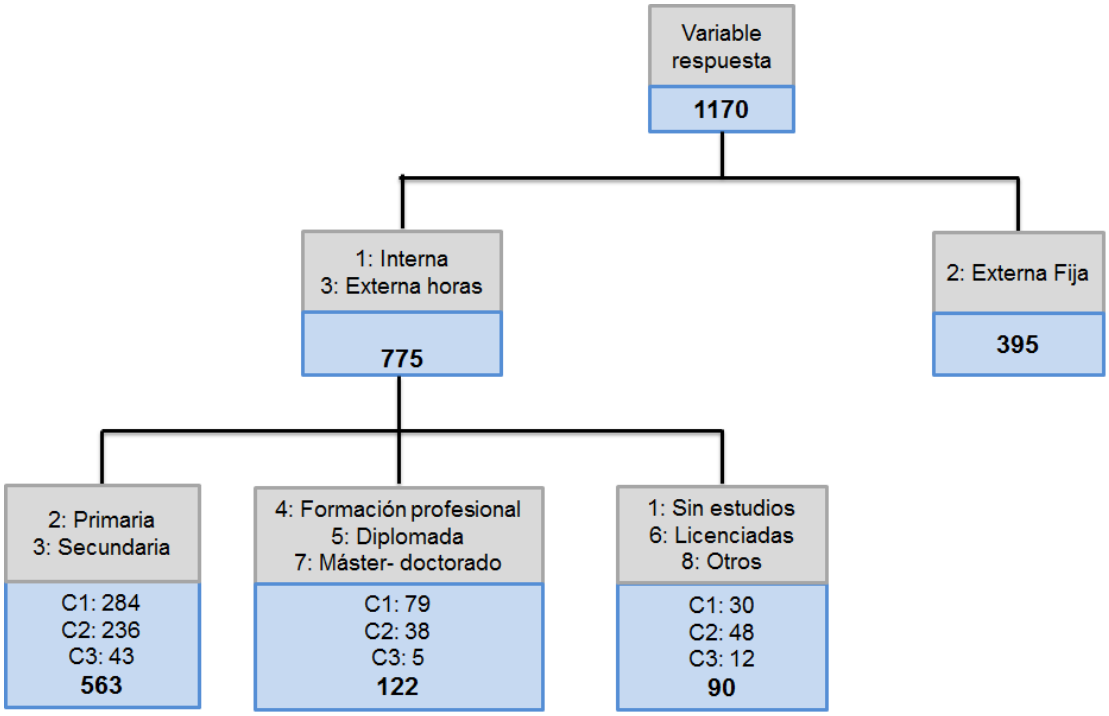

Figura 2.20: Primera parte del segundo nivel del árbol ternario de segmentación

Como ya hemos dicho anteriormente en este caso tenemos tres ramificaciones.

Las categorías débiles (categorías primaria y secundaria de la variable nivel de estudios) tienen un total de 563 mujeres, es decir, el 48% del total y aproximadamente el 73% de las mujeres internas y externas por horas, de las cuales poco más de la mitad pertenecen a la primera clase, más del 40% pertenecen a la segunda clase y finalmente menos del 10% pertenecen a la tercera clase.

Las categorías fuertes por la izquierda (categorías formación profesional, diplomadas y máster o doctorado de la variable nivel de estudios) cuentan con 122 encuestadas, es decir, más del 10% del total y del 15% de las mujeres internas y externas por horas. De estas 122 más del 64% pertenece a la primera clase, más del 31% a la segunda y el 4% a la tercera.

Las categorías fuertes por la derecha (categorías: sin estudios, licenciadas y otras, de la variable nivel de estudios) cuentan con 90 mujeres, es decir, menos del 10% del total por lo que el nodo será terminal y casi el 12% de las mujeres internas y externas por horas. De estas 90 más del 33% pertenece a la primera clase, más del 53% a la segunda y más del 13% a la tercera.

Ahora se puede seguir realizando el proceso en todos los nodos que sean no terminales. Recordemos que ya tenemos un nodo no terminal que cuenta con menos del 10% de la muestra total.

Árbol definitivo

Siguiendo el proceso hasta obtener que todos los nodos son terminales obtenemos el siguiente árbol:

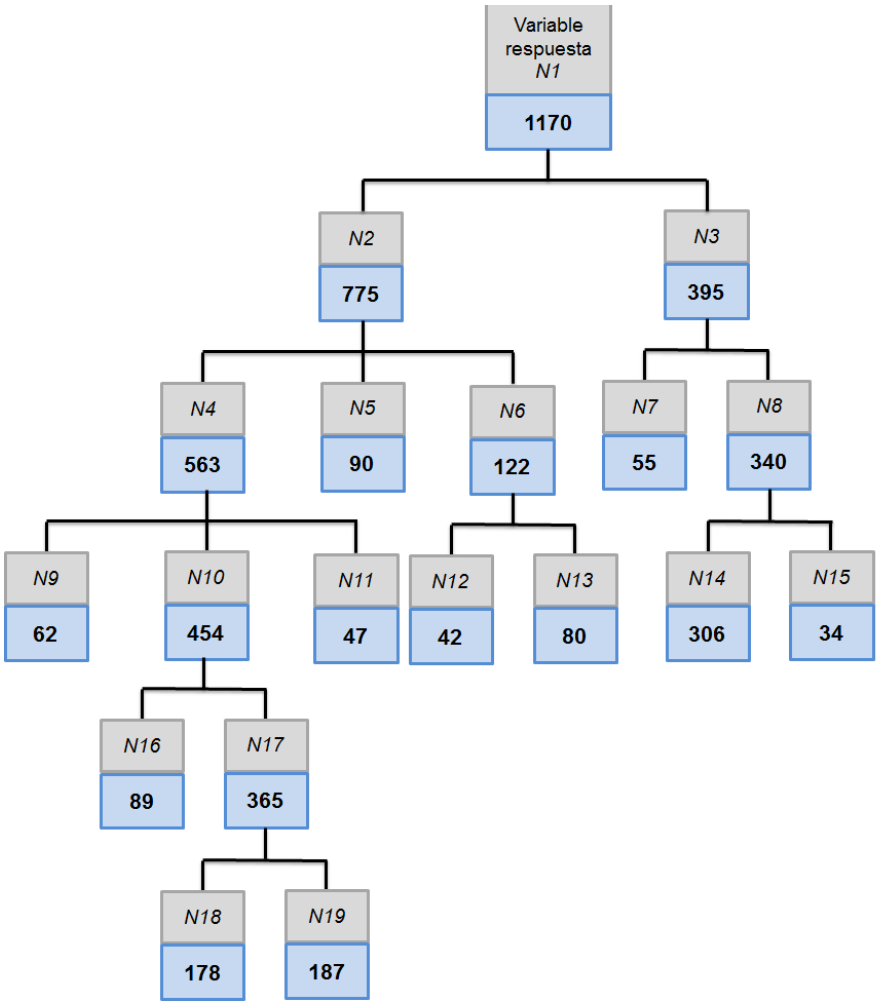

Figura 2.21: Árbol ternario de segmentación final

Estudiemos ahora los nodos que tenemos detenidamente. Empezaremos por los que no son terminales para ver cómo se sigue segmentando en cada uno.

Nodo	Mejor Predictor	Categorías Débiles	Categorías fuertes por la izquierda	Categorías fuertes por la derecha
N1	Modalidad empleo	1: Interna 3: Externa Fija	2: Externa por horas	No hay
N2	Nivel estudios	2: Primaria 3: Secundaria	4: Formación profesional 5: Diplomada 7: Máster-Doctorado	1: Sin estudios 6: Licenciada 8: Otro
N3	Edad	3: 30-45 años 4: 46-55 años 5: 56-64 años	No hay	1: <20 años 2: 20-29 años
N4	Edad	3: 30-45 años 4: 46-55 años	5: 56-64 años	1: <20 años 2: 20-29 años
N6	Modalidad empleo	3: Externa por horas	1: Interna	No hay
N8	Estado Civil	1: Soltera 2: Casada 3: Separada-Divorciada	No hay	4: Viuda 5: Pareja de hecho
N10	Modalidad empleo	3: Externa por horas	1: Interna	No hay
N17	Edad	4: 46-55 años	3: 30-45 años	No hay

Tabla 2.17: Descripción de los nodos no terminales del árbol final

Expliquemos esta tabla por columnas: la primera muestra el nodo en cuestión, la segunda el mejor predictor para segmentar y las otras tres el tipo de categorías según la agrupación de Siciliano y Mola para segmentar. Veamos ahora los nodos que sí son terminales y por qué no se segmenta en cada uno de ellos.

Nodo	% casos	Número encuestadas			% de cada nodo			Motivo Nodo Terminal
		Clase 1	Clase 2	Clase 3	Clase 1	Clase 2	Clase 3	
N5	7,69	30	48	12	33,33	53,33	13,33	<10%
N7	4,7	32	17	6	58,18	30,91	10,91	<10%
N9	5,3	18	33	11	29,03	53,23	17,74	<10%
N11	4,02	29	17	1	61,70	36,17	2,13	<10%
N12	3,59	32	7	3	76,19	16,67	7,14	<10%
N13	6,84	47	31	2	58,75	38,75	2,50	<10%
N14	26,15	257	36	13	83,99	11,76	4,25	NPS
N15	2,91	20	13	1	58,82	38,24	2,94	<10%
N16	7,61	63	23	3	70,79	25,84	3,37	<10%
N18	15,21	89	68	21	50,00	38,20	11,80	NPS
N19	15,98	85	95	7	45,45	50,80	3,74	NPS

Tabla 2.18: Descripción de los nodos terminales del árbol final

En esta tabla se muestra en la primera columna el número del nodo en estudio. Para cada nodo se tiene en la segunda columna el porcentaje de encuestadas sobre el total. En las tres siguientes columnas se representa el número de mujeres que pertenecen a cada clase en cada nodo y en las tres siguientes el peso (o porcentaje) que tienen cada clase en cada nodo. En la última columna se expone el motivo por el cual estos nodos son terminales donde hemos usado: <10% para expresar que el nodo no alcanza el 10% del tamaño muestral total y NPS si el nodo no presentaba más predictores significativos. Ahora vamos a ver, en la siguiente tabla la clase mayoritaria de cada nodo y su perfil.

Nodo	Clase mayoritaria	Perfil del nodo
N5	Clase 2	Trabajadoras internas o externas por horas, sin estudios, licenciadas o con otro nivel de estudios (diferente a los que tiene la categoría)
N7	Clase 1	Trabajadoras externas fijas menores de 30 años
N9	Clase 2	Trabajadoras internas o externas por horas con estudios primarios o secundarios menores de 30 años
N11	Clase 1	Trabajadoras internas o externas por horas con estudios primarios o secundarios mayores de 55 años
N12	Clase 1	Trabajadoras internas con formación profesional, diplomadas o con máster o doctorado
N13	Clase 1	Trabajadoras externas por horas con formación profesional, diplomadas o con máster o doctorado
N14	Clase 1	Trabajadoras externas fijas mayores de 29 años solteras, casadas o separadas
N15	Clase 1	Trabajadoras externas fijas mayores de 29 años viudas o con pareja de hecho
N16	Clase 1	Trabajadoras internas mayores de 29 años con estudios primarios o secundarios
N18	Clase 1	Trabajadoras externa por horas de 30 a 45 años con estudios primarios o secundarios
N19	Clase 2	Trabajadoras externa por horas de 46 a 55 años con estudios primarios o secundarios

Tabla 2.19: Perfiles de los nodos terminales del árbol final

Mapa Factorial

Para finalizar este estudio queremos ver la relación de las clases latentes con estos nodos terminales. Para esto vamos a crear la tabla cruzada de las clases con una nueva variable en la cual cada categoría será uno de los nodos terminales. Hecho esto realizaremos nuevamente un análisis de correspondencias no simétrico y

analizaremos el biplot que nos resulta como ya hemos hecho anteriormente. Mostramos inicialmente la tabla de contingencia que vamos a usar:

	N5	N7	N9	N11	N12	N13	N14	N15	N16	N18	N19
Clase 1	30	32	18	29	32	47	257	20	63	89	85
Clase 2	48	17	33	17	7	31	36	13	23	68	95
Clase 3	12	6	11	1	3	2	13	1	3	21	7

Tabla 2.20: Nodos terminales del árbol final

Mostramos ahora el mapa factorial. Vemos que hemos conseguido una magnifica absorción de inercia:

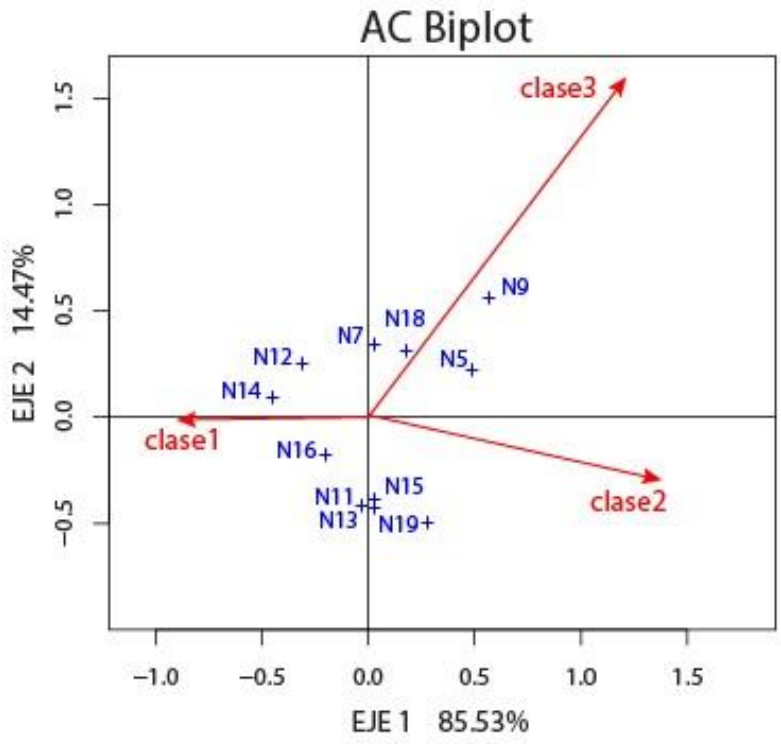

Figura 2.22: Mapa factorial de la variable respuesta con los nodos terminales

Los ejes factoriales muestran las diferencias entre nuestras clases: las encuestadas que cotizan por todas sus horas, tienen contrato de trabajo y tarjeta de la Seguridad Social como titulares y las que no cotizan y no tienen contrato ni tarjeta de la Seguridad Social son de eje 1 mientras que las que cotizan pero por menos horas de las que trabajan, que tienen acuerdos verbales con sus empleadores y no tienen tarjeta de la Seguridad Social como titulares son de eje 2.

Podemos decir que presentan un perfil similar: las trabajadoras que sean externas por horas con formación profesional, diplomadas o con máster o doctorado

con las que sean internas o externas por horas con estudios primarios o secundarios mayores de 55 años y con las externas fijas mayores de 29 años viudas o con pareja de hecho.

Otro grupo estaría constituido por las trabajadoras externas fijas con menos de 30 años y las trabajadoras externas por horas de 30 a 45 años con estudios primarios o secundarios.

Y por otro lado, forman un grupo las trabajadoras internas con formación profesional, diplomadas o con máster o doctorado y las trabajadoras externas fijas mayores de 29 años solteras, casadas o separadas.

Finalmente, vemos que las trabajadoras internas o externas por horas con estudios primarios o secundarios menores de 30 años son las que más se alejan de todas las demás.

También podemos decir que de estos tres grupos que hemos expuesto aquí, el primero son principalmente mujeres que no cotizan, no tienen contrato ni tarjeta de la Seguridad Social como titulares junto a las trabajadoras externa por horas de 46 a 55 años con estudios primarios o secundarios.

Por otro lado el segundo grupo junto al último, el más alejado (aunque este en mayor medida), son generalmente las que cotizan pero por menos horas de las que trabajan, teniendo un acuerdo verbal con sus empleadores y sin tarjeta de la Seguridad Social.

El grupo restante de los expuestos anteriormente, el segundo, junto a las trabajadoras internas mayores de 29 años con estudios primarios o secundarios son trabajadoras en su mayoría que cotizan por todas y cada una de las horas que trabajan, tienen contrato de trabajo y tarjeta de la Seguridad Social.

De entre todos los nodos destacados nos falta decir que las trabajadoras internas o externas por horas, sin estudios, licenciadas o con otro nivel de estudios son una mezcla entre las trabajadoras que no cotizan y están sin contrato y las que sí cotizan pero no por todas las horas que trabajan y tienen un acuerdo verbal.

2.7. Estudio comparativo de los algoritmos Chaid y Taid

Aunque ya los hemos analizado por separado, en este punto lo que vamos a confrontar los resultados obtenidos con el algoritmo Chaid con los resultados

obtenidos con el algoritmo Taid poniendo de manifiesta clasificación obtenida por cada uno de los métodos de las trabajadoras. Este análisis se realizará atendiendo a cada una de las variables respuestas:

Variable respuesta: *¿Tiene contrato de trabajo?*

En el caso del algoritmo Chaid hemos visto una clara predilección a que todos los grupos formados tengan, en mayor o menor porcentaje contrato de trabajo. Mientras que, en el algoritmo Taid los tres grupos claramente diferenciados, es decir, tienen contrato de trabajo las trabajadoras internas mayores de 29 años con estudios primarios o secundarios, externas fijas con menos de 30 años y externas por horas de 30 a 45 años con estudios primarios o secundarios, tienen un acuerdo verbal las trabajadoras internas con formación profesional, diplomadas o con máster o doctorado, externas fijas mayores de 29 años solteras, casadas o separadas e internas o externas por horas con estudios primarios o secundarios menores de 30 años y no tienen contrato las externas por horas con formación profesional, diplomadas o con máster o doctorado, internas o externas por horas con estudios primarios o secundarios mayores de 55 años, externas fijas mayores de 29 años viudas o con pareja de hecho y externas por horas de 46 a 55 años con estudios primarios o secundarios.

Variable respuesta: *¿Cotiza usted a la Seguridad Social?*

Como resultado del algoritmo Chaid hemos visto que las trabajadoras externas por horas que llevan en el servicio doméstico de 1 a 3 años no cotizan mientras que en el Taid las que no cotizan son las externas por horas con formación profesional, diplomadas o con máster o doctorado, internas o externas por horas con estudios primarios o secundarios mayores de 55 años, externas fijas mayores de 29 años viudas o con pareja de hecho y externas por horas de 46 a 55 años con estudios primarios o secundarios. Por otro lado, el Chaid nos lleva a afirmar que todas las demás cotizan por todas las horas que trabajan, no habiendo una mayoría que coticen por menos horas de las que trabajan en ningún caso. Sin embargo, el Taid sí hace esta diferenciación, clasifica como trabajadoras que cotizan por todas sus horas solo a las internas mayores de 29 años con estudios primarios o secundarios, externas fijas con menos de 30 años y externas por horas de 30 a 45 años con estudios primarios o secundarios y nos dice que cotizan por menos horas de las que trabajan las internas con formación profesional, diplomadas o con máster o doctorado, externas fijas mayores de 29 años solteras, casadas o separadas e internas o externas por horas con estudios primarios o secundarios menores de 30 años

Variable respuesta: *¿Recibe remuneración por su trabajo?*

En los dos casos se quedan todas las trabajadoras clasificadas en la categoría de sí reciben remuneración. Esto era algo evidente ya que casi el 96% de las encuestadas la recibe.

Variable respuesta: *¿Quién cotiza a la Seguridad Social?*

En este caso, hemos visto que en el Chaid predominaba la cotización a cargo de los empleadores. Sin embargo en el caso del Taid vemos que de las trabajadoras que cotizan las internas mayores de 29 años con estudios primarios o secundarios, externas fijas con menos de 30 años y externas por horas de 30 a 45 años con estudios primarios o secundarios cotiza su empleador mientras que en el otro caso, el de las internas con formación profesional, diplomadas o con máster o doctorado, externas fijas mayores de 29 años solteras, casadas o separadas e internas o externas por horas con estudios primarios o secundarios menores de 30 años son ellas quienes cotizan en su mayoría, seguido de cerca por las que cotizan a medias con su empleador y por las que cotiza el empleador.

Variable respuesta: *¿Por qué no cotiza a la Seguridad Social?*

En el caso del Chaid vemos que las externas fijas menores de 30 años no saben si tienen que hacerlo y en el resto de casos predomina el desinterés o no les compensa. Sin embargo, en el Taid de las que no cotizan que son las externas por horas con formación profesional, diplomadas o con máster o doctorado, internas o externas por horas con estudios primarios o secundarios mayores de 55 años, externas fijas mayores de 29 años viudas o con pareja de hecho y externas por horas de 46 a 55 años con estudios primarios o secundarios no lo hacen por desinterés o porque no les compensa.

Variable respuesta: *¿Tiene tarjeta propia de la Seguridad Social?*

En el caso del Chaid afirmamos que sí tienen tarjeta las mayores de 30 años o las menores de 30 años pero con un nivel de estudios secundarios mientras que en el Taid las que sí tienen tarjeta son las internas mayores de 29 años con estudios primarios o secundarios, externas fijas con menos de 30 años y externas por horas de 30 a 45 años con estudios primarios o secundarios. Por otro lado, en el Chaid vemos que las trabajadoras menores de 30 años con un nivel de estudios diferente a la secundaria (sin estudios, primarios, formación profesional, diplomadas, licenciadas, máster u otro) están incluidas en la tarjeta de la Seguridad Social de algún familiar o

directamente no tienen tarjeta. Como resultado del algoritmo Taid, este grupo lo forman las mujeres externas por horas con formación profesional, diplomadas o con máster o doctorado, internas o externas por horas con estudios primarios o secundarios mayores de 55 años, externas fijas mayores de 29 años viudas o con pareja de hecho, externas por horas de 46 a 55 años con estudios primarios o secundarios internas con formación profesional, diplomadas o con máster o doctorado, externas fijas mayores de 29 años solteras, casadas o separadas e internas o externas por horas con estudios primarios o secundarios menores de 30 años.

Resumiendo:

Hemos visto una clarísima diferencia entre los algoritmo Taid y Chaid. En el caso del Chaid las segmentaciones son mucho más precisas, los grupos son mucho más grandes, con un mayor número de encuestadas. Sin embargo en el caso del Taid se ha realizado una segmentación mucho más refinada de la población, con más ramificaciones, podemos decir que las suficientes para dividir bien a la población y que el árbol ternario resultante no fuese demasiado complejo de analizar.

Por otro lado, hemos visto que en el caso del algoritmo Taid no utiliza en ningún caso el predictor: tiempo que las empleadas llevan dedicándose al servicio doméstico. Sin embargo, el algoritmo Chaid sí utiliza todos sus predictores para alguna segmentación en alguna de las variables respuestas (esto era más fácil ya que son seis análisis distintos).

Finalmente, en nuestra opinión creemos que el algoritmo Taid es mucho más completo que el Chaid porque, como ya hemos visto, puede utilizar varias variables respuesta a la vez añadiendo solo un análisis de clases latentes, no verifica la paradoja de Simpson, como sí lo hace el Chaid, y, finalmente, como acabamos de ver, presenta una segmentación mucho más completa y eficiente.

Conclusiones

- 1.- A pesar de que el algoritmo CHAID presenta importantes limitaciones, entre las que cabe destacar que trabaja con tablas colapsadas y no captura el papel asimétrico de las variables en estudio, sigue siendo el método más utilizado en la bibliografía, en todas las ramas de la Ciencia.

- 2.- Los algoritmos DAVILA y DORADO, que desde el punto de vista teórico suponen un gran avance ya que superan varias de las limitaciones del CHAID, no se aplican en la práctica, posiblemente por la falta de software especifico para llevarlas a cabo.

- 3.- El algoritmo TAID, que supera todas las limitaciones de los anteriores ya que captura el papel asimétrico de las variables y permite trabajar con muchas variables respuesta simultáneamente, tampoco se ha aplicado en la práctica de forma generalizada, aunque existen aplicaciones puntuales. En este caso existe software, pero no es de uso libre, lo cual puede justificar su uso más local.

- 4.- Se han encontrado importantes diferencias en la definición de los perfiles de mujeres trabajadoras que se dedican a la actividad del servicio doméstico, al aplicar el algoritmo CHAID y el algoritmo TAID, lo cual pone de manifiesto la necesidad de extender los nuevos métodos de segmentación en los diferentes campos de la ciencia y de desarrollar software específico que contribuya a su divulgación y uso.

Bibliografía

Alvarado, E., Luyando, J. R., & Téllez, R. (2012). Caracterización del consumidor de la carne de pollo en el área metropolitana de Monterrey. *Región y sociedad*, *24*(54), 175–199.

Anderson, R. J., & Landis, J. R. (1980). Catanova for multidimensional contingency tables: Nominal-scale response. *Communication in Statistics Theory and Methods*, *9*, 1191-1206.

Anderson, R. J., & Landis, J. R. (1982). Catanova for multidimensional contingency tables: Ordinal-scale response. *Communication in Statistics Theory and Methods*, *11*, 257-270.

Araya, C. (2010). *Modelos de clases latentes en tablas poco ocupadas: una contribuciónbasada en Bootstrap*. Universidad de Salamanca, Salamanca.

Ávila, C. A. (1996). *Una alternativa al análisis de segmentación basada en el análisis de hipótesis de independencia condicionada*. Universidad de Salamanca, Salamanca.

Beh, E. J. (2008). Simple correspondence analysis of Nominal-Ordinal contingency tables. *Journal of Applied Mathematics and Decision Sciences*, 1-17.

Bell, R. L., Meier, R. J., & Guyot, W. (2013). A factorial analysis of gender and rank on business school faculty's salaries as a gauge for dissatisfaction. *Business Studies Journal*, *5*.

Benzecri, J. P. (1973). L'analyse des correspondences. En *L'analyse des Données* (Vol. 2). París.

Bichler, A., Neumaier, A., & Hofmann, T. (2014). A tree-based statistical classification algorithm (CHAID) for identifying variables responsible for the occurrence of faecal indicator bacteria during waterworks operations. *Journal of Hydrology*, *519*, 909-917.

Cabo, G., González, A., Roces, P., & Muñoz, S. (2002). *La presencia de las Mujeres en el Empleo Irregular*. Madrid: Centro de Estudios Económicos Tomillo.

Camminatiello, I., D'Ambra, L., & Ragione, L. (2011). A study of instantaneous emissions through the descomposition of directional measures for three-way contingency tables with ordered categories. *Journal of Applied Sciences*, *11*(4), 693-699.

Castro, C. R. (2005). *Contribuciones a la detección y análisis de variables relevantes en tablas de contingencia multivariante*. Universidad de Salamanca, Salamanca.

Chessel, G., & Gimaret, C. (1999). Analyse non symétrique des correspondances. *Business & Economics*.

D'Ambra, A., & Crisci, A. (2014). The confidence ellipses in decomposition multiple non symmetric correspondence analysis. *Communications in Statistics-Theory and Methods*, *43*(6), 1209-1221.

D'Ambra, A., D'Ambra, L., & Pasquale, S. (2010). Descomposition of the Gray-Williams «tau» in main and interaction effects by Anova in three-way contingency table. *Australian & New Zealand Journal of Statistics*, *52*(1), 55-75.

D'Ambra, L., Beh, E. J., & Amenta, P. (2005). Catanova for two-way contingency tables with ordinal variables using orthogonal polynomials. *Communications in Statistics - Theory and Methods*, *34*(8), 1755-1769.

Demspter, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39*(1), 1-38.

Dorado, A. (1998). *Métodos de búsquedas de variables relevantes en análisis de segmentación: aportaciones desde una perspectiva multivariante*. Universidad de Salamanca, Salamanca.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals os Statistics*, *7*(1), 1-26.

Escobar, M. (1998). Las aplicaciones del análisis de segmentación: el procedimiento Chaid. *Metodología de Ciencias Sociales*, *1*, 13-49.

Escrivá, M. A. (2000). ¿Empleadas de por vida? Peruanas en el servicio doméstico de Barcelona. *Revista de Sociología*, *60*, 327-342.

García, J. M. (2014). Una posible nueva clase trabajadora de servicios: evidencias a partir de un análisis del mercado de trabajo español entre 1999 y 2008. *Cuadernos de Relaciones Laborales*, *32*(2).

Gini, C. W. (1912). *Variabilitá e mutabilitá. Contributo allo studio delle distribuzioni e relazioni statistiche, studi economico-giuridici della R*. Universitá di Cagliari, Cagliari.

González, P. (1982). Notas sobre la condición de la mujer trabajadora en España durante los tres primeros decenios del siglo XX. En *Actas de las Primeras Jornadas de Investigación Interdisciplinaria. Nuevas perspectivas sobre la mujer.* (Vol. 2, pp. 97-104). Madrid: Universidad Autónoma de Madrid.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross-classifications. *Journal of the American Statistics Associacion*, *49*, 732-764.

Haberman, S. J. (1979). *Analysis of qualitative data: new developments* (Vol. 1). New York: Academic Pres.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple test of significance. *Biometrika*, *75*, 800-802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Sacandinavian Journal of Statistics*, *6*, 65-70.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, *29*(2), 127-199.

Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: a practical guide. *Frontiers in psychology*, *4*.

Kiranga, J. M. (2013). *Migration trends in Mombasa district*. University of Nairobi, Nairobi.

Lanza, S. T., & Rhoades, B. L. (2013). Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment. *Prevention Science*, *14*(2), 157-168.

Lauro, N. C., & D'Ambra, L. (1984). *L'analyse non symétrique des correspondences* (Vol. III). Amsterdam: Data analysis and Informatics.

Legohérel, P., Hsu, C. H., & Daucé, B. (2015). Variety-seeking: Using the CHAID segmentation approach in analyzing the international traveler market. *Tourism Management*, *46*, 359-366.

Light, R. J., & Margolin, B. H. (1971). An analysis of variance for categorical data. *Journal of the American Statistics Associacion*, *66*, 534-544.

Light, R. J., & Margolin, B. H. (1974). An analysis of variance for categorical data II. Small samples comparisons with chi-square and other competitors. *Journal of the American Statistics Associacion*, *69*, 755-764.

Marcu, S. (2009). Inmigrantes rumanas en el servicio doméstico y de cuidados de la Comunidad de Madrid: Estudio cualitativo. *Estudios Geográficos*, *LXX*(267), 463-489.

Mays, S. (2014). A test of a recently devised method of estimating skeletal age at death using features of the adult acetabulum. *Journal of forensic sciences*, *59*(1), 184-187.

Ma, Y. Z., & Ma, A. M. (2011). Simpson's paradox and other reversals in basketball: examples from 2011 NBA playoffs. *International Journal of Sports Science and Engineering*, *5*(3), 145-154.

Morgan, J., & Sonquist, J. (1963). Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association*, *58*, 415-434.

Oberski, D. L., van Kollenburg, G. H., & Vermunt, J. K. (2013). A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, *7*(3), 267-279.

Olmus, H., & Erbas, S. (2014). Catanova method for determining of zero partial association structures in multidimensional contigency tables. *Journal of Science*, *27*(3), 953–963.

Patino, M. C., Vicente, M. P., & Galindo, M. P. (2011). Perfil multivariante de las mujeres empleadas en el servicio doméstico. *Cuaderno de Relaciones Laborales*, *29*(2), 393-416.

Perrons, D., Plomien, A., & Kilkey, M. (2010). Migration and uneven development within an enlarged European Union: Fathering, gender divisions and male migrant domestic services. *European Urban and Regional Studies*, *17*(2), 197-215.

Platzer, E. (2006). From private solutions to public responsibility and back again: the new domestic services in Sweden. *Gender & History*, *18*(2), 211–221.

Popescu, M. E., Andreica, M., & Micu, D. (2014). A method to improve economic performance evaluation using clasification tree models. *European Journal of Business and Social Sciences*, *3*(4), 249-256.

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. *International Journal of Computer Science Issues*, *7*(1), 1002-1144.

Ramírez, G. J. (1995). *Contribuciones al análisis de segmentación*. Universidad de Salamanca, Salamanca.

Rodríguez, P., & Turégano, S. (2014). Tourism consumption and social inequality in Spain. *Revista de Turismo y Patrimonio Cultural*, *12*(1), 29–51.

Salvo, S. I. (2002). *Contribuciones al análisis de modelos para variables cualitativas que contemplan variable respuesta*. Universidad de Salamanca, Salamanca.

Sepúlveda, R. (2003). *Contribuciones al análisis de clases latentes en presencia de dependencia local*. Universidad de Salamanca, Salamanca.

Siciliano, R., & Mola, F. (1997). Ternary classification trees: a factorial approach. En *Visualization of categorical data* (pp. 311-323). Academic Press.

Singh, B. (2004). Catanova for analysis of nominal data from repeated measures design. *Journal of the Indian Society of Agricultural Statistics*, *58*(3), 257–268.

Soldic-Aleksic, J. (2012). Combined approach of kohonen som and chaid decision tree model to clustering problem: A market segmentation example. *Journal of Economics and Engineering*, *3*(1).

Suh, E., & Alhaery, M. (2015). Predicting cross-gaming propensity using Chaid analysis. *UNLV Gaming Research & Review Journal*, *19*(1).

Takane, Y., & Jung, S. (2009). Tests of ignoring and eliminating in nonsymmetric correspondence analysis. *Advances in data analysis and classification*, *3*(3), 315–340.

Tellez, R., Mora, J. S., Martínez, M. Á., García, R., & García, J. A. (2012). Caracterización del consumidor de carne Bovina en la zona matropolitana del Valle de México. *Agrociencia*, *46*(1), 75–86.

Vassilikou, C. (2007). Inmigrant women in greece. A biographical study of domestic workers. *Gender, rovne prilezitosti, vyzkum*, *8*(1), 40-45.

Von Davier, M. (2001). Winmira (Versión 1.45). Alemania.

Zübeyir, B. (2015). Study of some demographic properties ingluencing the burnout levels of nurses in public hospitals by chaid analysis. *Journal of The Faculty of Economics and Administrative Sciences*, *5*.

Anexo I

El cuestionario

ENCUESTA A MUJERES
EMPLEADAS DE HOGAR

Estamos realizando una investigación cuyo objetivo es recabar, analizar, y difundir información sobre la situación de las mujeres trabajadoras salmantinas EMPLEADAS DE HOGAR.

El estudio nos permitirá comparar las características socioeconómicas de las trabajadoras en situación irregular y regular; identificar los principales perfiles socioeconómicos de las trabajadoras en situación irregular; analizar las causas y consecuencias sobre la vida personal y laboral de dicha actividad para las mujeres afectadas, y detectar sus principales demandas.

Los resultados se harán llegar hasta las autoridades pertinentes que son quienes deben proponer políticas tendentes a mejorar la situación de las mujeres en los distintos ámbitos.

Para cumplir el objetivo, es absolutamente imprescindible su colaboración: *¿Sería tan amable de responderme a unas preguntas relacionadas con su situación?*

La encuesta es absolutamente anónima.

ENCUESTA A MUJERES
EMPLEADAS DE HOGAR

1. ¿Cuál es su nacionalidad?

☐	Española
☐	De un país de la Unión Europea
☐	De un país de Europa del Este
☐	De un país de Latinoamérica
☐	De un país de África
☐	Resto del mundo

2. ¿Cuál es su Estado Civil?

☐	Soltera
☐	Casada
☐	Separada /Divorciada
☐	Viuda
☐	Pareja de Hecho

3. ¿Qué nivel de estudios tiene terminados?

☐	Sin estudios
☐	Primarios (Graduado Escolar)
☐	Secundarios (Bachillerato, COU, ESO)
☐	Formación Profesional
☐	Diplomado, Arquitecto Técnico o Ingeniero Técnico
☐	Licenciado, Arquitecto o Ingeniero Superior
☐	Estudios de Postgrado (Master, Doctorado) o especialización
☐	Otros estudios no reglados

4. ¿Cuántos hijos/as tiene usted? *(Escriba el número de cada)*

☐	Hijos
☐	Hijas

5. Número de habitantes que hay en su localidad

☐	Menos de 100 habitantes
☐	Entre 100 y 500 habitantes
☐	Entre 500 y 2000 habitantes
☐	Entre 2000 y 5000 habitantes
☐	Más de 5000 habitantes

Escribir el nombre de la localidad donde vive

6.- ¿Recibe remuneración por su trabajo?

☐	Si
☐	NO

7. ¿Cotiza usted a la Seguridad Social?

Si, por todas las horas que trabajo	Si, pero por menos horas de las que realmente trabajo	NO COTIZO porque estoy exenta	NO COTIZO, pero debería cotizar
☐	☐	☐	☐

8.- ¿Tiene firmado contrato de trabajo?

Si	Tengo un acuerdo verbal con mi empleador	NO tengo contrato
☐	☐	☐

CUESTIONARIO GENERAL

CARACTERÍSTICAS SOCIOECONÓMICAS

9. ¿Qué edad tiene?

☐	Menos de 20 años
☐	De 20 a 29 años
☐	De 30 a 45 años
☐	De 46 a 55 años
☐	De 56 a 64 años

10. ¿Dónde vive actualmente?

☐	En mi propia vivienda
☐	En una vivienda alquilada
☐	En una habitación alquilada
☐	En una pensión
☐	En casa de familiares/amigos
☐	En casa de mi empleador
☐	En residencia, piso compartido, o similar

11. ¿Cuántas personas viven con usted? ……

12. Cuántas personas dependen económicamente de usted? ……

13. ¿Cuántas personas de su hogar enfermas, o con discapacidad, dependen de su atención y cuidado? *Indicar el número por tramo de edad*

☐	Ninguna
☐	Menores de 15 años
☐	De 16 a 64 años

☐ De 65 y más años

14. Indique quién realiza las tareas del hogar *(Puede responder varias a la vez)*

☐ Yo
☐ Mi marido (o pareja)
☐ Mis hijos/as
☐ Otros familiares
☐ Una empleada de hogar

15. ¿Quiénes tienen trabajo remunerado en su familia? *(Puede responder varias a la vez)*

☐ Mi marido (pareja)
☐ Yo
☐ Mis hijos (alguno de ellos, o todos)
☐ Mis padres (alguno de ellos, o ambos)
☐ Mis hermanos (alguno de ellos, o todos)
☐ Otros familiares
☐ Ninguno

16. ¿En cuál de los siguientes tramos situaría los ingresos netos mensuales percibidos exclusivamente por su trabajo?

☐ Menos de 300 €
☐ De 301 a 600 €
☐ De 601 a 1000 €
☐ De 1000 a 1500 €
☐ Más de 1500 €
☐ NS/NC

17. Aproximadamente ¿cuáles son los ingresos mensuales de su familia?

☐ Menos de 300 € al mes
☐ Entre 301 y 600 € al mes
☐ Entre 601 y 1.200 € al mes
☐ Entre 1.201 y 2.400 € al mes
☐ Más de 2.400 € al mes
☐ Lo desconozco

18.¿Quién le ha ayudado cuando ha estado sin trabajo o en perentoria situación económica?

☐ Nunca he necesitado ayuda
☐ Familiares o amigos
☐ Instituciones públicas
☐ ONGs o parroquias
☐ Nadie

19. ¿Qué tipo de tarjeta sanitaria de la Seguridad Social tiene?

☐ Tarjeta sanitaria propia como titular
☐ Estoy incluida en la tarjeta de un familiar
☐ No tengo tarjeta propia ni estoy incluida en ninguna
☐ NS/NC

20. ¿Pertenece a alguna de estas asociaciones?

- [ ] Asociación de mujeres rurales
- [ ] Asociación de mujeres empresarias
- [ ] Asociación de empleadas de hogar
- [ ] Asociación de inmigrantes
- [ ] Otra (especificar)

TEMAS LABORALES

21. ¿Cuántos años ha estado trabajando? *(Sin tener en cuenta los periodos de tiempo que no ha trabajado)*

22. ¿Ha trabajado alguna vez antes de estar en su actual empleo?

- [ ] Si
- [ ] No *(ir a la pregunta 24)*

23. Indique en qué sectores de actividad ha trabajado *(Puede responder más de una)*

- [ ] Agricultura
- [ ] Confección, textil, calzado, marroquinería
- [ ] Comercio
- [ ] Empleada de Hogar
- [ ] Limpieza de oficinas
- [ ] Hostelería
- [ ] Otro (especificar)

24. ¿En qué actividad trabaja actualmente? *(Puede responder más de una)*

- [ ] Agricultura
- [ ] Confección, textil, calzado, marroquinería
- [ ] Comercio
- [ ] Limpieza de oficinas
- [ ] Hostelería
- [ ] Otro (especificar)

- [ ] Empleada de Hogar

25. ¿Cuánto tiempo lleva en su empleo actual?

- [ ] Menos de 1 año
- [ ] De 1 a 3 años
- [ ] De 3 a 5 años
- [ ] Más de 5 años

26. ¿Cómo ha encontrado este trabajo?

- [ ] A través de familiares o amigos
- [ ] Por iniciativa propia
- [ ] A través de la prensa
- [ ] A través de una ONG o una parroquia
- [ ] Otros (especificar)

27. ¿Cuántos días a la semana trabaja?

28. ¿Cuántas horas semanales trabaja?

29. ¿Cuántas horas semanales le gustaría trabajar?

30. ¿Suele hacer más horas de las fijadas en su horario habitual? ¿Se las pagan aparte?

- [ ] No suelo hacerlas
- [ ] Las hago y me las pagan como corresponde
- [ ] Las hago pero no me pagan nada extra
- [ ] Las hago y me pagan según la ocasión
- [ ] Las hago y me compensan con un descanso

31. Indique el grado de satisfacción de los siguientes aspectos de su trabajo:
(Valorar de 0 a 5 cada aspecto)

- [ ] Salario
- [ ] Horario de trabajo
- [ ] Relación con su empleador/a
- [ ] Vacaciones/Permisos
- [ ] Pagas extraordinarias

32. ¿Podría decirme qué significa para usted tener un trabajo?
(Valorar de 0 a 5 el grado de importancia de cada uno de los siguientes aspectos

- [ ] Es un complemento para aumentar la renta familiar
- [ ] Reconocimiento Social
- [ ] Acceso a prestaciones sociales (jubilación, subsidio de desempleo…)
- [ ] Es un medio para desarrollarse como persona
- [ ] Independencia económica
- [ ] Otro (especificar)

33. ¿Si pudiera?

- [ ] Trabajaría por su cuenta
- [ ] Disminuiría el número de horas de dedicación al trabajo
- [ ] Aumentaría el número de horas de dedicación al trabajo
- [ ] Dejaría de trabajar para dedicarme a mi familia

34. ¿Dónde está su trabajo?

	En la localidad en la que vive
	A menos de 10 Km de la localidad donde vive
	Entre 10 y 20 Km de la localidad donde vive
	A más de 20 km de la localidad donde vive

35. ¿Cuánto tiempo lleva dedicándose al servicio doméstico?

	Menos de 3 meses
	De 3 a 6 meses
	De 6 meses a 1 año
	De 1 a 3 años
	Más de 3 años

36. ¿En que modalidad de empleo doméstico trabaja actualmente?

	Interna
	Externa Fija
	Externa por horas

37. ¿En cuántas casas trabaja actualmente?_____

38. Díganos con qué frecuencia realiza las siguientes tareas

Habitualmente,	*A veces*	*Nunca*	
			Cocinar
			Hacer la compra
			Limpiar la casa
			Planchar
			Cuidar los niños
			Cuidar enfermos, discapacitados o personas mayores

39. ¿Ha estado dada de alta en el antiguo Régimen Especial de Empleados de Hogar de la Seguridad Social?

	Si
	No *(Ir a la pregunta 42)*

40. ¿Quién cotizaba a la Seguridad Social?

	El/la empleador/a
	El/la empleador/a y usted (parte proporcional)
	Usted

41. ¿Porqué no cotiza a la Seguridad Social?

	No tiene obligación
	No le interesa o no le compensa
	No sabe si tiene que cotizar

42. ¿Conoce las nuevas prestaciones sociales a las que tiene/tendría derecho como cotizante a la Seguridad Social tras la inserción del Régimen Especial de Empleados de Hogar de la Seguridad Social?

☐	Si
☐	No

43. ¿Considera que con la nueva Ley de Empleadas de Hogar 2012 es totalmente equiparable su trabajo al resto de los trabajadores en otros sectores?

☐	Si
☐	No
☐	NS/NC

44. La cotización como empleada de hogar no da derecho al desempleo ¿considera que las Administraciones Públicas deberían considerarlo?

☐	Si
☐	No

¿por qué?

Las preguntas de la 45 a la 50, deben ser contestadas solamente por mujeres que **TRABAJEN Y NO COTICEN** (O COTICEN POR MENOS HORAS DE LAS TRABAJADAS).

45. Indique cuáles de las siguientes circunstancias han influido en sus condiciones laborales actuales (NO contrato y/o NO cotización a la Seguridad Social)

☐	El salario es tan bajo que no compensa cotizar a la Seguridad Social
☐	No tiene papeles y no tiene alternativa
☐	Dejó de trabajar hace algunos años y cuando ha querido volver a trabajar no ha encontrado otra alternativa
☐	Es ama de casa, sin experiencia laboral, y no ha encontrado otro tipo de trabajo
☐	Prefiere cobrar todo el salario y ha llegado a un acuerdo con el empleador/a
☐	Es un tipo de trabajo que le permite atender a la familia y obtener ingresos extra
☐	Su bajo nivel de formación le ha obstaculizado el acceso a otro tipo de empleo
☐	Su empleador/a le ha impuesto estas condiciones
☐	Otro *(especificar)*

CONSECUENCIAS SOBRE LA VIDA LABORAL Y PERSONAL

Díganos su grado de acuerdo, o desacuerdo, con las siguientes afirmaciones.

46. El desempeño de un trabajo en condiciones precarias:
(Valorar de 0 a 5 cada circunstancia)

	Impide beneficiarse de prestaciones sociales (subsidio de desempleo, jubilación...)
	Relega al trabajador a tareas tediosas, pesadas, etc
	Dificulta encontrar un empleo regularizado (con contrato, cotizando a la Seguridad Social, etc)
	Impide la promoción profesional
	Recorta la posibilidad de pedir mejoras en las condiciones laborales (salario, horario, vacaciones, etc)
	Quita aliciente a buscar un empleo que de acceso a prestaciones sociales
	Perjudica la valoración profesional que los demás tienen respecto de una misma

47. Y con respecto a la vida personal , el desempeño de un trabajo en condiciones precarias: *(Valorar de 0 a 5 cada circunstancia)*

	Aísla al trabajador socialmente
	Reduce la autoestima del trabajador
	Limita la independencia económica
	Reduce la valoración personal que la familia y los amigos tienen de una misma

DEMANDAS Y MEDIDAS

48. Si pudiera elegir, ¿le gustaría tener un empleo que le garantizara el acceso a prestaciones sociales? (con contrato y/o cotizando a la Seguridad Social)

	Si
	No *(fin de la encuesta)*

49. A continuación le voy a indicar una lista de posibles demandas que a las mujeres en condiciones laborales semejantes a la suya les gustaría hacer a la Administración, las empresas y/o a la sociedad. Indique el grado de importancia que tienen para usted cada una de ellas *(Valorar de 0 a 5 cada circunstancia)*

	Flexibilizar el horario laboral
	Obtener apoyo, asesoramiento y orientación en la búsqueda de empleo
	Que la Administración vaya asumiendo el cuidado de las personas dependientes (guarderías, servicios de ayuda a domicilio, residencias...)
	Realizar campañas de sensibilización social para cambiar los factores culturales que llevan a considerar el trabajo de la mujer como un complemento y no como un derecho
	Tener acceso a prestaciones sociales (jubilación, subsidio de desempleo etc)
	Otra (especificar)
	Exigir al empleador/a un contrato por escrito *Sólo para empleadas de hogar*
	Obtener el permiso de residencia *Sólo para inmigrantes sin papeles*

50. Por último, le voy a indicar una lista de posibles medidas para ayudar a las mujeres a mejorar sus posibilidades de acceso a empleos, que les aseguren prestaciones sociales y condiciones laborales dignas. Indique el grado de importancia que tienen para usted. *(Valorar de 0 a 5 cada circunstancia)*

	Programar cursos prácticos que verdaderamente sean v útiles para la inserción laboral
	Facilitar el acceso a cursos para mejorar la formación y la cualificación de las mujeres
	Ofrecer servicios de apoyo y orientación a las mujeres en la búsqueda de un trabajo
	Disminuir las cotizaciones sociales
	Reducir los impuestos exigidos para establecerse por cuenta propia
	Aumentar la intervención de la inspección laboral
	Aumentar las responsabilidades penales a las empresas que incurran en irregularidades
	Otra (especificar)

Muchísimas gracias por su colaboración