“Viejos y nuevos dilemas en la exploración de datos”

Desarrollo de metodología y producción /análisis de datos

Grupo de Trabajo Nº16. Metodología y epistemología de las ciencias sociales

Pablo Hein
Departamento de Sociología
Facultad de Ciencias Sociales
Universidad de la República

En la última década, tanto los avances tecnológicos, como informáticos, han llevado a que la metodología cuantitativa y específicamente al análisis de datos se haya visto colonizado por una suerte de técnicas multivariables. Algunas con mayor fuerza, otras con mayor seducción, se introducen a-críticamente en los informes de investigación y dan por sentado un mundo solidificado y plausible de ser comprendido en un universo de variables métricas.
La presente ponencia expone, tanto las ventajas como desventajas, de algunas técnicas de exploración-segmentación, así como sus diferentes alcances y niveles. Este conjunto de técnicas, resumidas en la “de segmentación”, es utilizada por la investigación aplicada y reintroducidas recientemente por la investigación científica. Mediante ejemplos concretos, se sintetizan las principales ventajas y desventajas en su uso en las cc.ss.

Palabras claves: metodología, análisis multivariado, segmentación

Introducción

El presente artículo describe a partir de ejemplo concreto las potencialidades y limitaciones de Análisis de Segmentación, conocido como el procedimiento Chaid.
Este procedimiento habitualmente se utiliza en los estudios de mercado, para segmentar clientes y/o productos.
En los estudios donde se utilizan cuestionarios estructurados, casi siempre se parte de un conjunto de hipótesis, teorías o un conjunto de conocimiento científico comprobado previamente. En algunos casos, como en los censos, contrariamente al devenir científico, este conjunto de hipótesis no se establecen claramente, ya que algunos estadísticos sostienen que los censos tienen la única finalidad de “construir marcos muestrales potentes”.
Cuando los metodólogos o analistas sociales se enfrentan a estas cuestiones, y con la necesidad de explicar y/o explorar un problema, se comienza a cruzar variables existentes (independientes), con la finalidad de “encontrar” asociaciones o correlaciones que luego “alimenten” los más diversos análisis multivariados.
Esta técnica, tiene la potencialidad de seleccionar variables, relevantes a priori teóricamente y/o empíricamente, para iluminar y describir aquellas variables que se asocian con el fenómeno y por otro lado, desarrolla las diferencias o semejanzas al interior de la matriz de datos, de los grupos o subgrupos, que presentan diferencias con el fenómeno.
En concreto se puede resumir esta técnica como un análisis exploratorio de datos, que a diferencia con la minería de datos, explorar una base a la vez.
Como sostiene Escobar *“Su potencia, al mismo tiempo que su peligro, reside en la selección automática de aquellas categorías que pronostican mejor los valores de la variable considerada*

objetivo. Además segmentar significa dividir y en consecuencia, permite que se hallen grupos…" (*M. Escobar 1998*)

En el presente texto, mediante dos ejemplos, se intenta despejar/segmentar diferentes grupos, que presenten diferencias entre ellos (características distintas en la variable dependiente) y determinados rasgos en comunes a su interior (similitudes en las independientes), para dos situaciones concretas.

En un primer caso, la producción científica de los docentes de la Universidad de la República, medida mediante el número de publicaciones. En el otro ejemplo, la reincidencia carcelaria, que se puede leer como reincidencia en el delito, de las personas privadas de libertad (reclusos).

Las dos cuestiones comunes es que, en primer lugar, en ambos casos, los datos son emanados de Censos, en el primer caso del II Censo de Funcionarios Docentes. Este Censo se llevó de forma conjunta a funcionarios, tanto técnicos, administrativos, servicios, así como a los funcionarios docentes, durante el mes de noviembre de 2009. En el caso de las personas privadas de libertad (reclusos) los datos provienen del I Censo de Población Privada de Libertad a nivel nacional, realizado en el mes de Setiembre del 2010.

En segundo lugar, y sucede comúnmente con los datos emanados de todo relevamiento censal, el difícil e intrincado camino de procesamiento e interpretación de los datos, dado que no necesariamente, los indicadores (preguntas) se construyen a partir de teorías estructuradas. En ambos casos se pretende un acercamiento exploratorio acerca de cuales son las factores/variables preponderantes en la explicación/comprensión de un fenómeno determinado.

La técnica

La técnica de árbol de segmentación, bien puede utilizarse para el estudio de variables nominales, ordinales hasta de razón.

El procedimiento en su fase final, crea un modelo de clasificación basado en "ramas y hojas" en la lógica de árboles, y las hojas o nodos terminales nos ejemplifican "casos en determinados grupos", con determinadas características semejantes que difieren con el resto de los grupos.

A su vez se puede analizar el pronostico de la variable dependiente (problema) basado en valores algorítmicos (entre otros) a partir de un conjunto de variables independientes (pronosticadoras).

Este análisis si bien puede ser confirmatorio, es recomendable que se utilice y evalúe como análisis predictivo.

En sus orígenes, esta técnica se basó en variables cualitativas, pero a comienzos de la década de los ´60, los estudios y modelizaciones se centraron en la profundización de los estudios que pudieran incluir variables cuantitativas, forzando de esta manera el desarrollo cuasi-matemático de los algoritmos y/o asociaciones. Los desarrollo recientes y los niveles de medición de las variables utilizadas la mayoría de la veces por los cientistas sociales, destacan los análisis basados en el estadístico Chi cuadrado, permitiendo de esta manera la incorporación de variables nominales u ordinales, en conjunto con variables más potentes. Esto implica, hasta la fecha la reducción de las variables métricas, a ordinales para su correcta inclusión en los modelos.

El algoritmo denominado CHAID, como sostiene Escobar "…. Desarrollada por Cellard (1967), Bourouche y Tennenhaus (1972), Kass (1980) y Magidson (1989,1993 y 1993) … tiene como principal característica distintiva de otros algoritmos de segmentación el que la muestra no se segmente de modo binario …. O que se puede formar segmentos con dos categorías al unísono. Al igual que otras prácticas de segmentación, las operaciones elementales que realiza son: a) la agrupación de las categorías de variables pronosticadoras; b) la comparación de efectos entre distintas variables, y c) la finalización del proceso de segmentación. (Escobar, M. 1998).

Este algoritmo o procedimiento, como ya se adelantó, tiene como limitantes el uso de las variables de razón deberán recategorizarse en variables de menor nivel de medición, léase ordinales. A su vez los

grados de libertad de las tablas, y por ende del estadístico de asociación, es similar a todas las tablas desarrolladas, ya que la variable dependiente (problema) es la misma y las variables independientes "son pares de categorías".[1]
A nivel general, las ventajas de la técnica, entre otras, es la rápida segmentación de la matriz de origen, y la inclusión y exclusión de variables predictoras teóricamente seleccionadas como independientes.
A su vez permite segmentar en grupos, estratificar de los mismos en relación al problema, la predicción de eventos futuros, reducción de datos y clasificación de variables y la fusión de variables continuas y categorías de respuestas (Manual de SPSS).
Es importante señalar que las variables independientes o denominadas pronosticadoras, se puede combinar sus categorías de respuestas mediante dos procedimientos básicos, el primero es el libre, que puede resultar tanto pares de combinación (en su recodificación) como combinaciones de categorías. Este procedimiento puede fusionar categorías de respuestas extremadamente opuestas, por ende se corre ciertos riesgos metodológicos para su análisis posterior y la opción monótona que intenta limitar la obtención de fusiones lógicas. En este caso se forman los grupos deseados, es decir la recodificación de una variable ordinal de cuatro respuestas, en tres variables nominales y cada una con dos categorías de respuestas, fácilmente interpretables.

Las poblaciones de estudio

En el caso de los funcionarios docentes de la UdelaR, se aplicó un formulario auto administrado, vía página web. La información de base se alimentaba de los registros contables de la Universidad. Se censaron 8628 docentes en la UdelaR. Existen 195 casos que comparten la condición de ser funcionarios docente y no docente. Luego de contar con la matriz de datos, se planteo la necesidad de realizar diversos informes entre otros, los solicitados por rectorado que solicitaba un informe detallado sobre la producción de los docentes (léase publicaciones) y la formación académica
En el caso de las personas privadas de libertad, el objetivo general fue relevar y explorar los rasgos básicos de la totalidad de la población carcelaria, conocer las condiciones básicas de privación de libertad y determinar las condiciones sociales, culturales y económicas de dicha población. Se aplicó un formulario administrado, diseñado específicamente para estas instancias. Se censaron 5821 personas privadas de libertad, 2661 rechazaron brindar información, constituyendo 8492 reclusos. Para el análisis se tomaron 3910 casos, que constituyen una población con respuestas válidas para las variables incluidas en el modelo
En este caso se intenta aportar como se despejaron las variables que tienen una mayor incidencia en el tema de la reincidencia carcelaria,

Las variables dependientes (pronosticada)

Para el caso de los docentes la variable dependiente, se construyó a partir de un conjunto de preguntas en las cuales se les interrogaba a los docentes si habían realizado algún tipo de publicaciones en los últimos cinco años. Las publicaciones relevadas fueron cuatro; Documentos de trabajo e informes de investigación, Coautoría de artículos en revistas, Artículos en revistas científicas y Autoría/Coautoría de libros (incluye artículos en libros).
Luego se construyó un índice sumatorio simple y posteriormente se trabajó con un único indicador simple, que distingue si los docentes han realizado o no algún tipo de publicación relevados, con

[1] Por más que la variable ordinal sea de tres o más categorías el procedimiento podrá reagrupar las categorías en su dos tramos siempre y cuando las categorías sean acumulables, como por ejemplo tramos de horas dedicadas a la investigación, de 0 a 10, de11 a 30, 31 a 40 y 41 y más.

independencia de cantidad. En este sentido, un 61% de los docentes manifestó que había realizado algún tipo de publicación en los últimos cinco años.
Para el caso de los reclusos la variable dependiente, se construyó a partir del indicador simple que fue la pregunta "¿UD es primario?" Y con su contraste "¿Con esta cuantas veces estuvo recluido?". A nivel general, un 48% de los censados manifestó ser primario, por lo que nuestra variable dependiente asume un valor positivo en el 52% de los casos. Por otro lado se realizo un estudio detallado de la variable edad y se selecciono para esta técnica/modelo a aquellos individuos que cuenten como mínimo 25 años de edad.

Las variables independientes (pronosticadoras)

Para el caso de los docentes, luego de realizado el censo y con los primeros resultados, las variables que se listaban en las diferentes interpretaciones eran entre otras, la carga horaria del docente, la incidencia o el peso de la horas dedicadas a la investigación, el ser full-time, la participación en proyecto internacionales, el grado docente[2], la propia área de conocimiento, el sexo, la edad, estudios de posgrado y hasta el tipo de hogar. En relación a algunos indicadores o frecuencias simples de la variables independientes un 33% de los docentes pertenece al área de Ciencias de la Salud y en porcentajes cercanos al 30% se encuentran los docentes del área de Ciencias Sociales y Humanas y Ciencias y Tecnológicas. Por último los docentes del área Agrarias concentran casi el 9%. Un 60% aproximadamente, cuenta con una carga horaria menor a 29 horas semanales, esto se da tanto en hombres como en mujeres. Aquellos docentes que tienen una alta dedicación (mayor a 40 horas) con el 25% y 23% de los hombres y mujeres respectivamente.
La participación por grado docente, muestra claramente una concentración en los grado intermedios (2 y 3 respectivamente) Los docentes de grados superiores (4 y 5) son apenas el 7% de la población. Esta distribución es bastante similar en todas las áreas de conocimiento. Casi todas las áreas coinciden entorno al 35% de docentes que cursaron posgrado, salvo el área artística que tan solo el 11% manifestó positivamente dicha pregunta. En casi todas las áreas, salvo Ciencias de la Salud, los estudios de Maestría y Doctorado adquieren porcentajes significativos. Por último un 59% manifestó haber realizado algún tipo de actividad vinculada a la investigación en los últimos tres años.
En el caso de la población reclusa, las variables que se listaban en las diferentes interpretaciones eran de las más diversas "procedencias", sean estas estructuradoras de la conducta, de carácter psicosocial o psíquico. Siguiendo esta línea de trabajo, se comenzaron a delimitar diferentes conjuntos de factores sociales, personales y por último redes familiares y de amigos, que se pueden pensar como influyentes, con diferente "peso", de la reinserción carcelaria. Para esto se pensó en un conjunto de indicadores dentro de las dimensiones mencionadas. Para la dimensión individual se incluyó el sexo, delitos cometidos bajo efectos de sustancias tóxicas (drogas y/o alcohol), estado civil, nivel educativo y si contaba con trabajo al momento de la detención.
La dimensión psicosocial incluye las variables; violencia familiar cuando el recluso era un niño/a y si recibió castigo físico en la infancia. Por último la dimensión familiar y de amigos cercanos constó de dos planos y se definió de la siguiente manera; el primero de estos planos incluía la constatación de círculos sociales con antecedentes carcelarios y/o delictivos, internación en hogares de INAU o similar en su infancia, y el segundo con la composición familiar durante su infancia, el tipo de vivienda donde residía en el momento de ser privado de libertad y el barrio donde se ubicaba la misma (asentamiento – no asentamiento).

[2] En la UdelaR existe una división docente (carrera docente) en cinco tramos desde los docentes grado 1, inician su carrera, ayudantes pasando por los grados 2 y 3, en los cuales se incorporan mayores responsabilidades hasta llegar a los grados 5, docente denominados profesores titulares. Para nuestro objeto es una variable continua.

El modelo

Como ya se sostuvo, para analizar tanto la producción de los docentes, como los factores que inciden en la reincidencia carcelaria, se presenta la siguiente tabla en la cual se especifican las variables incluidas, el método final de crecimiento y los casos mínimos finales en las hojas (nodos terminales).
El método Chaid exhaustivo fuerza el modelo en las variables ordinales, hasta la fusión continua de pares de valores (categorías) hasta lograr una única dicotomía de valores. En concreto se limita a la obtención de segmentaciones binarias. (Biggs et. al 1991).
Por otro lado se especifico el corte del modelo, en un nivel máximo de tres (profundidad), a los efectos de lograr la mejor combinación del conjunto de variables independientes propuestas y se intentó terminar la segmentación con un número no menor de 400 casos, para que la conclusiones o posibles vías de explicación conjugara un número suficiente de casos.

Tablas resumenes de los modelos

Resumen del modelo		
Especificaciones	Método de crecimiento	CHAID exhaustivo
	Variable dependiente	Prod_
	Variables independientes	Posgrado, Otra ocupación, Grado, Edad en tramos, Sexo, Area de conocimiento, Gestión, Extensión, Investigación, docencia, Horas en tramos, Tipo de Hogar, Numero de hijos
	Máxima profundidad de árbol	3
	Mínimo de casos en un nodo filial	400
	Mínimo de casos en un nodo parental	100
Resultados	Variables independientes incluidas	Investigación, Edad en tramos, Posgrado, Grado
	Número de nodos	23
	Número de nodos terminales	14
	Profundidad	3

(Fuente: Elaboración propia. 2012)

Resumen del modelo		
Especificaciones	Método de crecimiento	CHAID exhaustivo
	Variable dependiente	Primario No Primario

	Variables independientes	Sexo, Cuando cometió este delito lo hizo bajo el defecto del alcohol o de laguna droga, Familiares o amigos con antecedentes, En su infancia o adolescencia estuvo internado/a en un hogar del INAU o similar, Personas con las que vivía entre los 8 y los 10 años, En esos años vio o escucho a sus padres, mayores maltratarse físicamente, Cuando era niño, en su casa le pegaban o castigaban habitualmente , Estado Civil, Donde vivía antes de ingresar a este establecimiento , Su vivienda se ubicaba en un asentamiento, Nivel educativo más alto que ha alcanzado, Trabajaba antes de ingresar a este establecimiento.
	Máxima profundidad de árbol	3
	Mínimo de casos en un nodo filial	1000
	Mínimo de casos en un nodo parental	100
Resultados	Variables independientes incluidas	En su infancia o adolescencia estuvo internado/a en un hogar del INAU o similar, Sexo, Trabajo anterior, Círculos con antecedentes
	Número de nodos	13
	Número de nodos terminales	8
	Profundidad	3

(Fuente: I Censo de Población Privada de Libertad M.I. – FCS Setiembre 2010.Elaboración propia.)

Tabla resumen de resultados I

	Pronosticado		
Observado	No	Si	Porcentaje correcto
No	2177	1202	64,4
Si	873	4376	83,4
Porcentaje global	35,4	64,6	**76,0**

(Fuente: I Censo de Población Privada de Libertad M.I. – FCS Setiembre 2010.Elaboración propia.)

Como se explicita en la cuadro el modelo resultante, pronostica correctamente, un 76% de los casos y los nodos terminales (grupo resultantes) son 14.

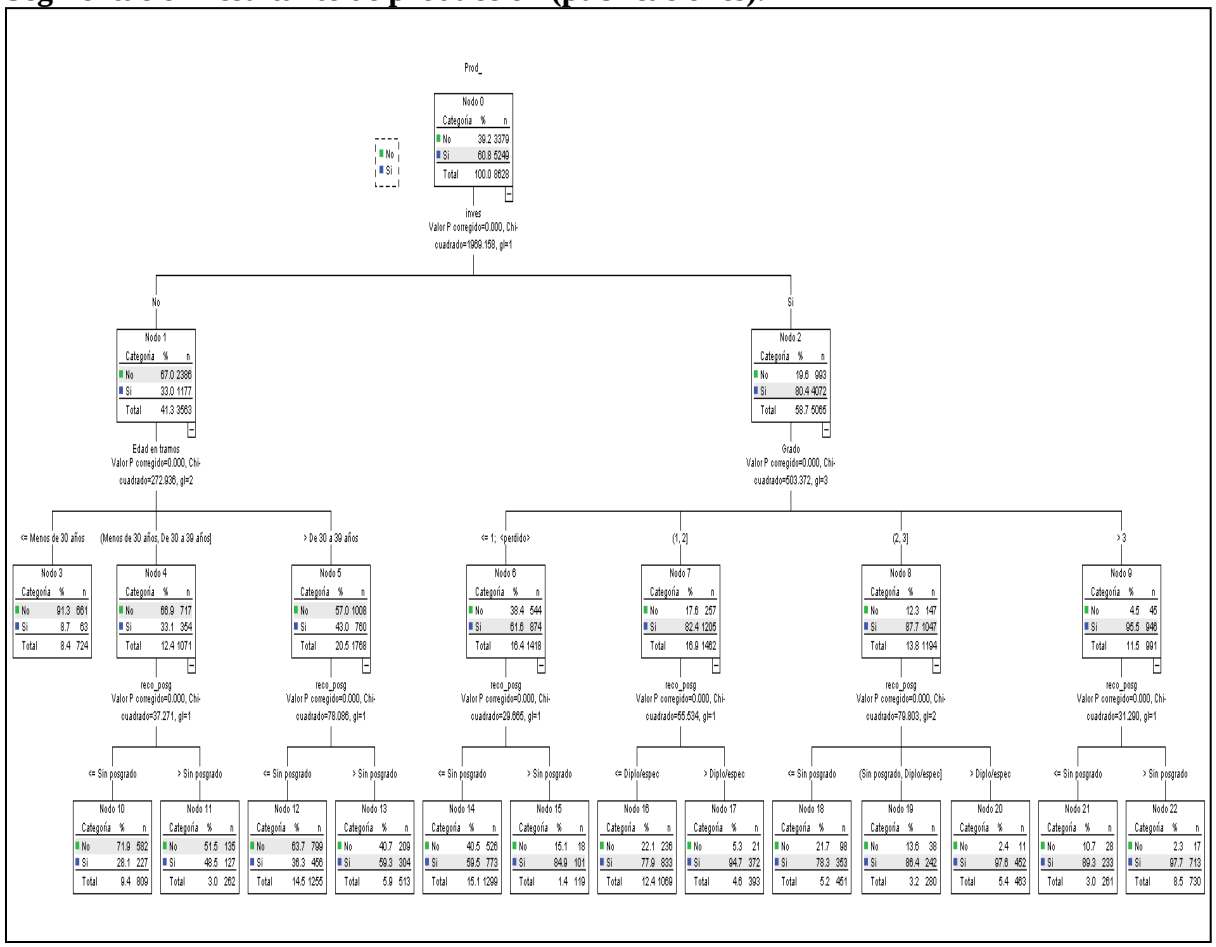
Tabla resumen de resultados II

	Pronosticado		
Observado	Primario	No primario	Porcentaje correcto
Primario	1426	434	76,7
No primario	857	1193	58,2
Porcentaje global	58,4%	42,6%	**67,0**

(Fuente: Elaboración propia. 2012)

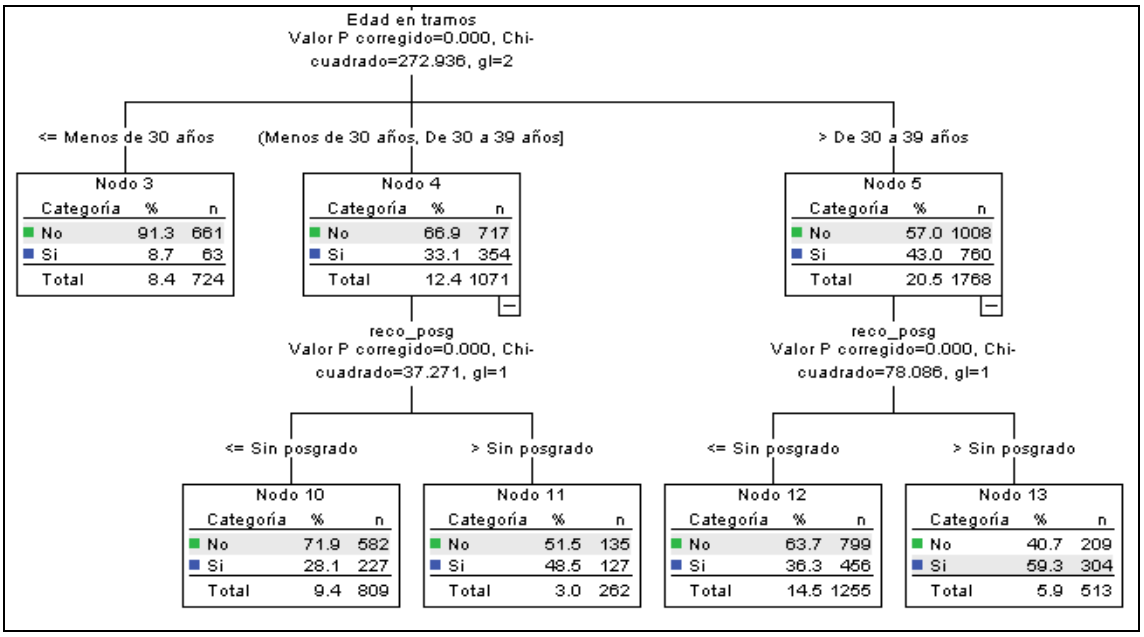
Como se explicita en la cuadro el modelo resultante, pronostica correctamente, un 67% de los casos y los nodos terminales (grupo resultantes) son 8.

Segmentación resultante de producción (publicaciones).



(Fuente: II Censo de Funcionarios Docentes UdelaR Noviembre 2010 Elaboración propia. 2012)

Segmentación Parcial de los docentes que no realizan investigación



(Fuente: II Censo de Funcionarios Docentes UdelaR Noviembre 2010 Elaboración propia. 2012)

En una primera lectura, que se puede realizar siguiendo la segmentación, solo con fines descriptivos, se puede afirmar que, la primera segmentación para los que no publican y no realizan actividades de investigan acumulan 3563 docentes, de los cuales un 67% no publican. Estos a su vez se dividen en tres nodos según la edad mostrando un perfil marcadamente "joven" como era de esperar y prevaleciendo mayorías de docentes que no publican. Luego se produce la última segmentación interviniendo la variable posgrado, generado cuatro nodos terminales (hojas) predominando los docentes que no publican, salvo en el nodo terminal 5, con un 59% de docentes que publican, que son; con algún posgrado, mayores de 39 años y no realizan investigación. Es importante señalar que se genero un nodo terminal en el caso de los menores de 30 años y que no investigan. En este el 91% no publica.

Otra de las miradas sobre la capacidad que tiene tanto la segmentación, como de las variables en su conjunto, se puede dar, con las tablas construidas a continuación, la variable dependiente con cada grupo de los nodos terminales. A los efectos de presentación se subdividió la tabla en tres partes. Por un lado están los nodos terminales, con sus grupos característicos, que no realizan actividades de investigación, que son cinco grupos bien definidos. Por otro los nodos terminales de los que si realizan investigación, que son nueve grupos.

Esta mirada es sumamente útil, para efectuar una descripción y /o simplificación del análisis, es decir interpretar los grupos terminales.

Más allá de este hecho es bueno señalar en esta instancia que en el modelo se han incorporado la gran mayoría de variables independientes, que estaban en cierta discordia en el grupo de trabajo. Este hecho puede ser también objeto de interpretación ya que el análisis en cuestión se encarga de filtrar las más relevantes (esto se detallara en la conclusiones).

En la primera tabla, se aprecia claramente que los tres primeros grupos, predominan los docentes que no publican, salvo en aquellos mayores a 39 años de edad y que cuentan con algún tipo de posgrado terminado. Más allá de esta afirmación es interesante señalar que el porcentaje de quienes publican en este grupo, no supera, en porcentajes a ninguno de los nueve grupos terminales de docentes que si realizan actividades de investigación.

Producción (publicaciones) para los cinco grupos terminales (nodos) de la segmentación, no investigan

				No realizan actividades de investigan				
Publicaciones		Total	Subtotal no investiga	> 30	Hasta 39 Sin posgrado	Hasta 39 Algún posgrado	Mayor 39 Sin posgrado	Mayor 39 Algún posgrado
	No	39%	67%	91.3%	71.9%	51.5%	83.7%	40.7%
	Si	51%	33%	8.7%	28.5%	48.5%	36.3%	59.3%
Total		**(8628)**	**(3583)**	**(724)**	**(809)**	**(262)**	**(1255)**	**(513)**

(Fuente: II Censo de Funcionarios Docentes UdelaR Noviembre 2010 Elaboración propia. 2012)

En la segunda tabla se aprecian claramente nueve grupos terminales, pero un hecho importante es que la segunda variable de segmentación es producida, no por las edades, sino por los grados docentes. Esta segunda segmentación nos esta indicando, que para los docentes que investigan y se esta pensando en las publicaciones, la mirada o las hipótesis tendrían que "correr" por el lado de los grados docentes.
Por otro lado, la tercera segmentación arrojo cuatros grandes grupos, que son los becarios (docentes sin grado), los grados uno y dos, los grados tres y los grados superiores a tres, es decir cuatro y cinco.
En cualquiera de dichos grupos terminales, el porcentaje de docentes que publican sobrepasa el 60%, es decir 6 de 10 publicaron algún tipo de comunicación en los últimos tres años.
Por último se observa que los docentes, independientemente de grado aquellos que cuentan con un título de posgrado presentan porcentajes superiores en su grupo terminal.

Producción (publicaciones) para los nueve grupos terminales (nodos) de la segmentación, si investigan (parte I)

				Realiza actividades de investigación						
Publicaciones		Total	Sub total Investiga	Becarios Sin posgrado	Becarios Cursando posgrado	Grado 1 y 2 Sin o cursando diplomas	Grado 1 y 2 Con diploma o superior	Grado 2 o 3 Sin posgrado	Grado 2 o 3 cursando posgrado	Grado 2 o 3 Con posgrdo
	No	39%	19.6%	40.5%	15.1%	22.1%	5.3%	21.7%	13.6%	2.4%
	Si	51%	80.4%	59.5%	94.9%	77.9%	94.7%	78.3%	86.4%	97.5%

Total		**(8628)**	**(5065)**	**(1299)**	**(119)**	**(1069)**	**(393)**	**(451)**	**(280)**	**(463)**

(Fuente: II Censo de Funcionarios Docentes UdelaR Noviembre 2010 Elaboración propia. 2012)

Producción (publicaciones) para los nueve grupos terminales (nodos) de la segmentación, si investigan (parte II)

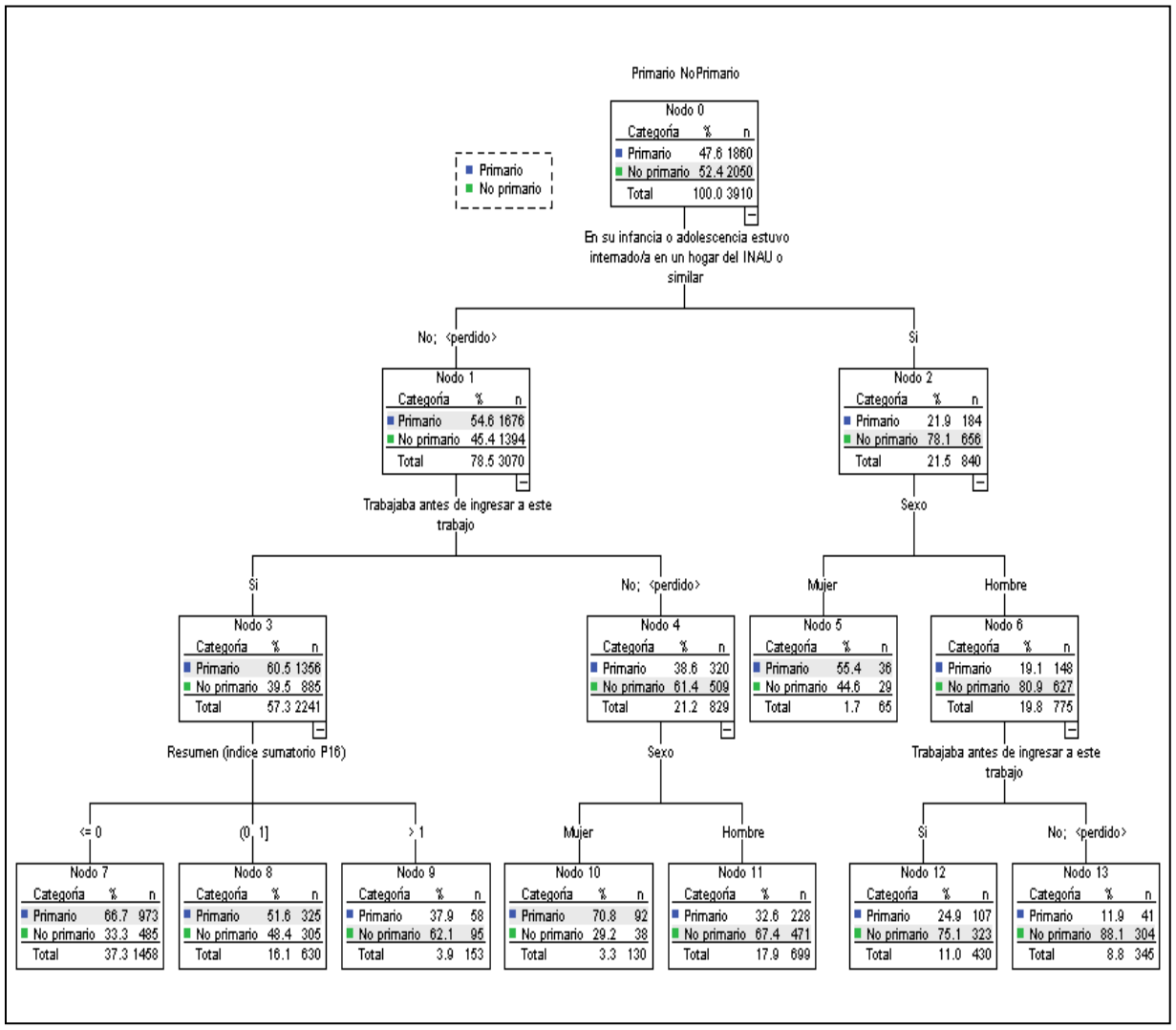
				Realiza actividades de investigación	
Publicaciones		Total	Sub total Investiga	Superior a grado 3 Sin posgrado	Superior a grado 3 Con posgrado
	No	39%	19.6%	10.7%	2.3%
	Si	51%	80.4%	89.3%	97.7%
Total		**(8628)**	**(5065)**	**(261)**	**(730)**

(Fuente: II Censo de Funcionarios Docentes UdelaR Noviembre 2010 Elaboración propia. 2012)

Este análisis de segmentación, nos permite realizar descripciones de los grupos resultantes, con características distintas entre ellos. Por su lógica de segmentación encontró grupos muy distintos entre ellos, como ser para los que no realizan actividades de investigación, la segmentación sucesiva se realizó por edades, mientras que en los docentes que si realizan actividades de investigación la segmentaciones produce por los grados docentes, una variable que marca de alguna manera el perfil de los docentes que publican.

Otro hecho importante en esta práctica es la nítida separación de los grupos resultantes y la marcada porcentualización del grupo que publica.

Segmentación resultante de reincidencia.



(Fuente: I Censo de Población Privada de Libertad M.I. – FCS Setiembre 2010.Elaboración propia.)

En primer lugar, es oportuno señalar que el modelo pronostica un 67% de los casos.
En segundo lugar, y observando el gráfico anterior, las variables que mejor predicen o reagrupan casos y por ende con un mayor "impacto" sobre la dependiente son: "Internaciones en un hogar de INAU o similar durante su niñez y/o adolescencia", "Trabajaba al momento de la detención" y "Sexo". Cabe aclarar que dichas variables poseen diferentes magnitudes y su "comportamiento" para la segmentación es diferencial. En concreto, por un lado existe una variable que da cuenta para el total de la población que es las "Internaciones en hogares del INAU o similar durante la infancia y/o adolescencia". Todo parece indicar que esta variable juega un papel fundamental en la determinación de los nodos y por ende en la segmentación de la población (casos).
Es así que se constituyen dos nodos centrales, de los cuales derivan dos variables, (en términos del árbol, se "abre" dos ramas) indicando que la variable "Trabajaba al momento del ingreso" y "Sexo", impacta diferente para sobre la primer "rama" que son los "Antecedentes en INAU o similar". A los que no contaban con internaciones se le asocia la variable "Trabajaba al momento de la detención" y para aquellos que si tuvieron experiencias de internación en hogares de INAU se les asocio la variable "Sexo".

Para aquellos que, no tenían antecedentes de INAU pero si trabajan al momento de la detención, la variable resumen de los "Familiares o amigos cercanos con antecedentes penales" fue la que más se asocio a diferencia de los que no tenían trabajo se les asocio el "Sexo".
Si se analiza la información bajo la óptica de los nodos terminales, en el caso de los primarios (no reincidentes) el nodo 10 compuesto por mujeres sin Internación INAU y que no trabajaban al momento de su detención y el nodo 7 integrado por personas que no estuvieron internados INAU que si trabajaban al momento de su detención y que no tenían círculos sociales con antecedentes delictivos, son los nodos que mas explican los primarios, es decir sin reincidencia.
Para los reincidentes (no primarios), el nodo 13 constituido por hombres previamente internados en el INAU o similar y sin trabajo al momento de la detención, el nodo 12 el mismo perfil que el anterior pero con trabajo, el nodo 11 hombres que no estaban internados INAU y no trabajaban y el nodo 9, personas sin antecedentes de internación en el INAU que si trabajaban al momento de su internación y con mas de dos personas en sus círculos sociales con antecedentes son los que mas "*aportan al modelo"*, para predecir a los No primarios es decir, los reincidentes .

Ganancias para los nodos Primarios

	Nodo		Ganancia			
Nodo	N	Porcentaje	N	Porcentaje	Respuesta	Índice
10	**130**	**3,3%**	**92**	**4,9%**	**70,8%**	**148,8%**
7	**1458**	**37,3%**	**973**	**52,3%**	**66,7%**	**140,3%**
5	**65**	**1,7%**	**36**	**1,9%**	**55,4%**	**116,4%**
8	**630**	**16,1%**	**325**	**17,5%**	**51,6%**	**108,4%**
9	153	3,9%	58	3,1%	37,9%	79,7%
11	699	17,9%	228	12,3%	32,6%	68,6%
12	430	11,0%	107	5,8%	24,9%	52,3%
13	345	8,8%	41	2,2%	11,9%	25,0%

(Fuente: I Censo de Población Privada de Libertad M.I. – FCS Setiembre 2010.Elaboración propia.)

Método de crecimiento: CHAID Variable dependiente: Primario – No Primario

Ganancias para los nodos NO Primarios (reincidentes)

	Nodo		Ganancia			
Nodo	N	Porcentaje	N	Porcentaje	Respuesta	Índice
13	**345**	**8,8%**	**304**	**14,8%**	**88,1%**	**168,1%**
12	**430**	**11,0%**	**323**	**15,8%**	**75,1%**	**143,3%**
11	**699**	**17,9%**	**471**	**23,0%**	**67,4%**	**128,5%**
9	**153**	**3,9%**	**95**	**4,6%**	**62,1%**	**118,4%**
8	630	16,1%	305	14,9%	48,4%	92,3%
5	65	1,7%	29	1,4%	44,6%	85,1%
7	1458	37,3%	485	23,7%	33,3%	63,4%
10	130	3,3%	38	1,9%	29,2%	55,8%

(Fuente: I Censo de Población Privada de Libertad M.I. – FCS Setiembre 2010.Elaboración propia.)

Método de crecimiento: CHAID Variable dependiente: Primario NoPrimario

Breves conclusiones

Esta técnica basada en la dependencia entre variables, que tiene como objetivo conformar grupos a partir de los valores de la variable independiente y que a su vez sean muy distintos en la variable dependiente (Escobar M. 1991).
En este caso concreto su utilidad (como técnica exploratorio) antecedió a un diagnóstico más detallado y profundo por parte del equipo de economistas.
Una de las grandes ventajas consistió en despejar rápidamente que variables independientes eran las que presentaban mayor poder descriptivo ante este hecho.
Así se descartaron variables tales, el pertenecer al régimen de dedicación total (full time) que se argumentaba como un indicador potente ante la variable dependiente. Otras de variables que en principio no tuvieron impacto u asociación con el fenómeno fueron, el sexo del docente donde las diversas teorías y/o argumentos esgrimían diferencias, el área de conocimiento en el cual existe un marcado consenso de que en el área básica la producción (publicaciones) es un hecho distintivo, ante el resto de las demás áreas, dado que las publicaciones internacionales constituye el intercambio fundamental en dicho campo, el tipo de hogar y el número de hijos con que cuenta el docente, estas variables estaban sustentadas por teorías que hablan de la escasa dedicación y por ende otorgando ciertas diferencias y por último los docentes que presentan otra ocupación, en contraposición a los full time.
En concreto, muchas de las variables que se había pensando como estructurados de los grupos dejaron de tener peso fundamental e incluso relativo, antes las tres que operaron con mayor fuerza para la segmentación. Todo indica a partir de este análisis preliminar que los estudios de posgrados, y el avance de la carrera docente, son dos elementos fundamentales para la producción de los docentes de la UdelaR. Este hecho cobro relativa fuerza, al interior del grupo interdisciplinario, ya que son dos indicadores en los cuales la política universitaria de promoción y formación puede incidir de manera cuasi-directa.
En tanto como sostiene Escobar "……La utilidad del análisis de segmentación es múltiple. Está especialmente diseñado para propósitos descriptivos, exploratorios e incluso pronosticadores. Además, con ciertas cautelas, también puede ser útil para un previo análisis causal de las variables " (Escobar, M. 1998).

Bibliografía básica

- Aaker, D. Day, G. (1993) "Investigación de Mercados" McGraw Hill, Buenos Aires.

- Bayce, Rafael (1983) "La investigación contemporánea en Educación: una evaluación epistemológica de teoría y métodos". CIESU/ ACALI Uruguay

- Belson, A. (1961) "Matching and prediction on the principie of clasification" *Applied Statistics.* Uk.

- Biggs, D., De Ville, B. y Suen E., " A Method of Choosing Multiway Partitions of Clasification and Decsion Tree" In Journal of Applied Statictis pp- 48 - 62 Num.18 1991

- Diez Medrano, Juan (1992) "Métodos de análisis causal". Cuadernos metodológicos del CIS. N° 3. España

- Escobar, Modesto (1998) Las aplicaciones del análisis de segmentación: El procedimiento Chaid. *EMPIRIA Revista de Metodología de Ciencias Sociales* N°1 1998 pp 13 a 49 España.

- Fowler , Floyd J (1995) "Improving Survey Questions Design and Evaluation Applied" Social Reserch Methods . V 38 SAGE United Kingdom

- García Ferrando, M. (1985) "Estadística descriptiva III: Tres o más variables" En Socioestadística : Introducción a la Estadística en Sociología" Editorial Alianza Universidad Textos. España.

- Madgison, J. (1993) "SPSS for Windows chaid release 6.0" Chicago SPSS Inc.

- Primer Censo de personas privadas de libertad Publicaciones web de Ministerio del Interior.

- Ruiz Maya L (et.al) (1990) "Metodología estadística para el análisis de datos cualitativos". CIS España

- SPSS Inc. Manual de User´s Guide USA.

- Zaltman, G y Burger, P. C. "Investigación de Mercados. Principios y dinámica". Editorial Hispano Europea, España. 1980.