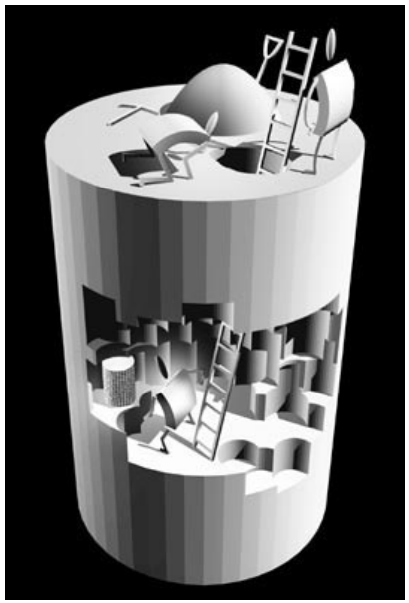
Minería de Datos Aplicada a la Encuesta Permanente de Hogares



Disertante:
Luis Alfonso Cutro
Adscripto a la asignatura
*"**D**iseño y **A**dministración de **D**atos".*

Prof. Coordinador:
Mgter. David Luís la Red Martínez

Conceptos de minería de datos

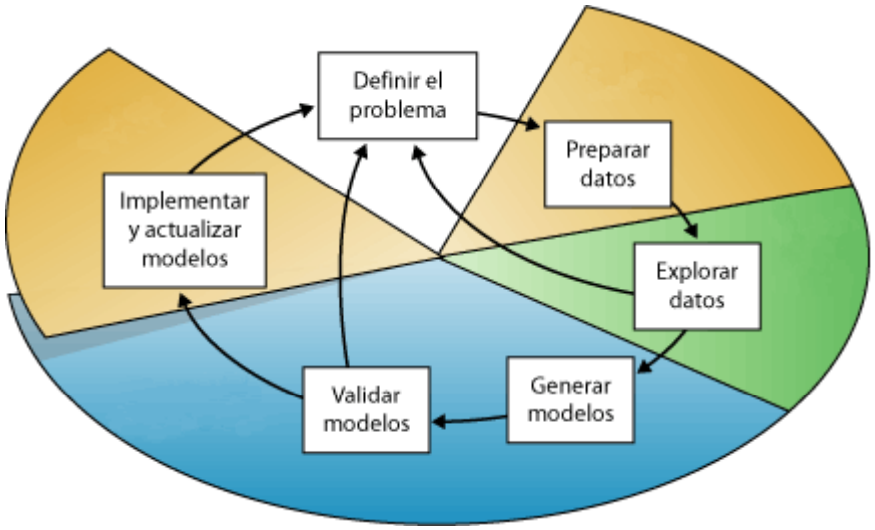
- **La *minería de datos* suele describirse como *"el proceso de extraer información válida, auténtica y que se pueda procesar de las bases de datos de gran tamaño."* En otras palabras, la minería de datos deriva patrones y tendencias que existen en los datos.**
- **Estos patrones y tendencias se pueden recopilar y definir como un modelo de minería de datos.**

Conceptos de minería de datos

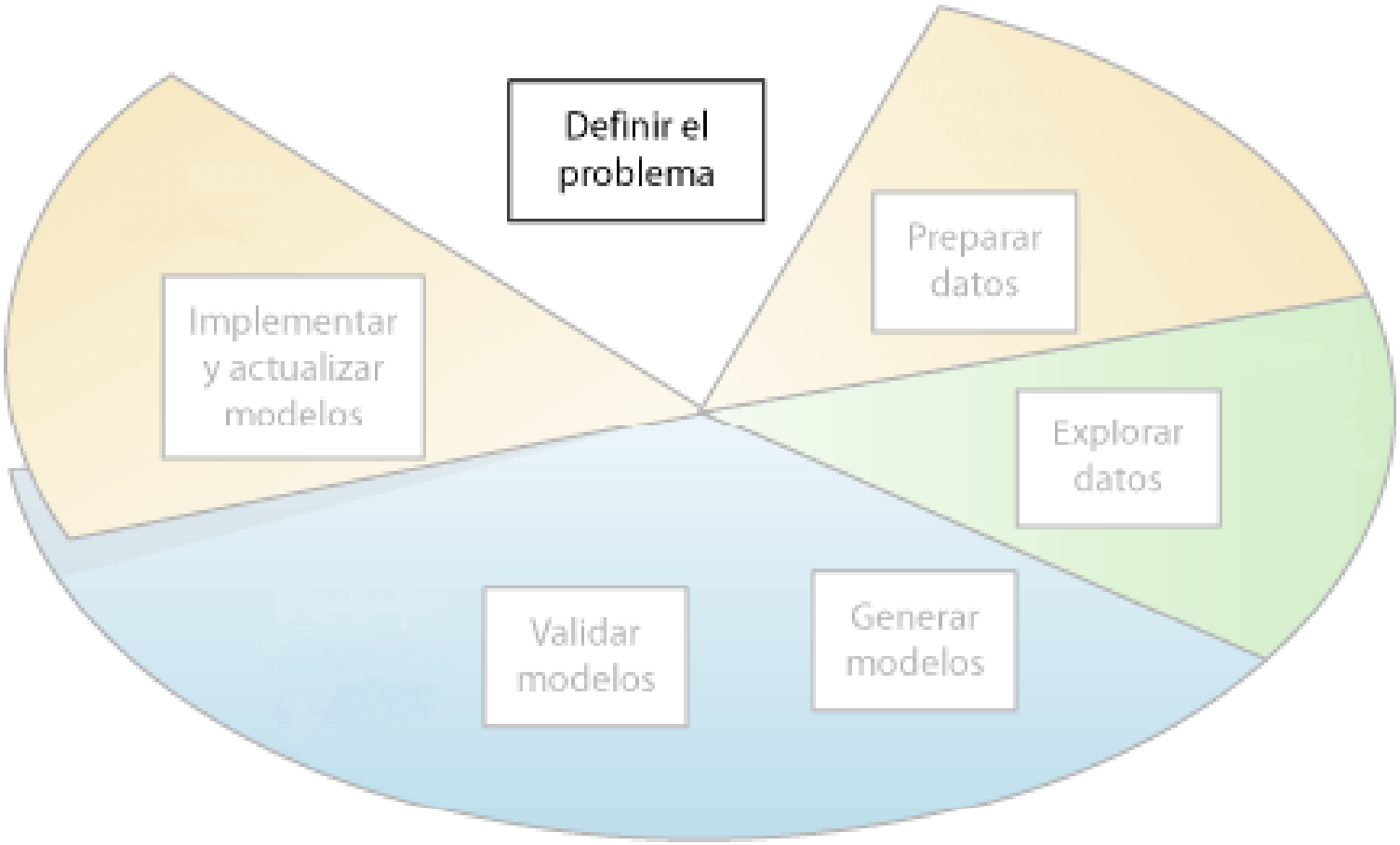
- **Los modelos de minería de datos se pueden aplicar a situaciones empresariales como las siguientes:**
- Predecir ventas.
- Dirigir correo a clientes específicos.
- Determinar los productos que se pueden vender juntos.
- Buscar secuencias en el orden en que los clientes agregan productos a una cesta de compra.

Etapas de Proceso de Minería de Datos

- **Definir el problema**
- **Preparar los datos**
- **Explorar los datos**
- **Generar modelos**
- **Explorar y validar los modelos**
- **Implementar y actualizar los modelos**



Definir el problema



Definir el problema

- Este paso incluye analizar los requisitos de la organización, definir el ámbito del problema y definir el objetivo final del proyecto de minería de datos.
- Estas tareas se traducen en preguntas como las siguientes:
- ¿Que se está buscando?.
- ¿Que atributo del conjunto de datos se desea intentar predecir?.
- ¿Como se distribuyen los datos?.

Definir el problema

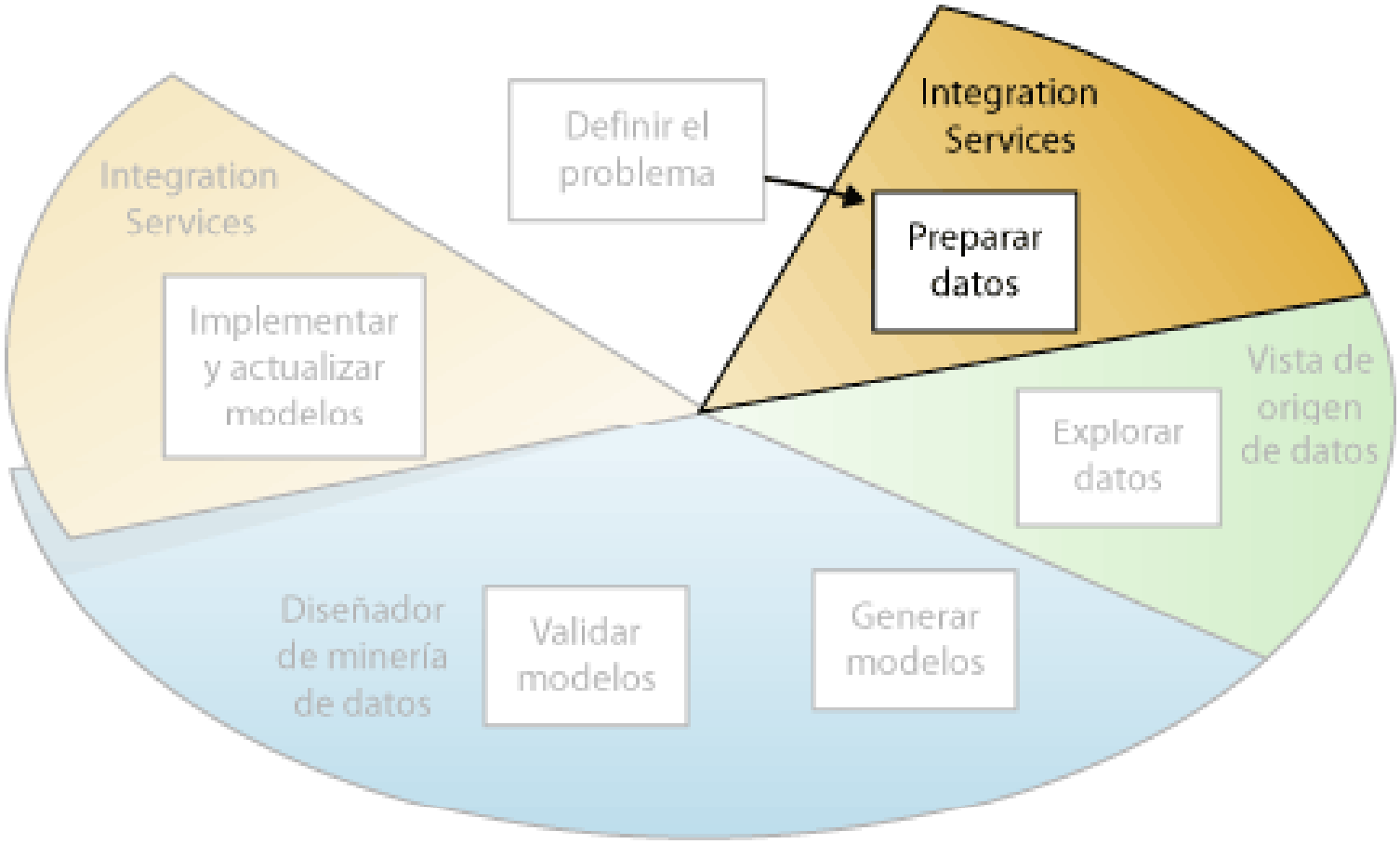
- Problema: extracción de patrones socio - demográficos, educativos y de ingresos de la provincia de Corrientes que se hallan ocultos en la Encuesta Permanente de Hogares EPH.
- Objetivos Generales: caracterizar y describir el empleo público de la Provincia de Corrientes a través de la utilización de técnicas de Minería de Datos.

Definir el problema

Objetivos Específicos:

- Describir la composición del empleo en Corrientes.
- Conocer los perfiles sociodemográficos de los Planes Jefes y Jefas.
- Indagar los perfiles educativos de los Planes Jefes y Jefas.
- Clasificar a los individuos, a partir de sus principales características académicas.

Preparación de los Datos



Preparación de los Datos

- Inicialmente se dispone de 12 bases de datos (a partir del primer trimestre del 2003 al primero de 2007) en el formato *Microsoft Access.*
- La misma contiene información de la nueva ***EPH (Encuesta Permanente de Hogares),*** *cuya muestra incluye 25.000 familias de las 28 aglomerados urbanos de la República Argentina con una frecuencia de cada tres meses.*

Preparación de los Datos

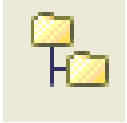
En el ambiente de trabajo del *Centro de depósito de datos*  se deberá definir:

- Fuentes de depósitos.
- Destino de depósitos.
- Esquemas de depósitos.
- Administración.

Fuentes de depósitos

- Inicialmente se dispone de 12 bases de datos (a partir del primer trimestre del 2003 al primero de 2007) en el formato *Microsoft Access.*
- La misma contiene información de la nueva ***EPH (Encuesta Permanente de Hogares),*** *cuya muestra incluye 25.000 familias de las 28 aglomerados urbanos de la República Argentina con una frecuencia de cada tres meses.*

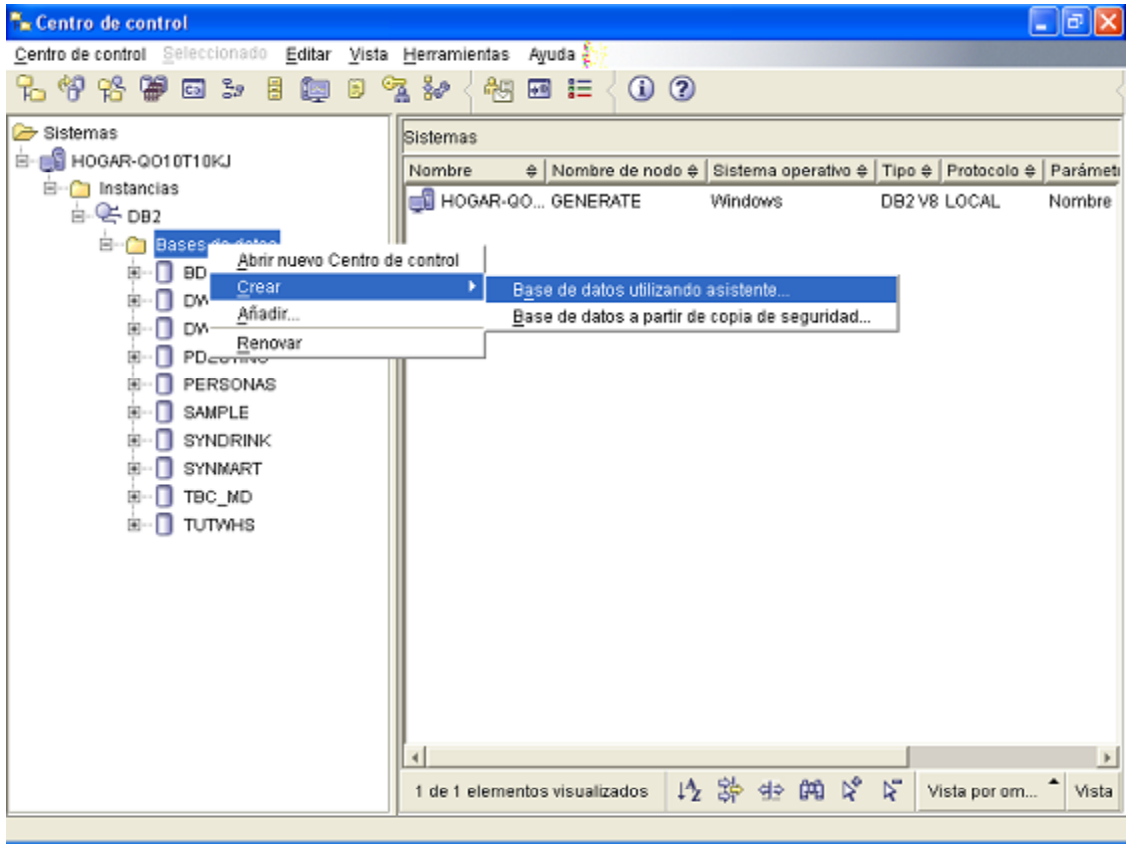
Fuentes de depósitos

Los pasos que se llevan a cabo en el ambiente de trabajo  *Centro Control* :

- Creación de la base de datos denominada EPH (Encuesta Permanente Hogares).
- Creación de la tabla USP.
- Importación de datos.
  - Microsoft Access → IBM DB2 UDB.
- Visualización del muestreo del contenido.

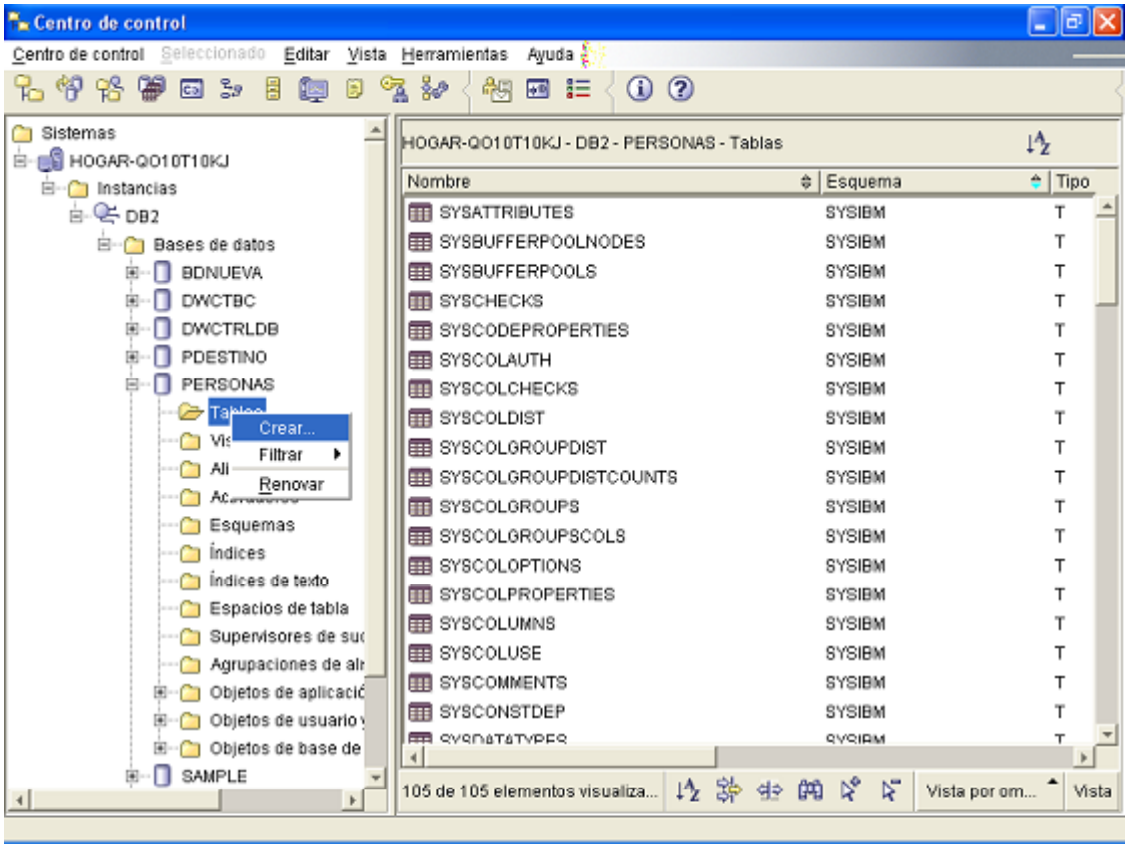
Fuentes de depósitos

- Creación de la base de datos denominada EPH.



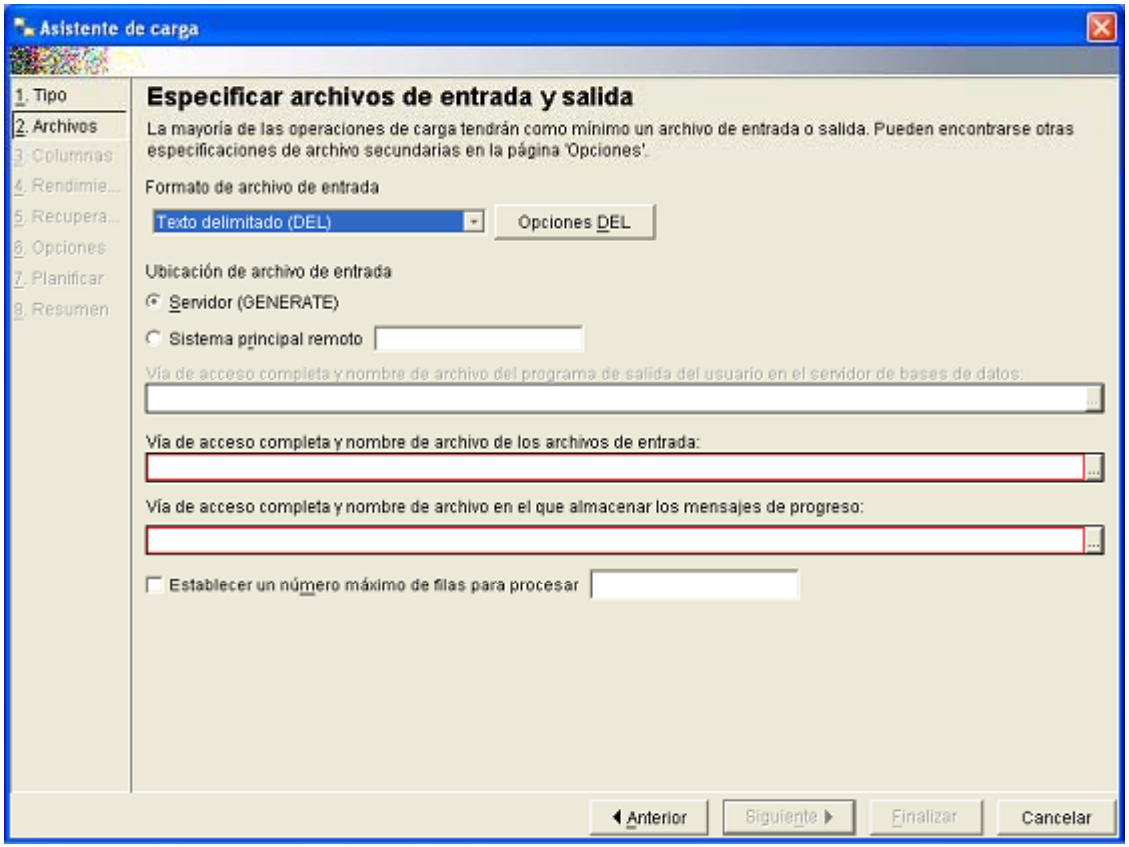
Fuentes de depósitos

- Creación de la tabla USP



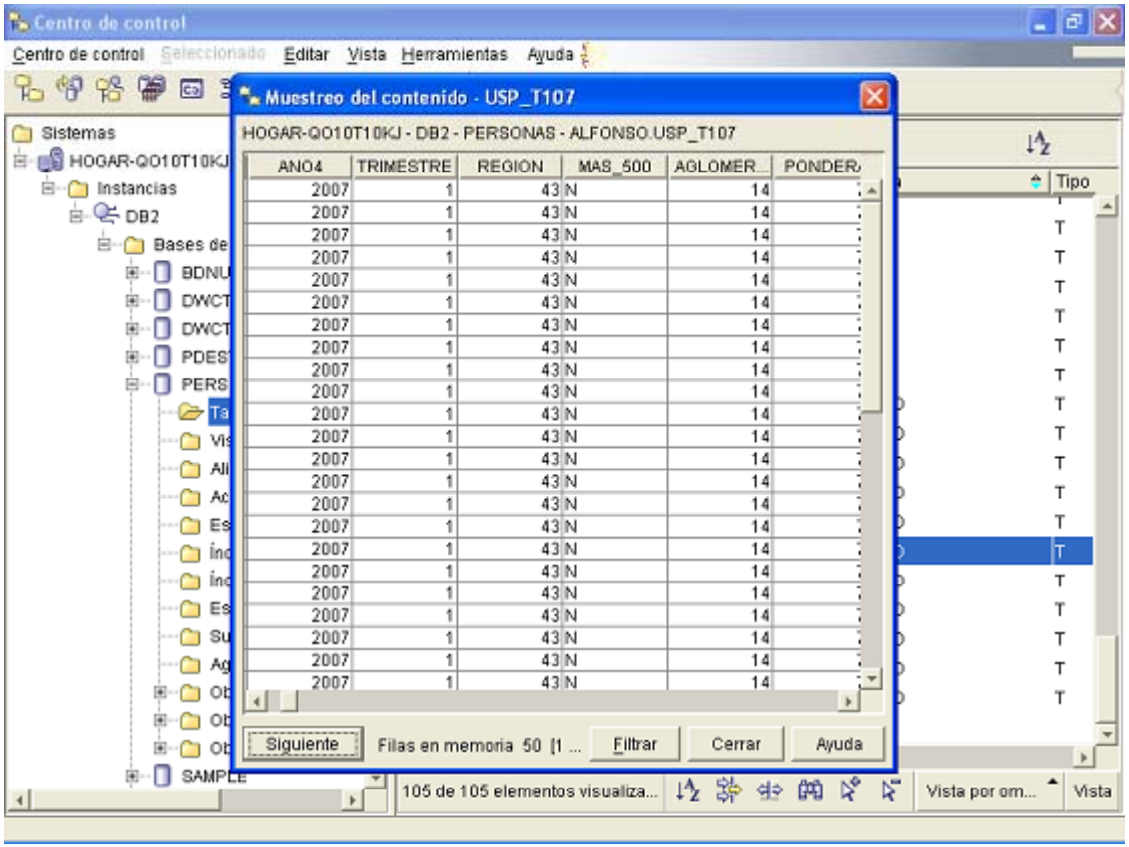
Fuentes de depósitos

- Importación de datos



Fuentes de depósitos

- Visualización del muestreo del contenido



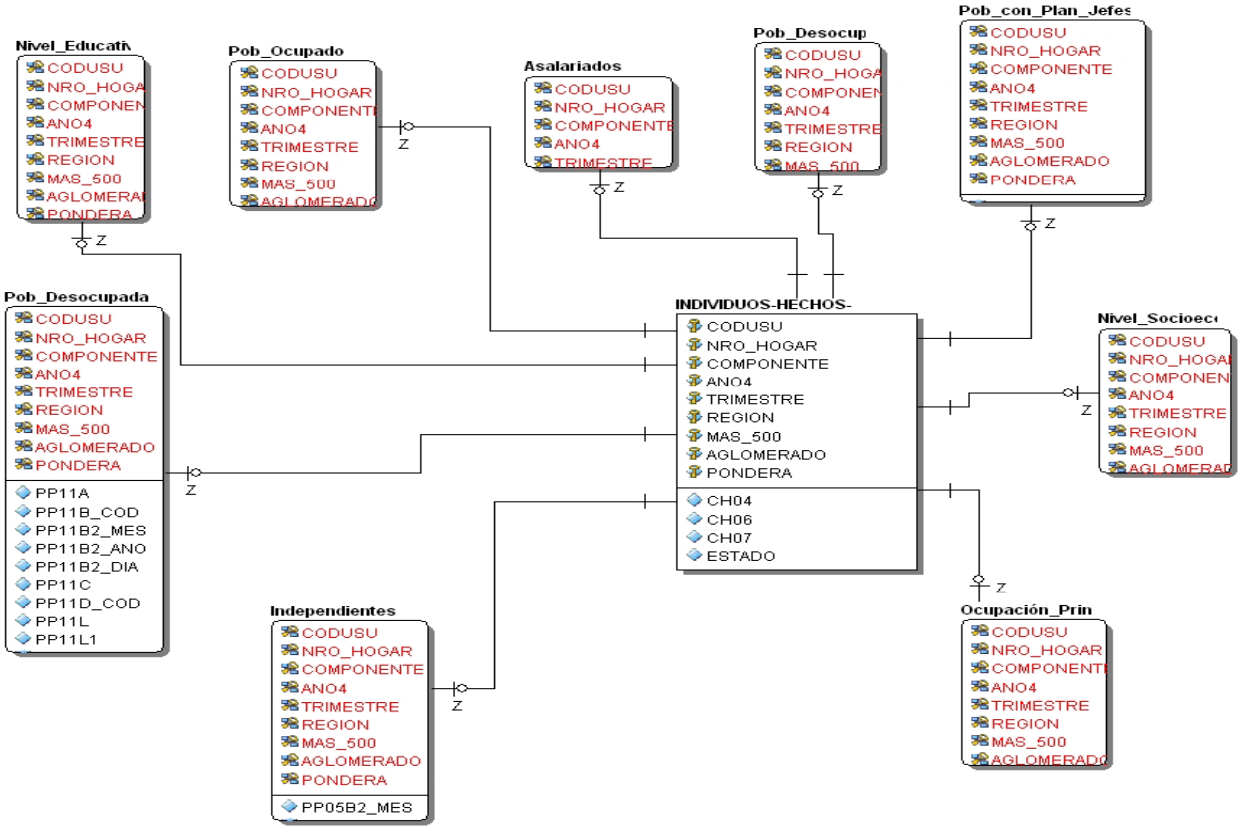
Destino de depósitos.

Los destino de depósitos son las siguientes Tablas o Dimensiones:

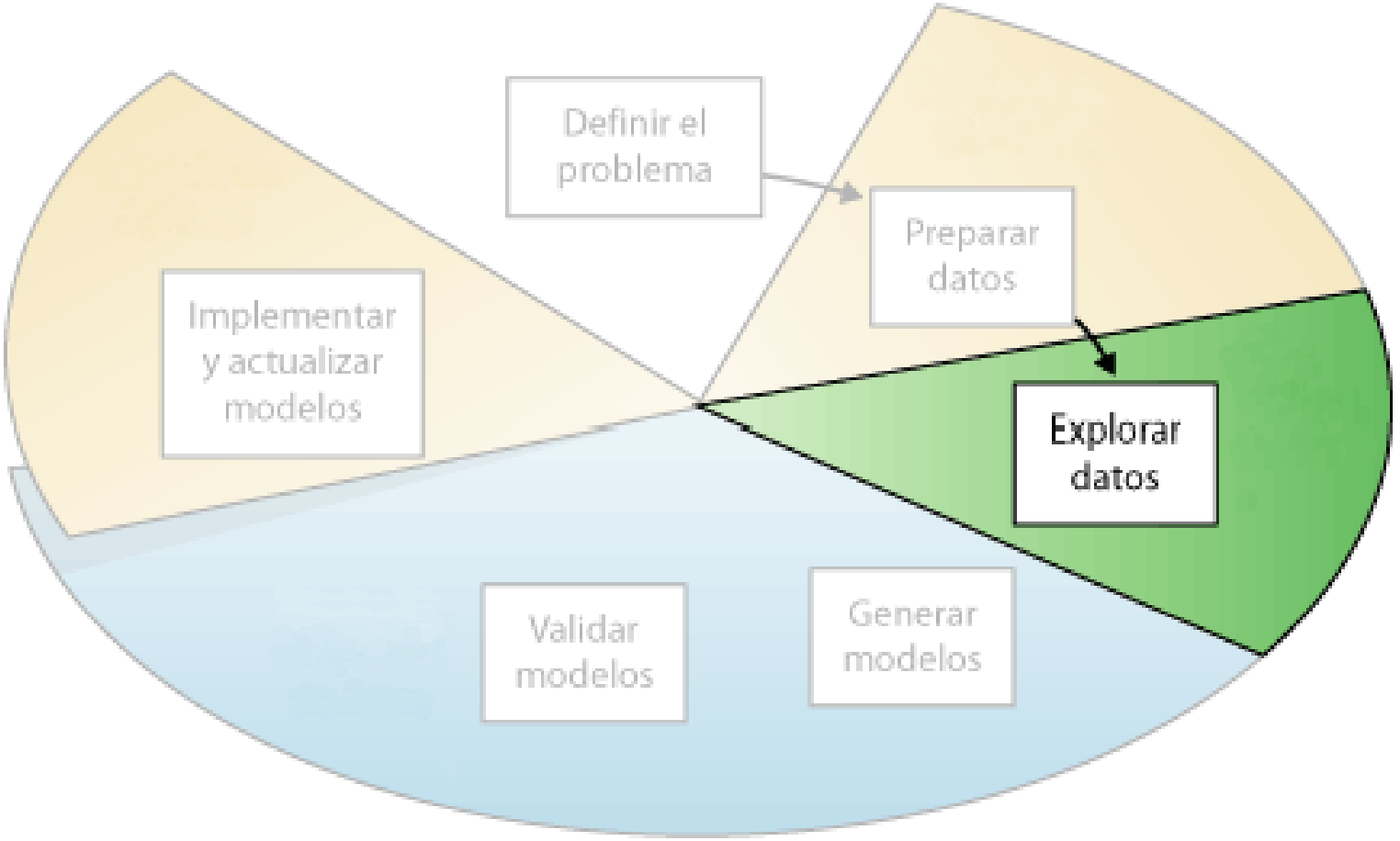
- Nivel Educación
- Población de Asalariados
- Población de Independientes
- Población Desocupada
- Población Desocupada c/Empleo Anterior
- Población Ocupados
- Ocupación Principal
- Individuos

Esquemas de depósitos

- Esquemas de Depósitos



Explorar los datos



Explorar los datos

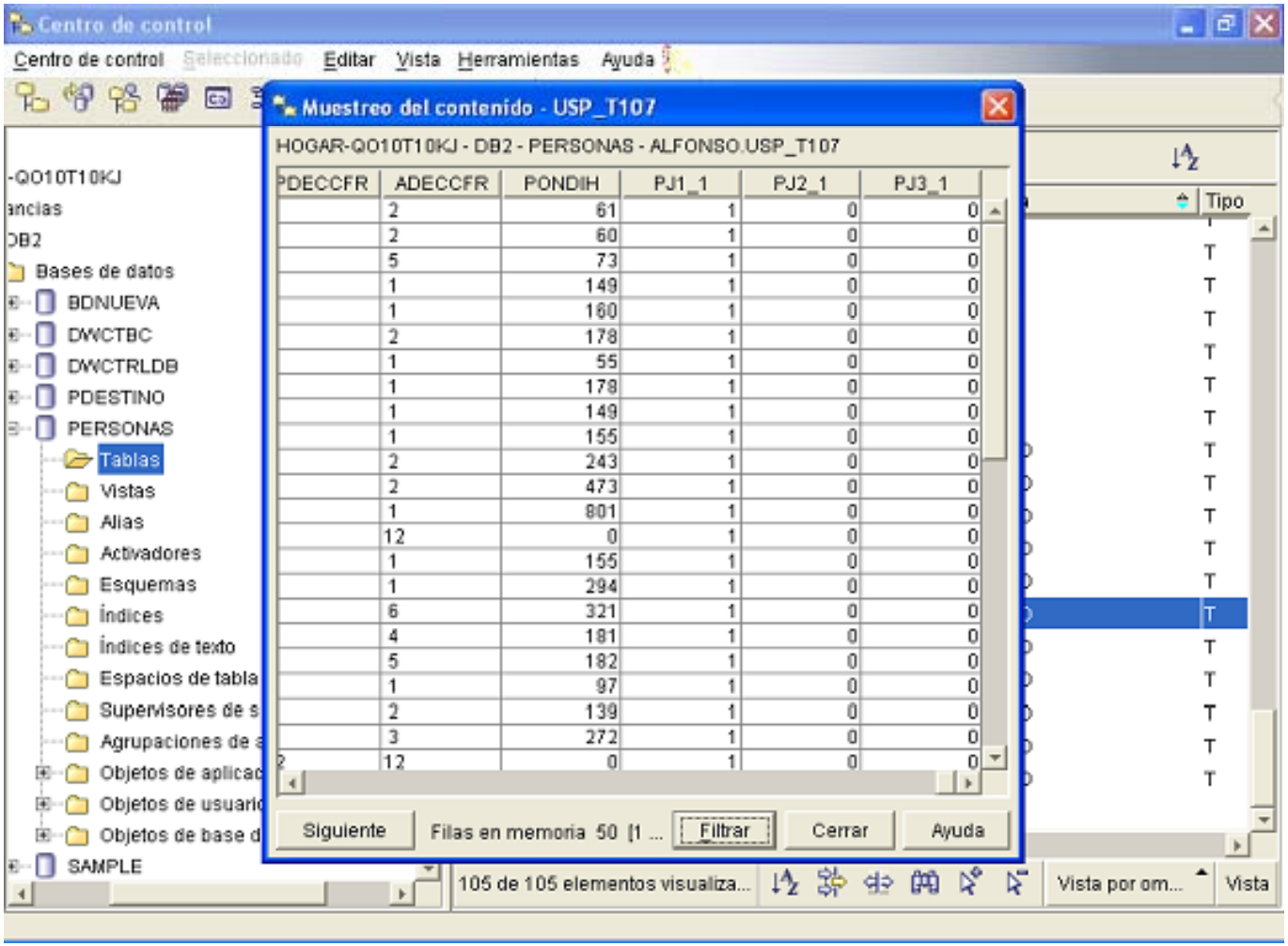
- Como se hizo mención al principio de este apartado, ``la creación de un modelo de minería de datos es un proceso dinámico e iterativo".
- Lo que implica que si en el transcurso de la exploración de los datos no se encuentra una coherencia de los datos logrados, sería conveniente volver a redefinir el problema a tratar. Para luego continuar con el ciclo de vida del Proyecto.

Explorar los datos

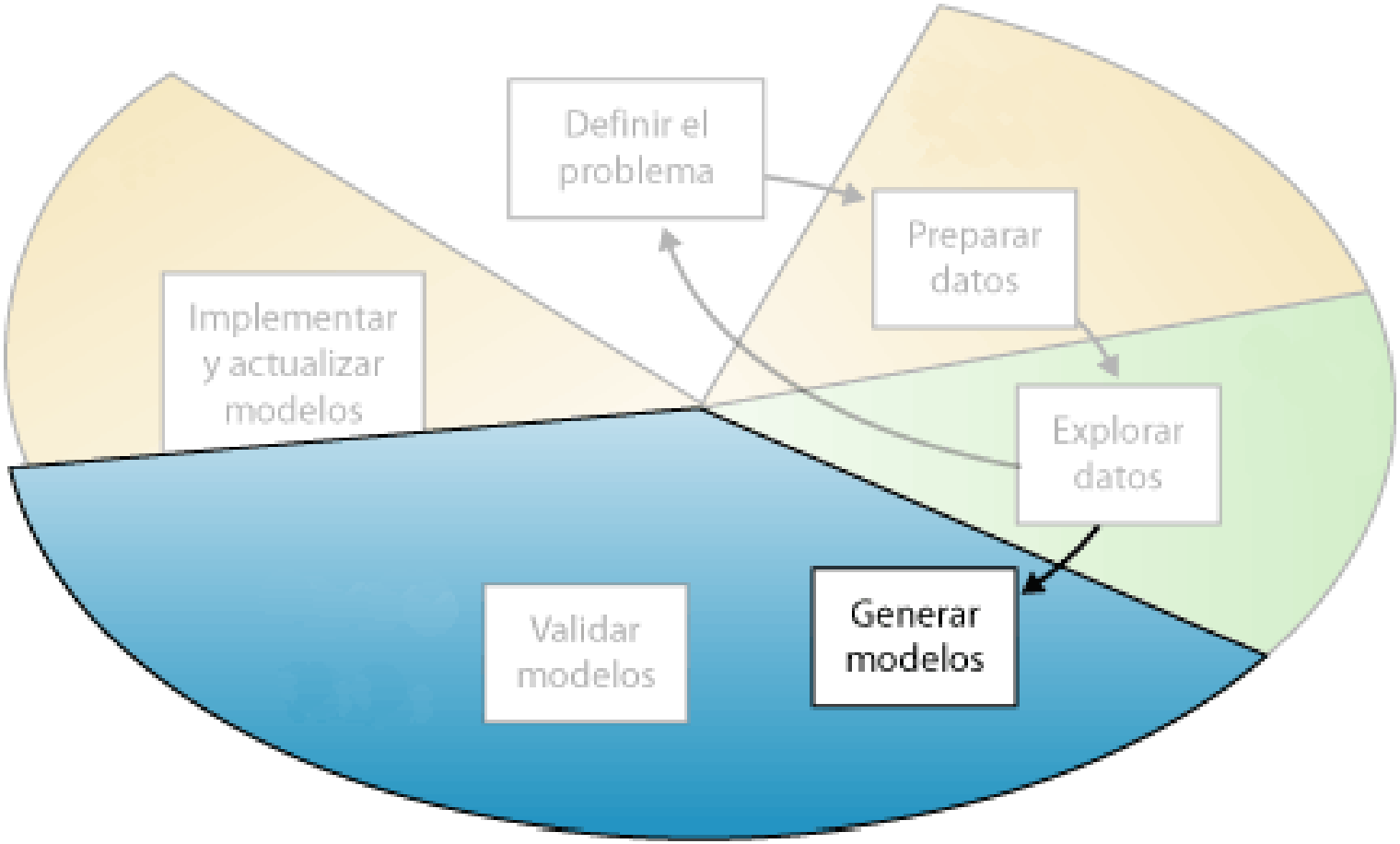
- *Por ejemplo*:
- ✓ ***Conocer los perfiles socio demográficos de los Planes Jefes y Jefas.***

  Se tendría que verificar la existencia de la variable que determina si la persona encuestada es poseedora de ese plan social. Dicha variable en este caso es la PJ1_1.

Explorar los datos



Generar modelos



Generar modelos

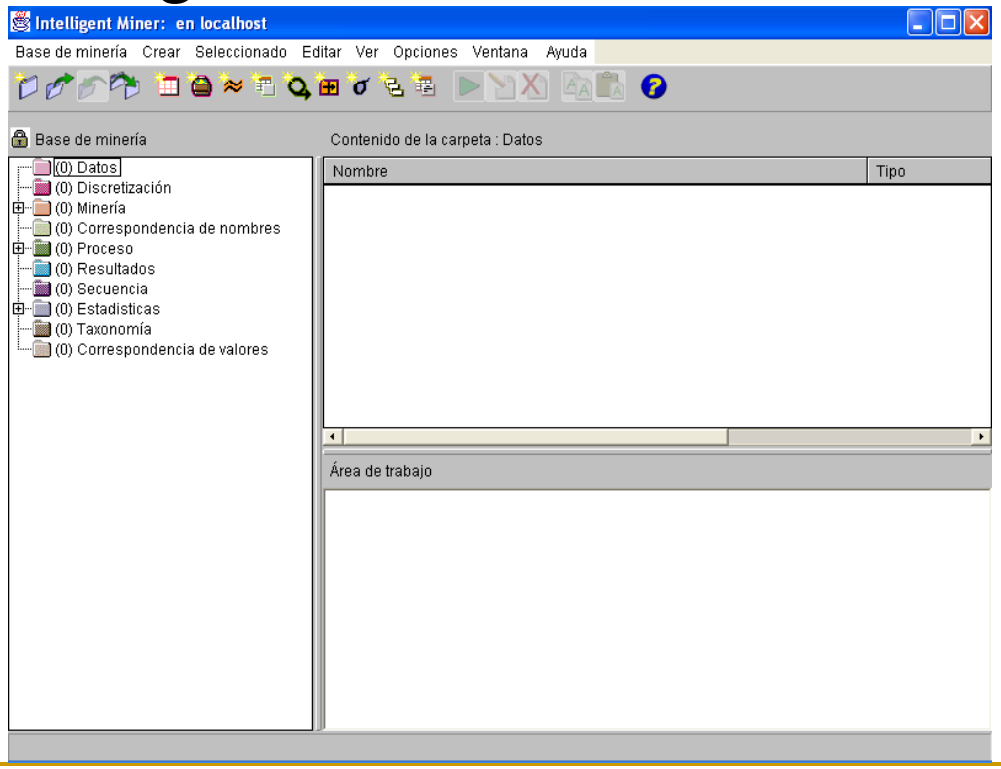
- *Conocer los Perfiles Socio Demográficos de los Planes Jefes y Jefas.*
- *Indagar los Perfiles Educativos de los Planes Jefes y Jefas.*
- *Clasificación del Ingreso de Cada Individuo, en Base a sus Principales Características Sociodemográficas.*
- *Clasificación del Ingreso de Cada Individuo, en Base a sus Principales Características Educacionales.*

Generar modelos

- Para los dos primeros se utilizaran la técnica de Minería de Datos, *Clustering.*
- Y las demás emplearan la técnica de *Árboles de Decisión.*

Generar modelos

- Iniciación con el ambiente de trabajo *IBM DB2 Intelligent Miner for Data*



Generar modelos

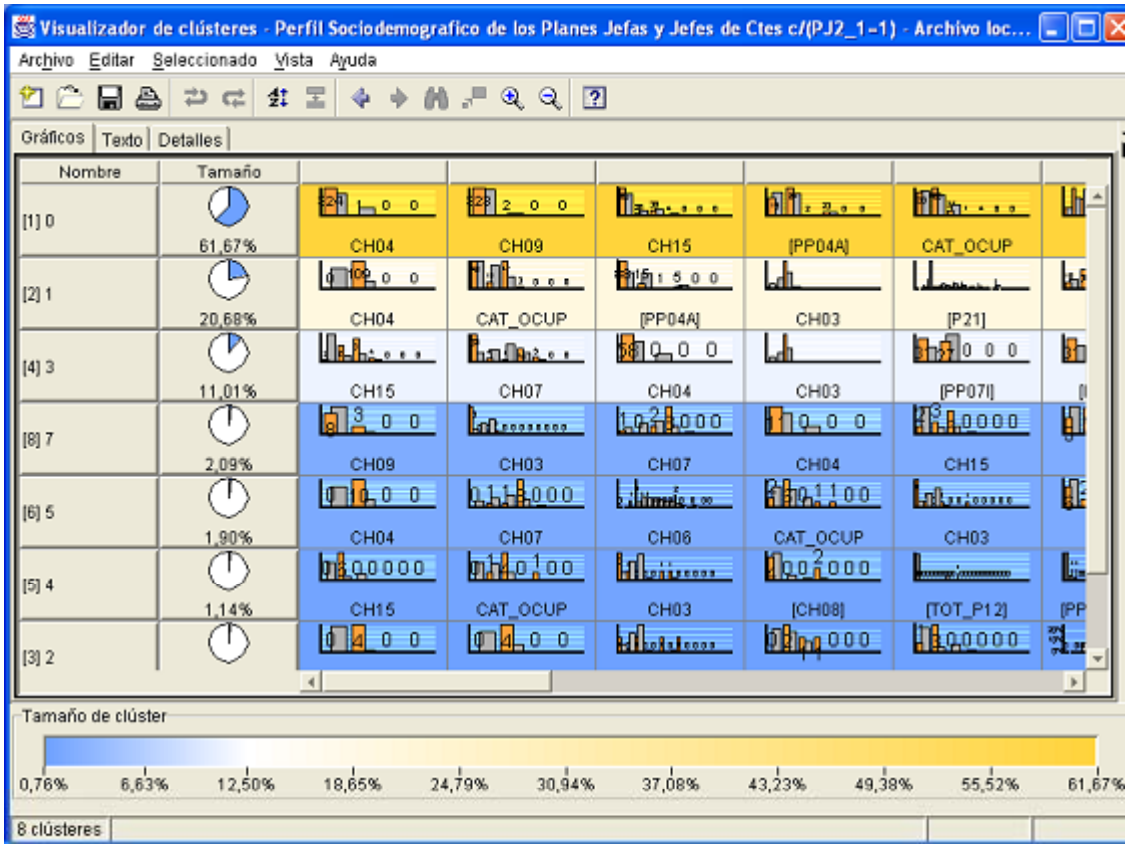
Se nota la existencia de 8 clusters identificados por la ejecución de minería.

En cada cluster, los diagramas y gráficos de barras representan los campos activos y suplementarios utilizados.

Los campos con mayor influencia en la formación del cluster se visualizan a la izquierda (CH15, CH09, CH04, CH07, CH03), mientras que los campos con menor influencia se visualizan a la derecha (PP04A, CH08,etc.).

Generar modelos

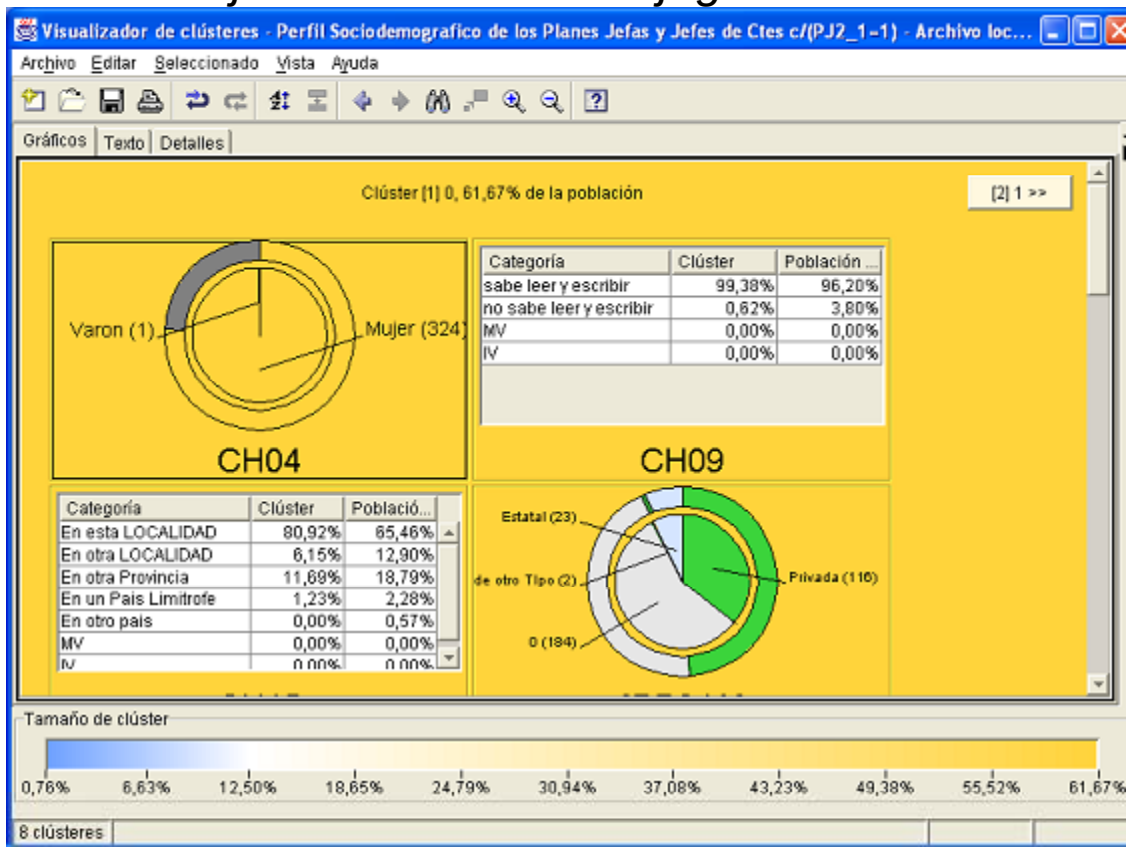
Se nota la existencia de 8 clusters identificados por la ejecución de minería de datos
Donde prácticamente un 93,36% de
la población está representada sólo
por estos tres primeros clústeres, dividiéndose el 6,64% restante
entre los demás.



Generar modelos

Visualización del cada Clúster ***Nº1 con 57.89%:***

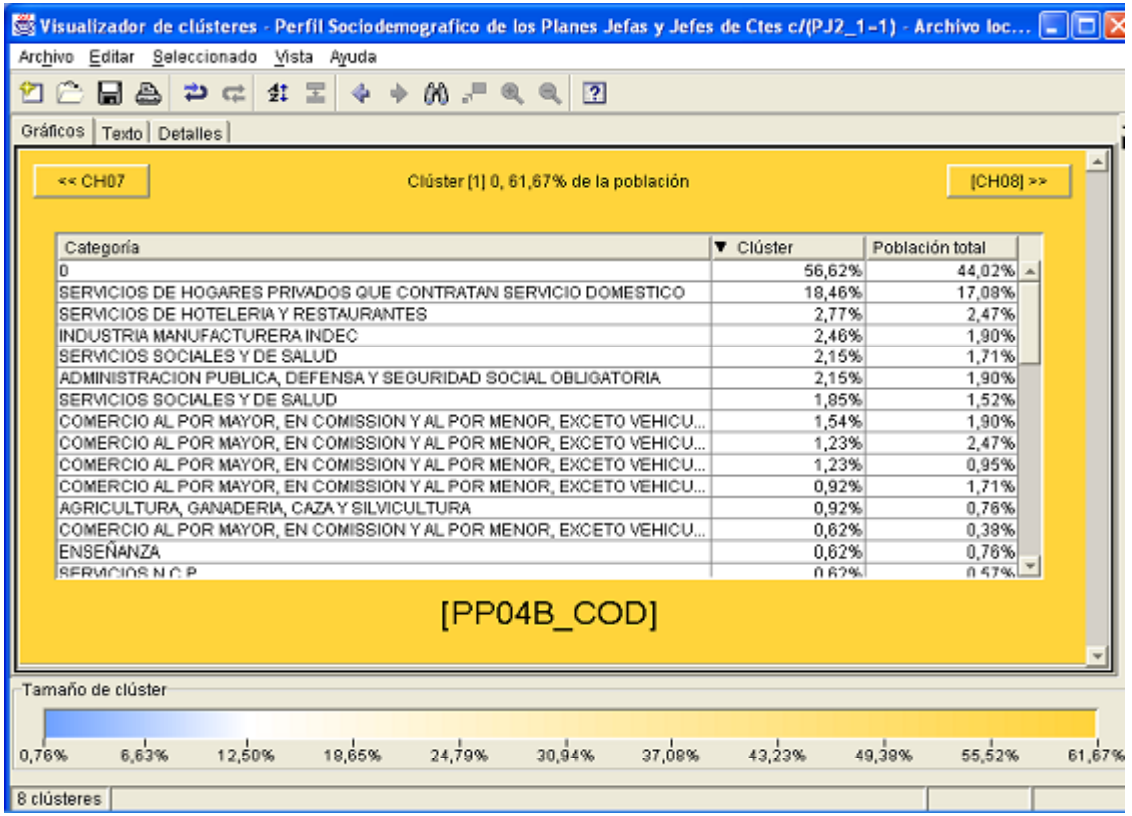
El primer grupo está represento por una población en su mayoría formada por mujeres, de 25 a 30 años de edad, que son residentes de Corrientes Capital y se encuentran unidas o juntadas con su conjugué.



Generar modelos

Visualización del Clúster ***Nº1 con 57.89%:***

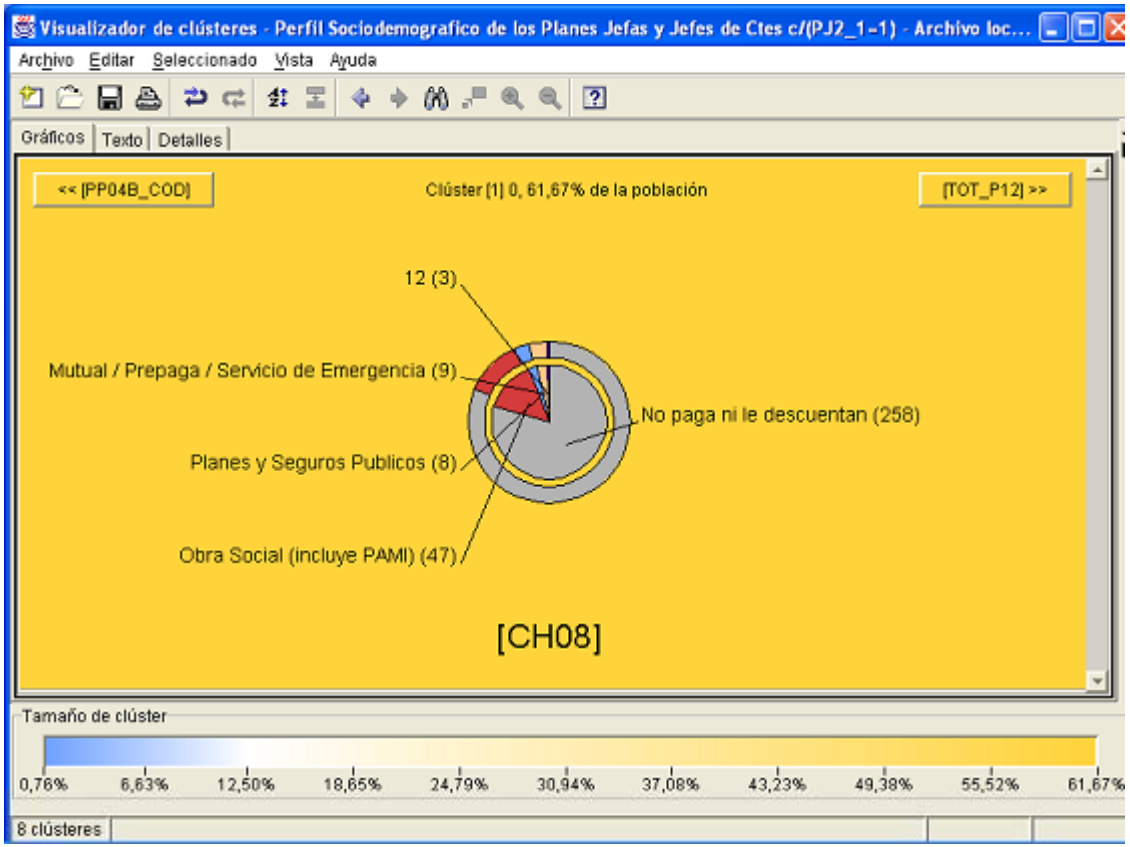
Con respecto a lo laboral, estás personas trabajan en hogares privados como servicio doméstico donde no paga ni le descuentan mensualmente una cobertura médica, tampoco tiene contrato de trabajo ni obra social y mucho menos descuento jubilatorio.



Generar modelos

Visualización del Clúster ***Nº1 con 57.89%:***

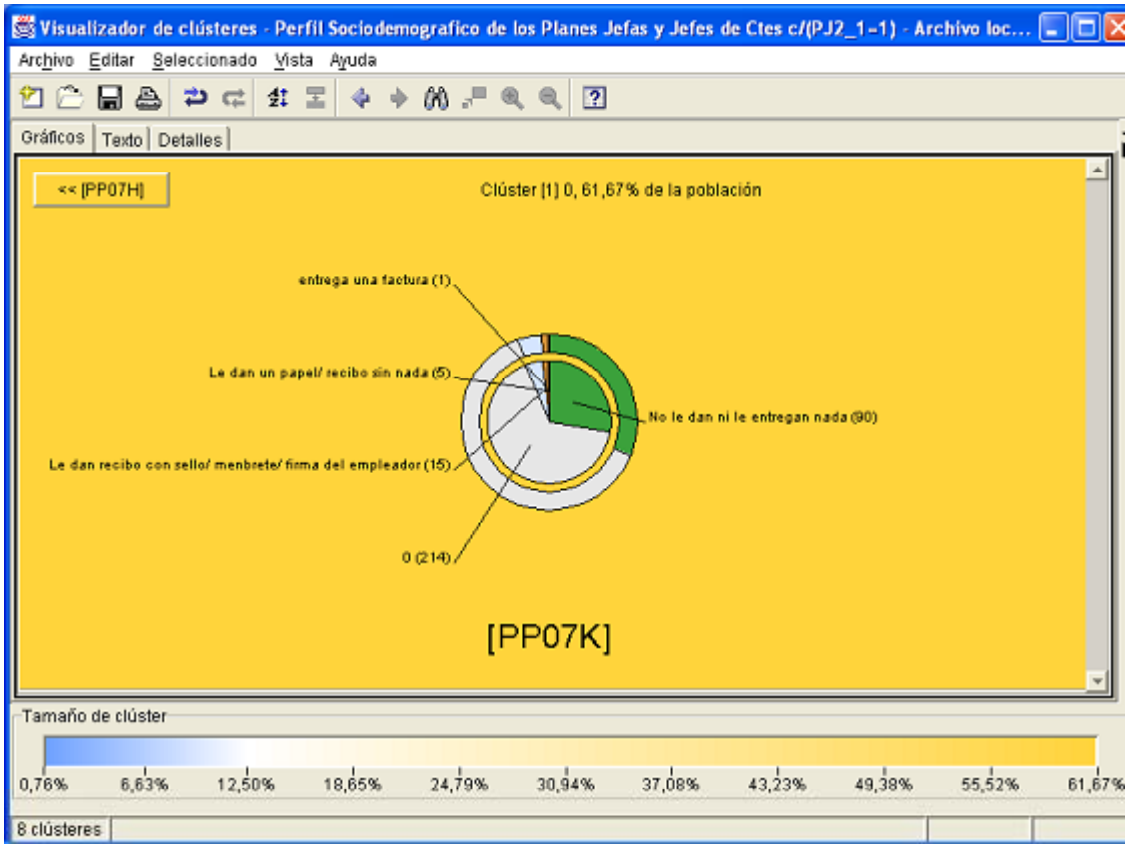
*No paga ni le descuentan mensualmente una **cobertura médica**.*



Generar modelos

Visualización del Clúster ***Nº1 con 57.89%:***

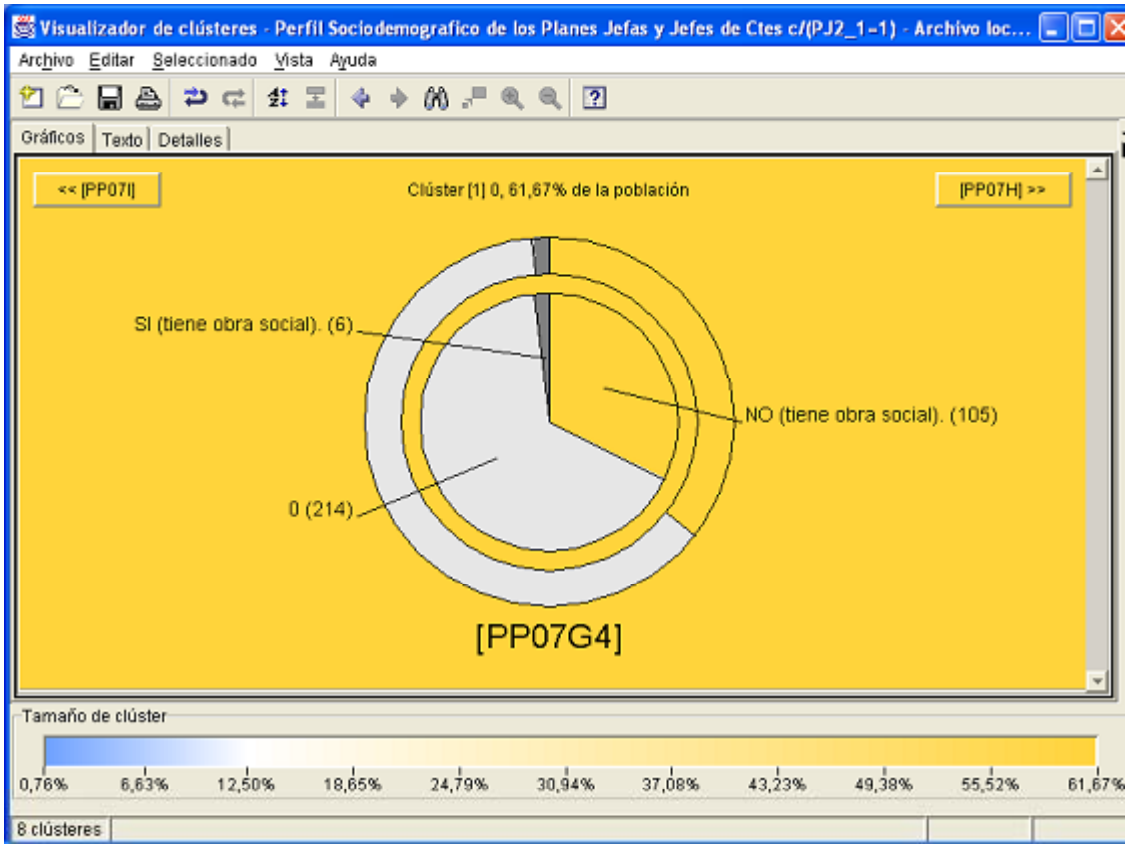
*No poseen **contrato de trabajo** ni obra social.*



Generar modelos

Visualización del Clúster ***Nº1 con 57.89%:***

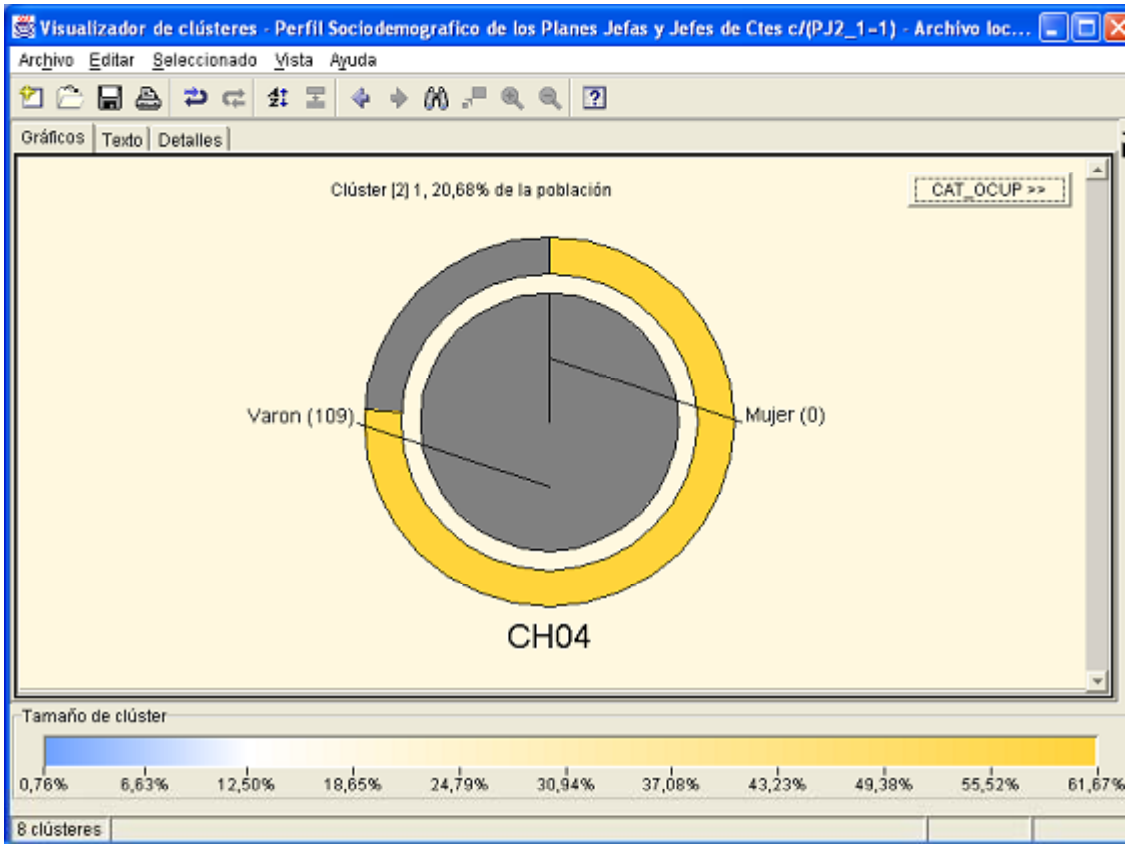
*No poseen contrato de trabajo ni **obra social**.*



Generar modelos

Visualización del Clúster ***Nº2 con 20,68%:***

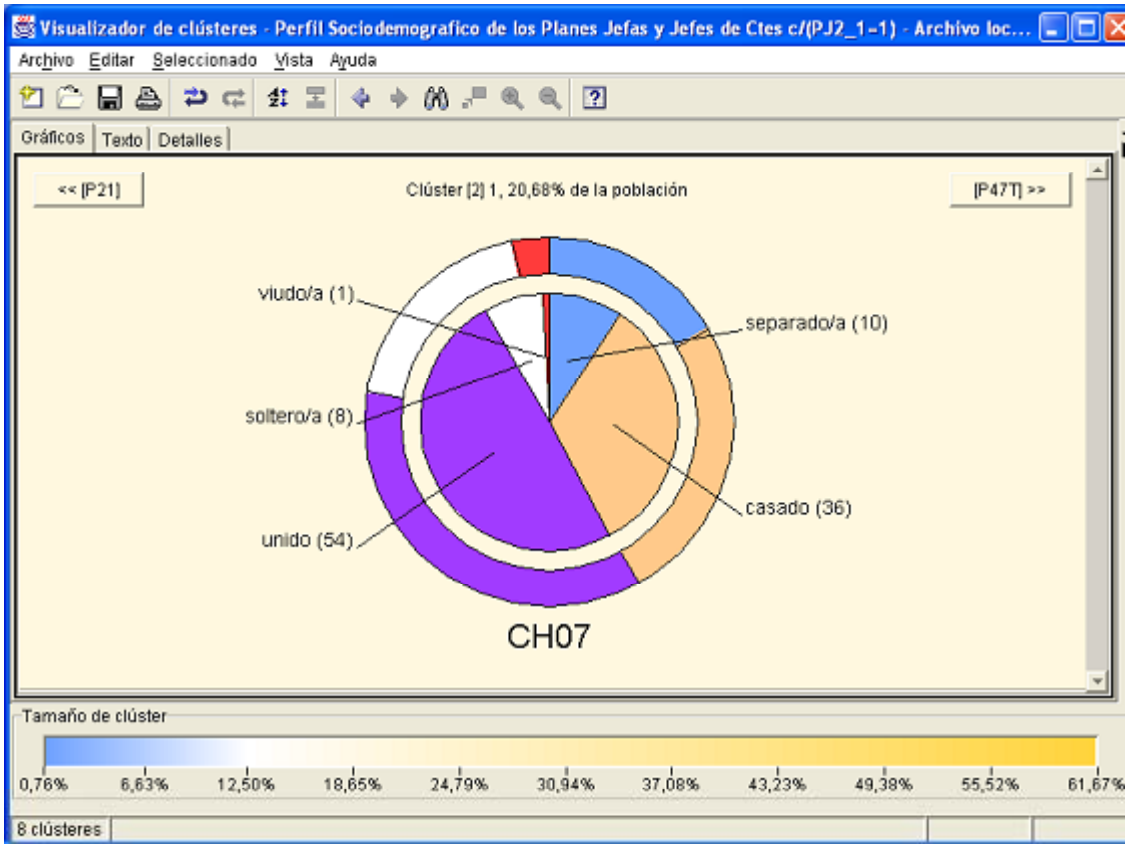
*El sexo predominantemente es el **masculino***



Generar modelos

Visualización del Clúster ***Nº2 con 20,68%:***

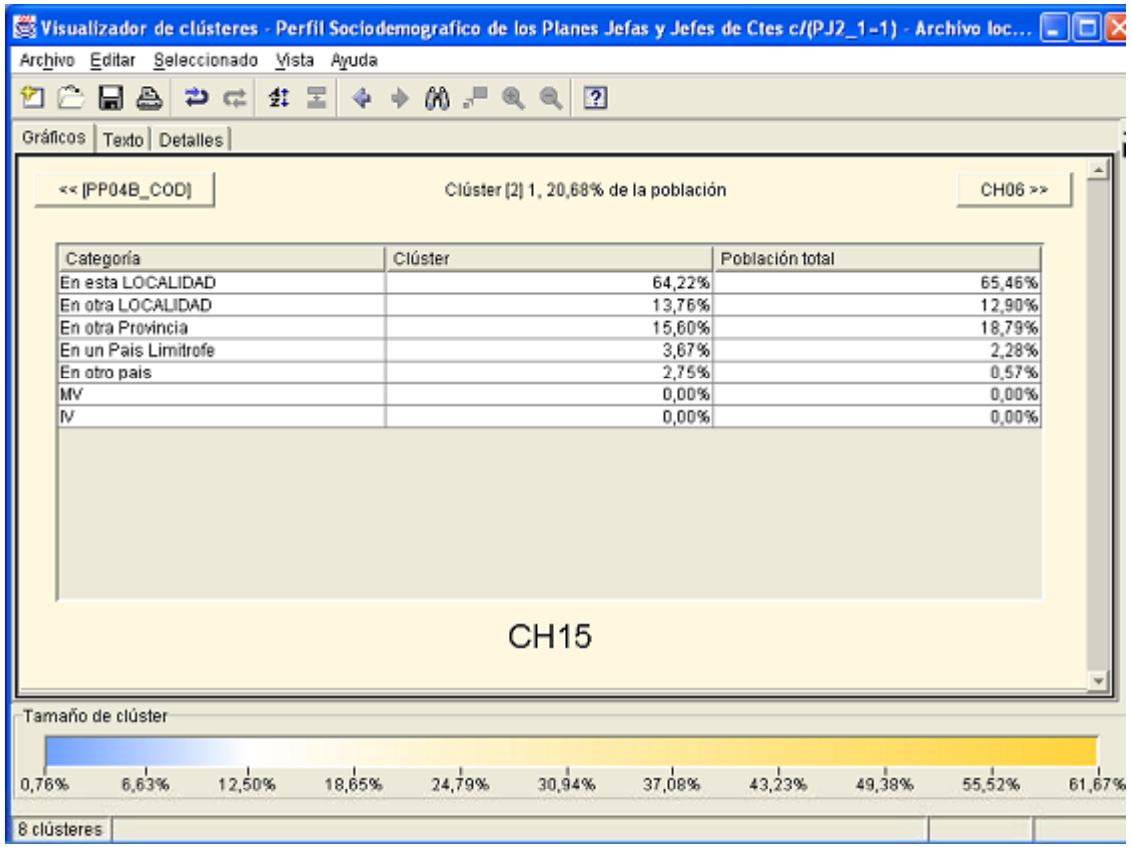
*El **estado civil** de unido y con una **edad** sobresaliente de 46 años.*



Generar modelos

Visualización del Clúster ***Nº2 con 20,68%:***

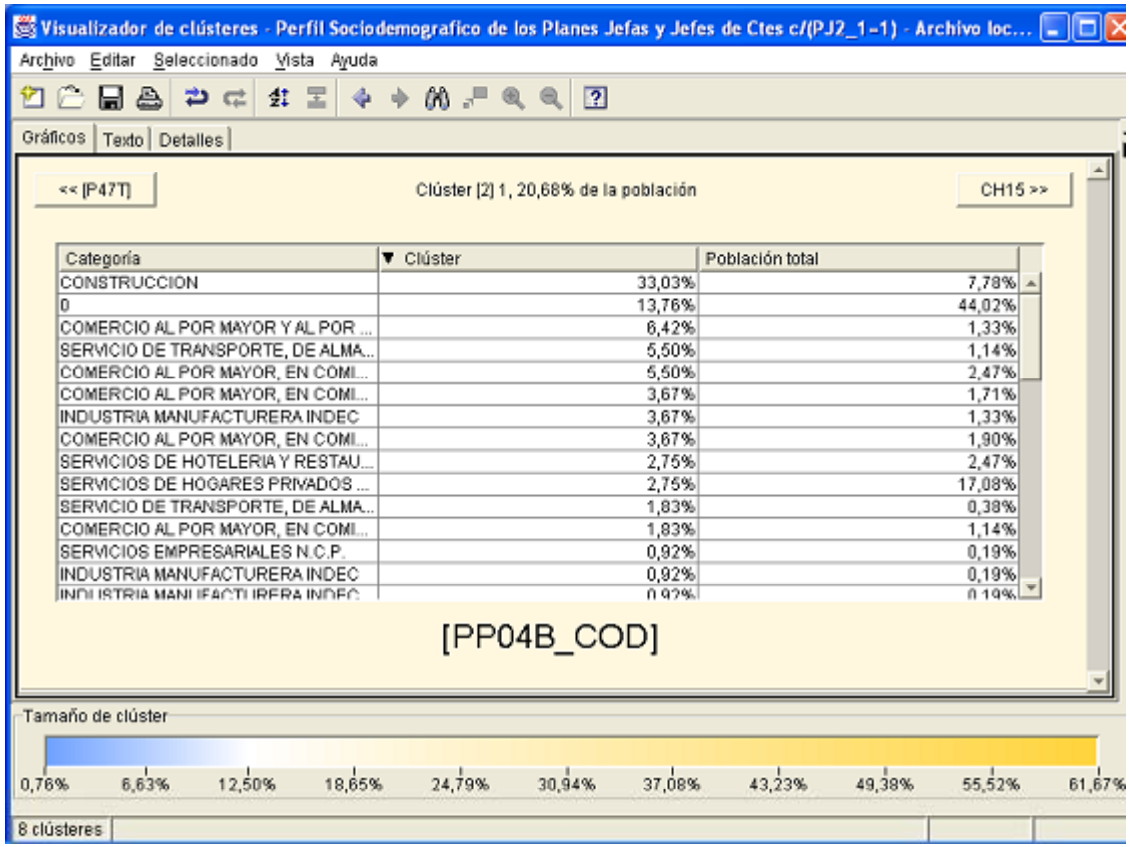
Sin diferenciarse con el primer clúster, en este en su mayoría siguen siendo de esta localidad o sea Corrientes Capital.



Generar modelos

Visualización del Clúster ***Nº2 con 20,68%:***

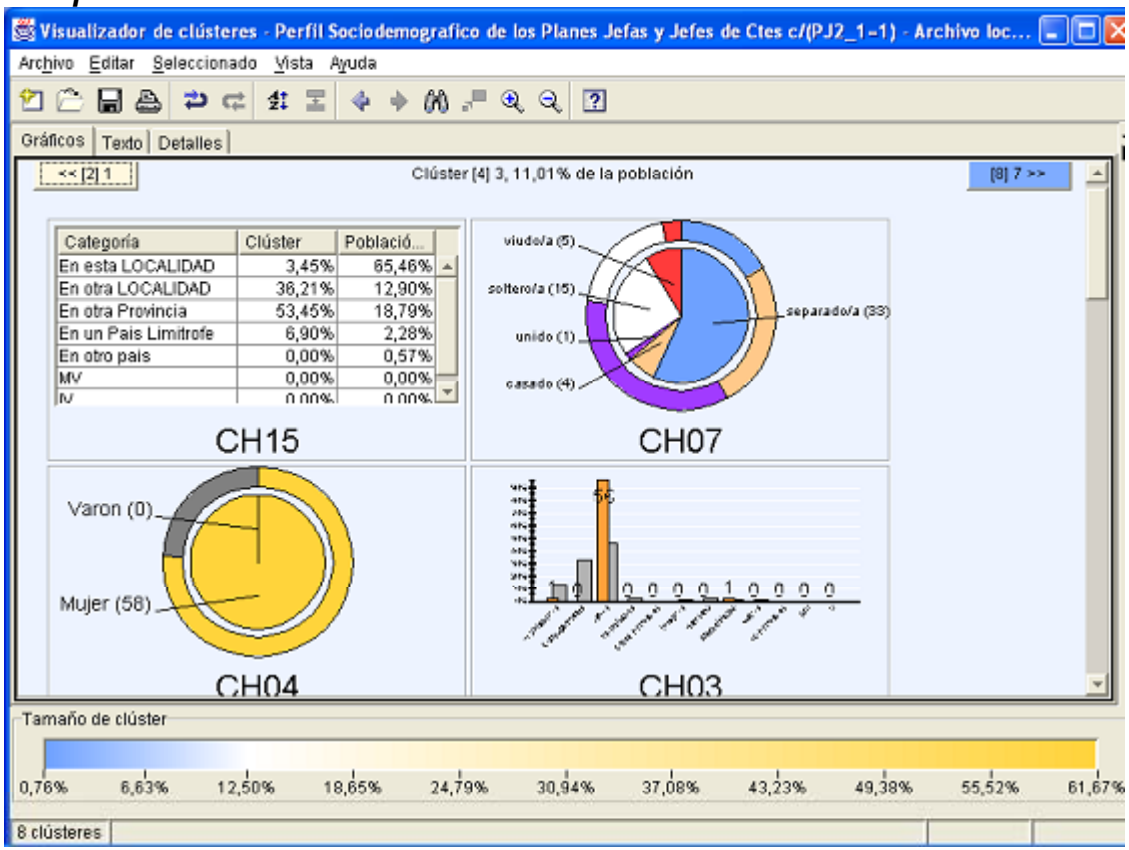
*A diferencia con el primer clúster, en este los individuos se dedican al rubro de la **construcción**.*



Generar modelos

Visualización del Clúster ***Nº3 con 11,01 %:***

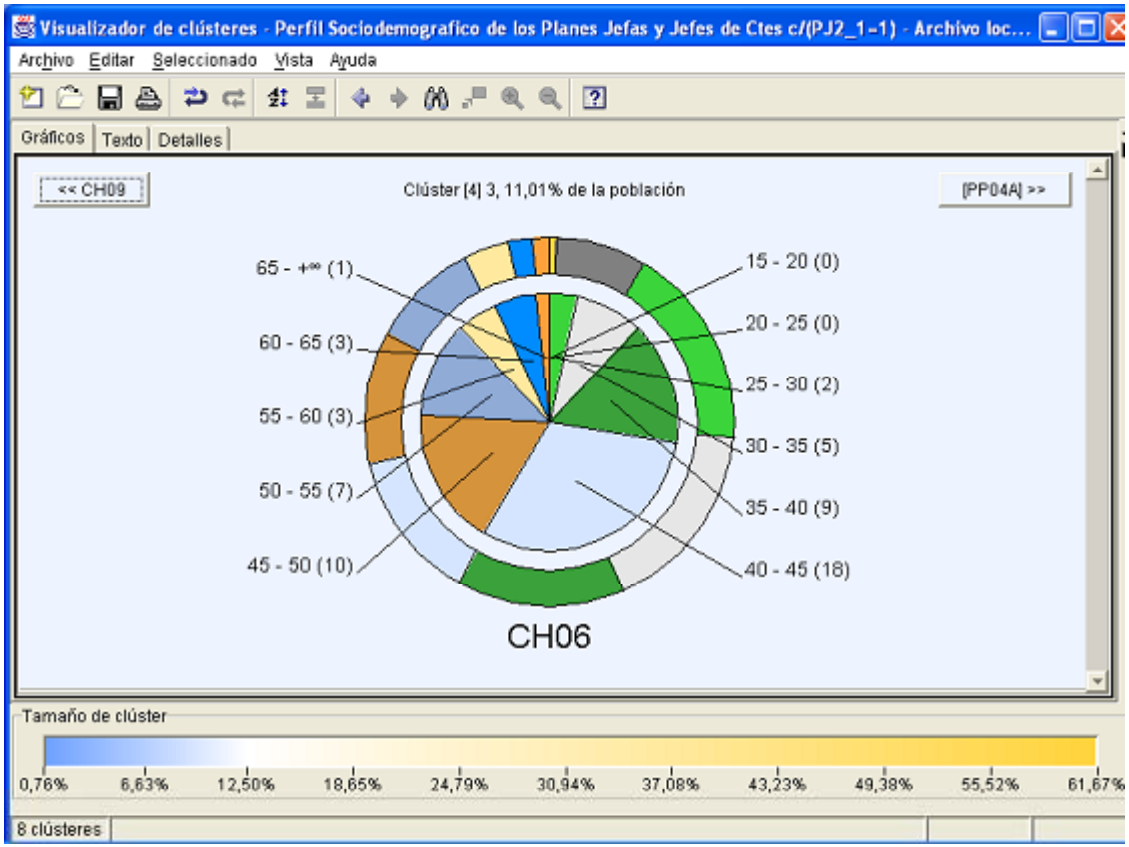
En este clúster se tiene como predominante a la mujer en la variable sexo la misma es separada con una edad que ronda los 40 a 45 años y ha nacido en otra provincia.



Generar modelos

Visualización del Clúster ***Nº3 con 11,01%:***

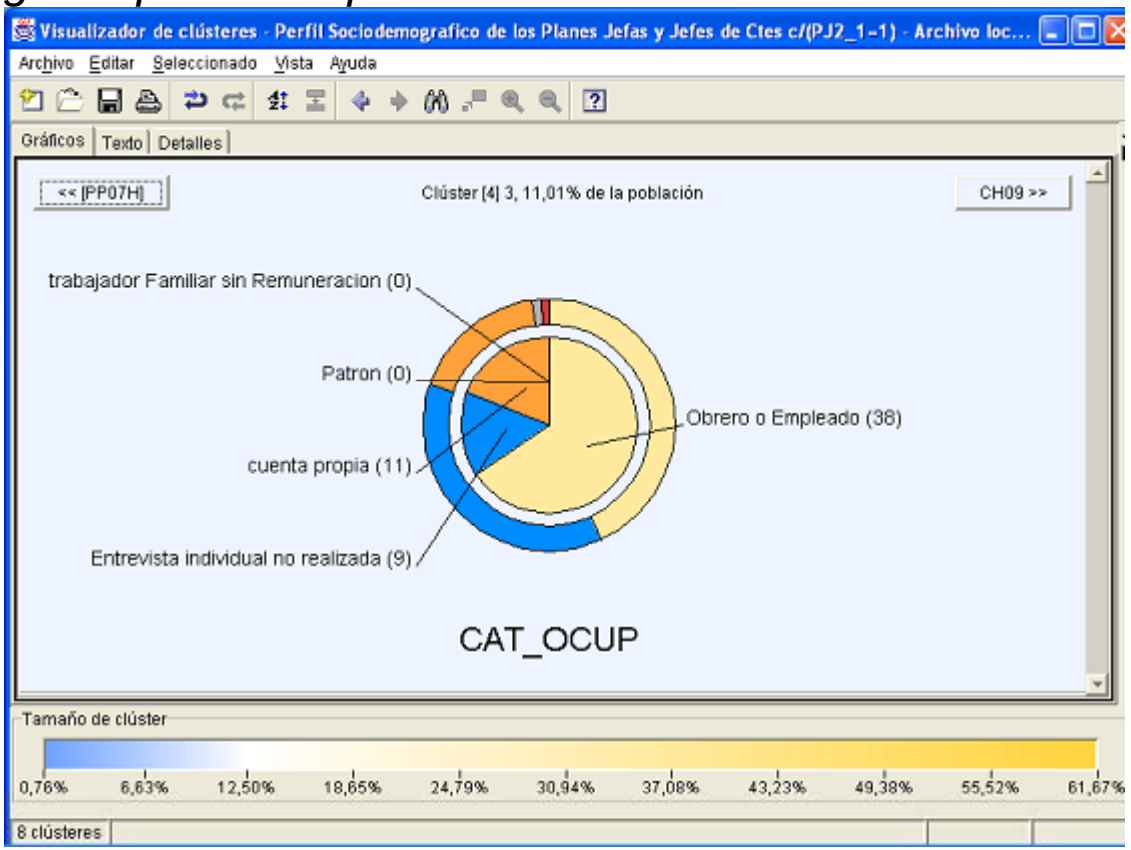
En este diagrama circular se puede observar que el rango de edad con mayor frecuencia es el [40-45].



Generar modelos

Visualización del Clúster ***Nº3 con 11,01%:***

La categoría ocupacional que sobresale es la de ***"obrero o empleado"*** *con un rubro de actividad económica como la "servicios de hogares privados que contratan servicio domestico".*

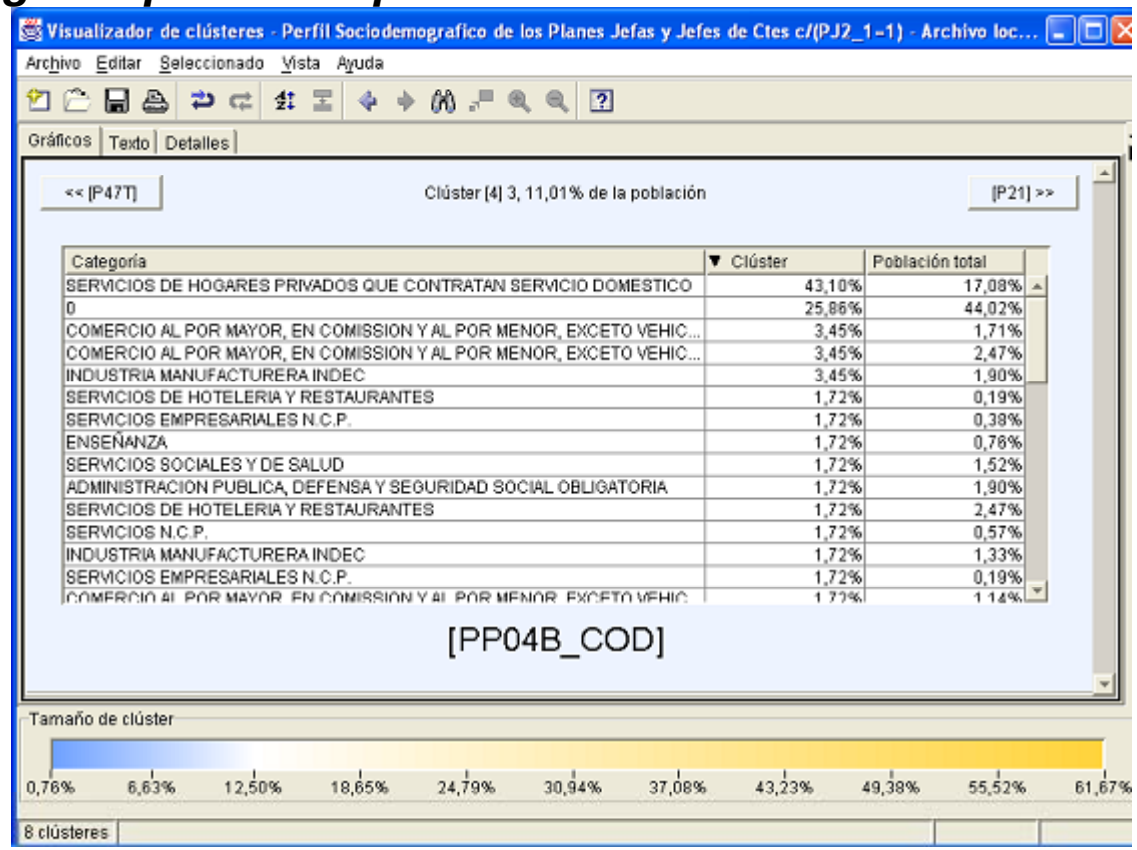


Generar modelos

Visualización del Clúster ***Nº3 con 11,01%:***

La categoría ocupacional que sobresale es la de "obrero o empleado" con un rubro de actividad económica como la

"servicios de hogares privados que contratan servicio domestico".



Generar modelos

- ***Clasificación del Ingreso de Cada Individuo en Base a sus Principales Características Sociodemográficas.***

Luego de obtener los diferentes perfiles de los individuos, en este caso los que posean planes asistenciales, será de sumo interés conocer las relaciones existentes entre el ingreso total de cada individuo con sus respectivas características sociodemográficas.

Generar modelos

La técnica que permitirá realizarlo, será la de Árboles de Decisión con el *DB2 Intelligent Miner for Data* 

Está es una técnica predictiva con supervisión, que permitirá obtener como resultado reglas que explican el comportamiento de una variable target con relación a otras predictoras.

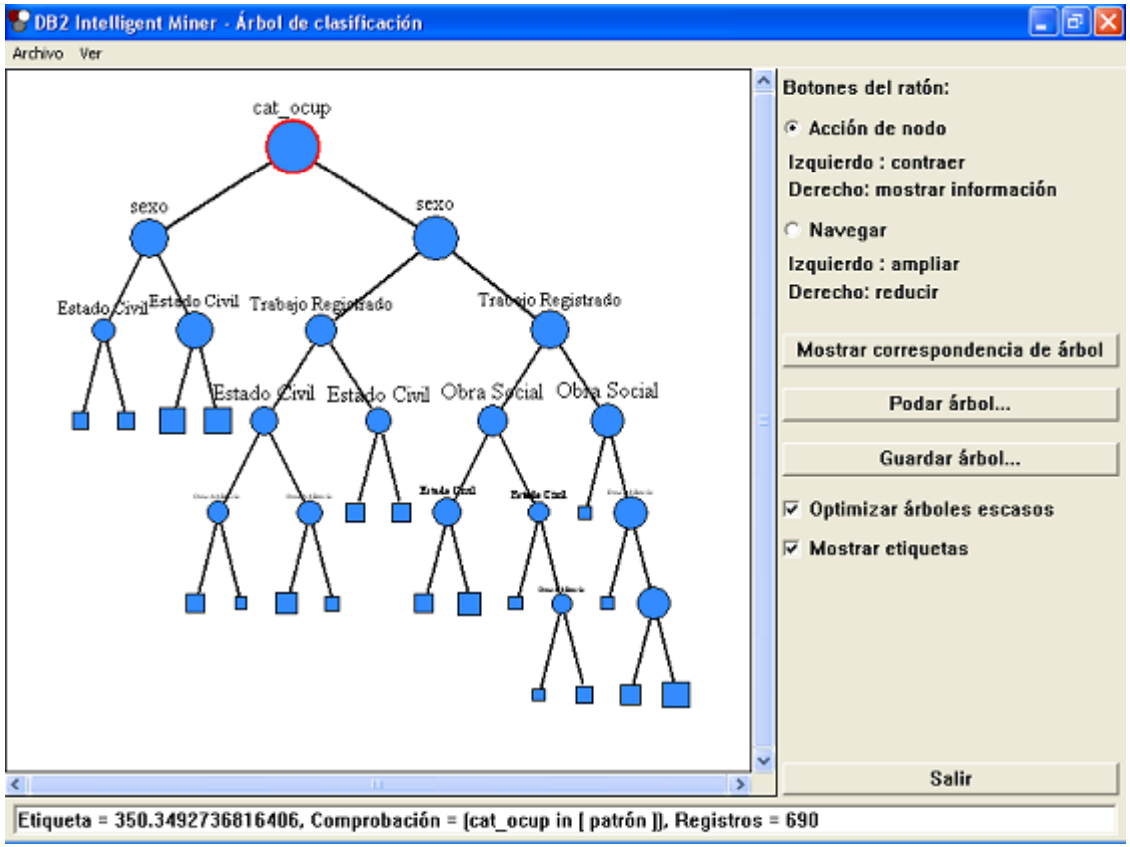
Generar modelos

El resultado obtenido es un modelo que clasifica a los individuos con sus respectivos ingresos y sus principales características sociodemográficas.

Se identifican diecinueve reglas que explican el perfil de estos individuos, determinadas por los nodos de desarrollo del Árbol (mayor cantidad de individuos y mayor pureza).

Generar modelos

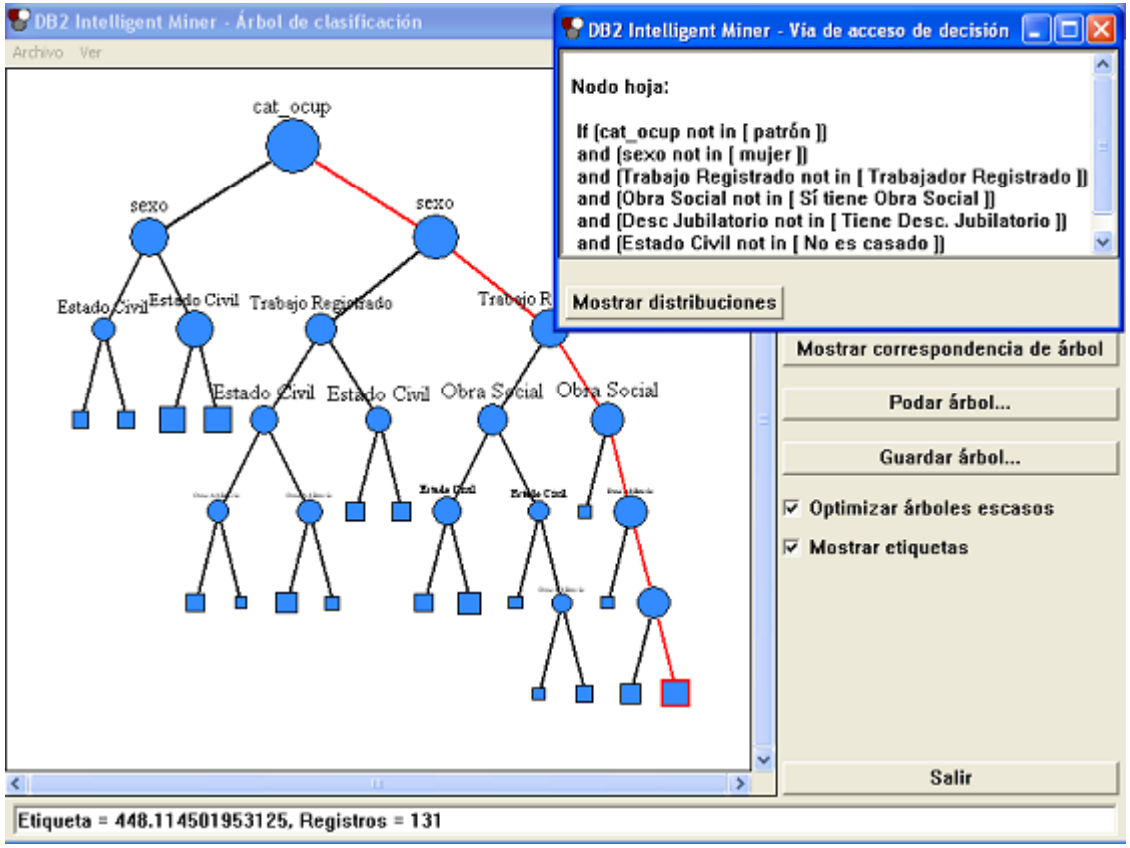
Se identifican diecinueve reglas que explican el perfil de estos individuos, determinadas por los nodos de desarrollo del Árbol (mayor cantidad de individuos y mayor pureza).



Generar modelos

Regla Nº 1

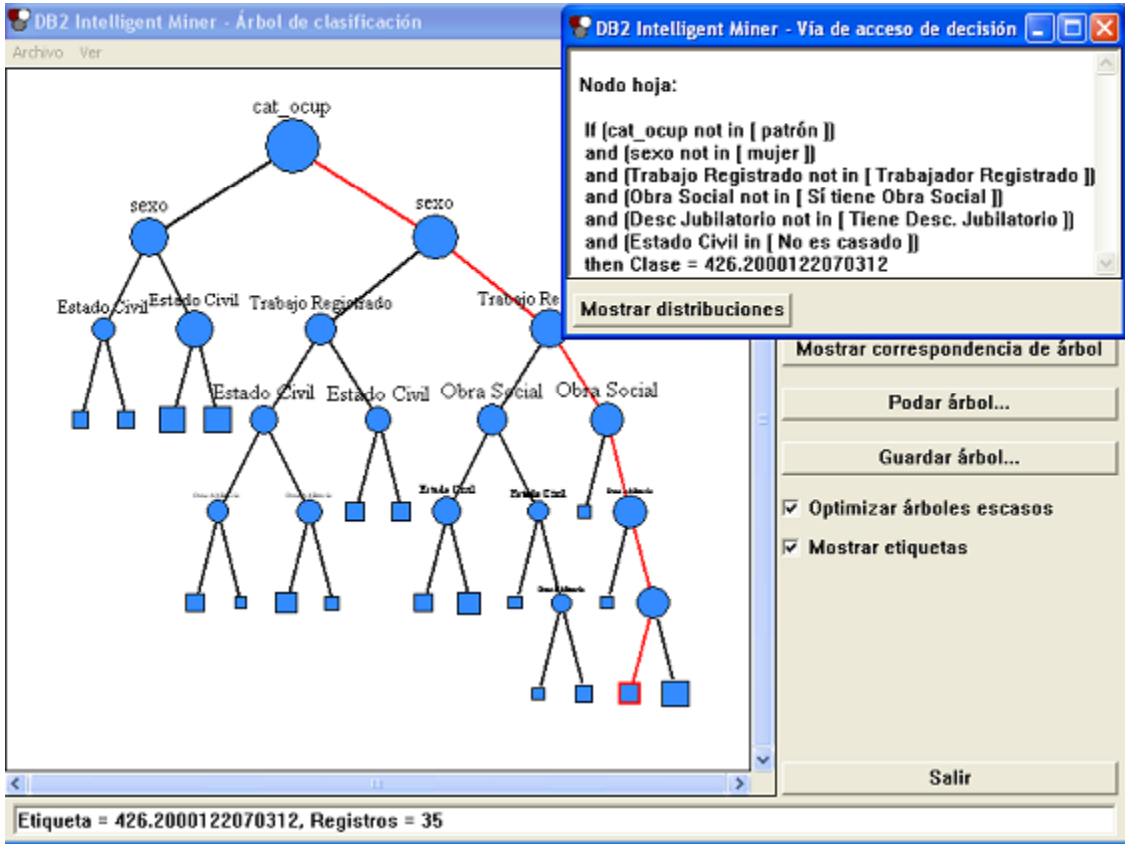
Si el individuo de estudio es de sexo femenino, no es patrón, no tiene trabajo registrado, ni obra social, ni descuento jubilatorio y su estado civil no es el casados entonces el ingreso total individual es de 448.11.



Generar modelos

Regla Nº2

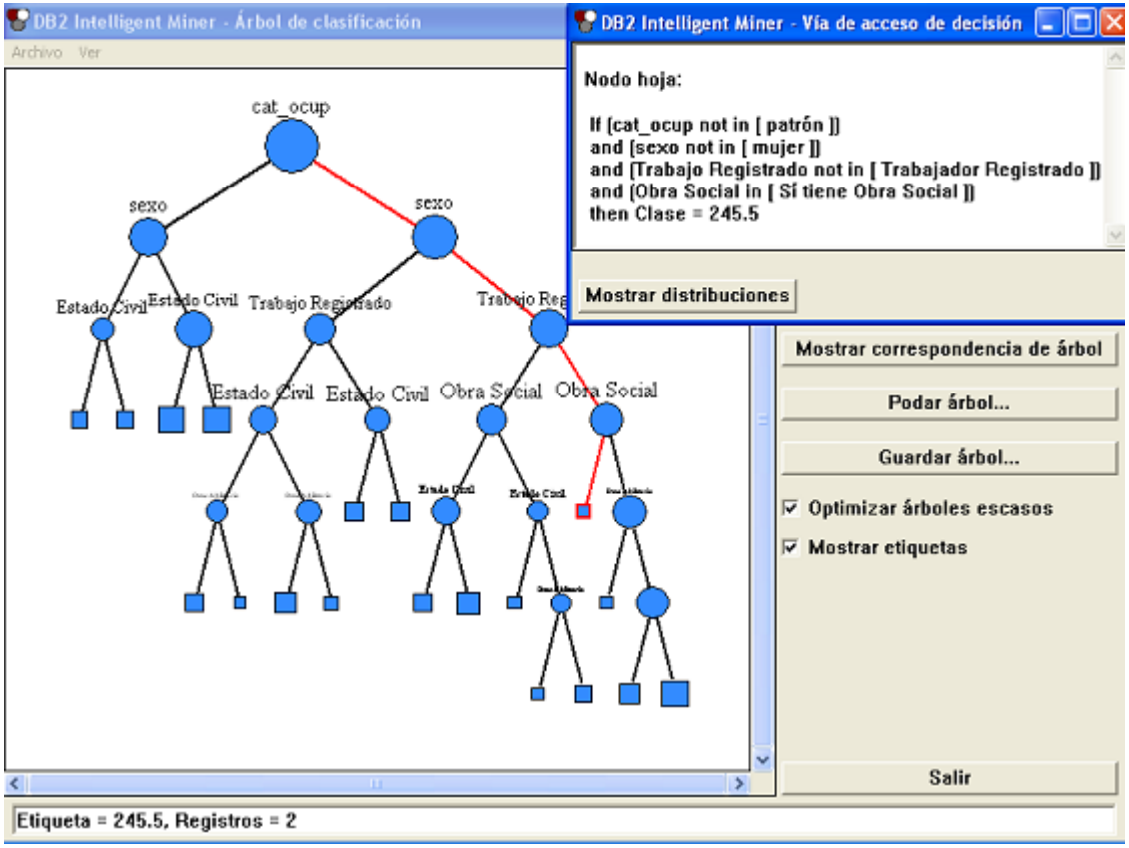
Si el individuo de estudio es de sexo femenino, no es patrón, no tiene trabajo registrado, ni obra social, ni descuento jubilatorio y su estado civil es el casados entonces el ingreso total individual es de 426.20.



Generar modelos

Regla Nº 4

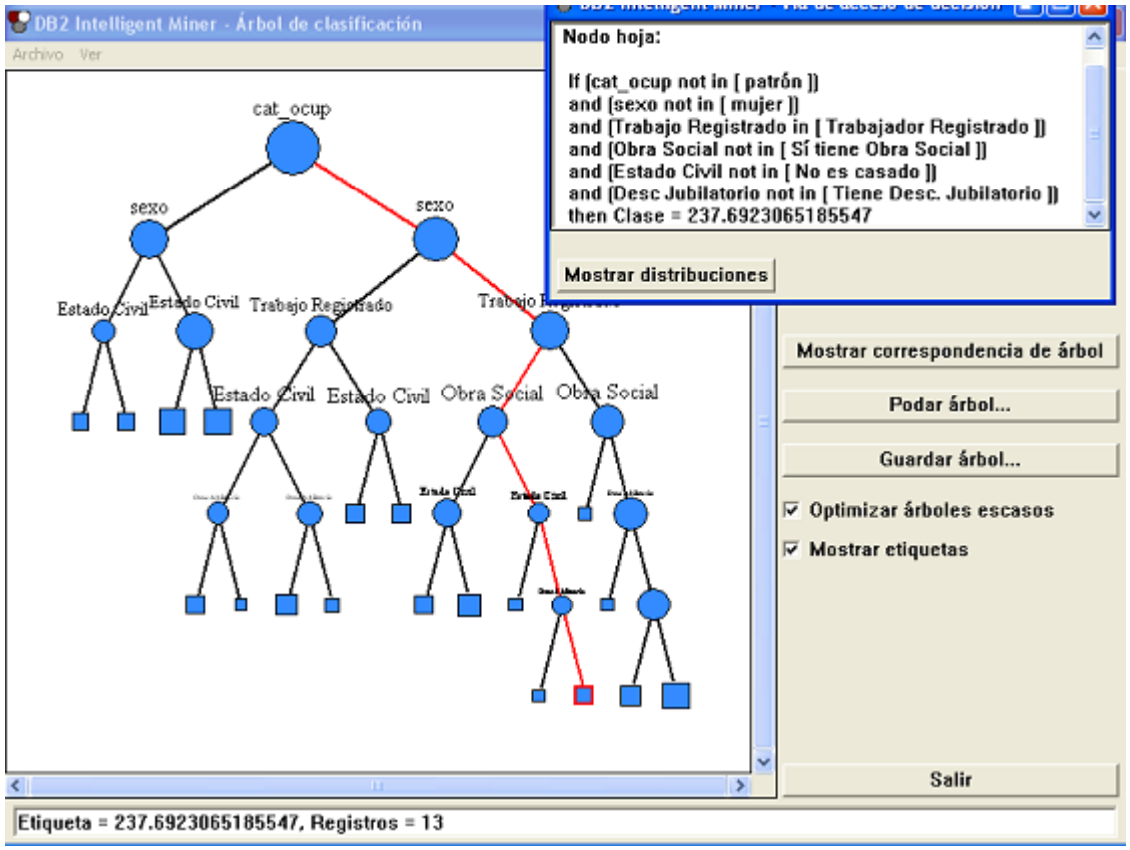
Si el individuo de estudio es de sexo femenino, no es patrón, no tiene trabajo registrado, pero sí posee obra social, entonces el ingreso total individual es de 245.5.



Generar modelos

Regla Nº5

Si el individuo de estudio es de sexo femenino, goza de un trabajo registrado, no es patrón, no posee obra social, ni descuento jubilatorio y su estado civil no es casados entonces el ingreso total individual es de 237.69.



Conclusiones

- Partiendo de los datos suministrados por el *Instituto Nacional de Estadística y Censos* (http://www.indec.mecon.ar/), se pudieron extraer patrones sociodemográficos y económicos de la una muestra de la población total de la republica Argentina, en este caso el aglomerado de Corrientes.
- Empleando técnicas de "*Clustering*" se obtuvo como resultado un modelo con todos los perfiles de los individuos que poseen planes asistenciales en la cuidad de Corrientes.
- Utilizando el algoritmo de "*Árboles de decisión y clasificación*" se obtuvo como resultado un modelo que clasifica a los individuos con sus respectivos ingresos y sus principales características sociodemográficas.

BIBLIOGRAFIA

- Jhon Wiley W.H. Inmon and Sons. Building the Data Warehouse. USA 1996.
- Fayyad, U.M. Piatetskiy-Shapiro, G. Smith, P. Ramasasmy, Advances in Knowledge Discovery and Data Mining. USA 2006.
- W. H. Inmon, Jhon Wiley and Sons. Data Warehouse Performance, USA 1992.
- IBM Press 2001. IBM DAB2 UDB Business Intelligence Tutorial.

FIN
GRACIAS POR SU ATENCIÓN

Mail : alfonsocutro@gmail.com