Universidad Nacional del Nordeste
Facultad de Ciencias Exactas, Naturales y Agrimensura

Trabajo Final de Aplicación

Minería de Datos Aplicada a la Encuesta Permanente de Hogares

Luis Alfonso Cutro - L.U.: 29027
Prof. Orientador: Mgter. David Luis la Red Martínez
Prof. Orientador: Expto. Osvaldo Pantaleón Quintana

Licenciatura en Sistemas de Información
Corrientes - Argentina

2008

Para los que me guían desde el más allá

Prefacio

Con la denominada *sociedad de la información* se está produciendo un fenómeno curioso, día a día se multiplica la cantidad de datos almacenados. Sin embargo, contrariamente a lo que pudiera esperarse, esta explosión de datos no supone un aumento de nuestro *conocimiento*, puesto que resulta imposible procesarlos con los métodos clásicos.

La información que se genera diariamente dentro de la organización es uno de sus activos principales, por lo que se debe orientar los recursos tecnológicos de manera que ayuden a los ejecutivos a tomar decisiones estratégicas y oportunas.

La capacidad de solucionar problemas de *decisión*, y la calidad de las decisiones tomadas, tienen grandes repercusiones en la organización y en su correcto funcionamiento, de modo que actualmente las organizaciones se enfrentan a la paradoja de que, cuantos más datos están disponibles, menos información se tiene.

Para enfrentar estos problemas, en los últimos años han surgido una serie de técnicas que facilitan el procesamiento avanzado de los datos y permiten realizar un análisis en profundidad de los mismos de forma automática. La idea clave es que los datos contienen más información oculta de la que se ve a simple vista.

En este trabajo se propone desarrollar un proceso de *extracción de conocimiento* a partir de los datos de la *Encuesta Permanente de Hogares (EPH)* suministrados por el *Instituto Nacional de Estadística y Censos (http: //www. indec. mecon. ar/).*

Se contempla desde la exportación de las bases de datos, la creación del Almacén de Datos (Data Warehouse), como así también el desarrollo de un sistema Web que permita la visualización de los resultados alcanzados en la etapa de extracción de conocimiento.

Objetivos Logrados

Se han alcanzado plenamente la totalidad de los objetivos planteados para el presente trabajo:

- Realización del Almacén de Datos (Data Warehouse) partiendo de la fuente antes indicada.

- Extracción de conocimiento mediante ténicas de minería de datos.
- Desarrollo de un sistema que permita visualizar los resultados alcanzados con los respectivos productos de extracción de conocimientos.

Clasificación del Trabajo

Utilización de software de base que permite la extracción de conocimiento en bases de datos mediante técnicas de minería de datos.

Desarrollo de una aplicación Web para la visualización de los resultados de las técnicas empleadas.

Etapas de Desarrollo

- Se ha efectuado una amplia recopilación bibliográfica específica de los temas pertinentes a la tarea planificada y a los productos de software que se emplearon para la concreción del Trabajo Final.
- Gracias de las gestiones realizadas por el Profesor Orientador Mgter. David Luís la Red Martínezante ante IBM Argentina se han recibido materiales tanto en CD's como en libros de dicha empresa, en el marco del Scholars Program de la misma, destinado a Universidades de todo el mundo; se destacan por ser necesarios para la realización del presente Trabajo Final los referentes a productos de software tales como los siguientes:
  - WebSphere Studio Application Developer versión 5.0 y 5.1.2.
  - DB2 UDB WorkGroup Server Edition versión 8.1.0.
  - DB2 UDB Personal Edition versión 8.1.0.
  - DB2 DB2 Intelligent Miner for Data versión 8.1.0.
- Se obtuvieron bases de datos como también documentación referente a la Encuesta Permanente de Hogares (EPH).
- Se realizaron las traducciones de los manuales correspondientes al software de extracción de conocimientos IBM DB2 Intelligent Miner for Data, como así también las correspondientes documentación sobre la base de datos IBM DB2 UDB.
- Se realizaron las traducciones de los manuales correspondientes a la suite Open Source Pentaho Business Intelligences.

- Se ha realizado el estudio del Manejador de Bases de Datos (DBMS) multiplataforma IBM DB2 UDB.

- Se ha desarrollado un Almacén de Datos (Data Warehouse) con la suite de IBM DB2 UDB WorkGroup Server Edition versión 8.1.0.

- Se ha logrado extraer patrones de conocimientos ocultos sobre los datos de la Encuesta Permanente de Hogares (EPH) gracias al IBM DB2 Intelligent Miner for Data.

- Se ha realizado un detallado estudio del entorno de trabajo Scientific WorkPlace 2.5.0 para la escritura del libro correspondiente al informe final.

- Se ha realizado un detallado estudio del lenguaje Java, para la utilización de la herramienta de desarrollo WebSphere Studio Application Developer, versión 5.0 para Windows.

- Se ha realizado el desarrollo de la aplicación utilizando páginas HTML.

- Una vez finalizada la aplicación se realizó la grabación en DVD de todo el material correspondiente al trabajo final: una versión de la aplicación, otra referente al informe final en formato Latex y el PDF generado. También se incluyó los instaladores de los productos utilizados para el desarrollo, es decir IBM DB2 UDB , DB2 Intelligent Miner for Data y WebSphere Studio Application Developer.

Organización del Trabajo Final de Aplicación

El trabajo final de aplicación comprende el informe final impreso y un DVD, además de un resumen breve y de un resumen extendido.

El *informe final* está organizado en capítulos, los que se indican a continuación:

- Introducción a la *minería de datos*: presenta una visión general sobre el descubrimiento de conocimiento en bases de datos (KDD).

- Introducción al *motor de bases de datos* DB2 UDB: detalla las más relevantes caracteríticas de esta familia de productos de gestión de bases de datos multiplataforma.

- Introduccion al *entorno de desarrollo Java* WebSphere Studio: presenta los principales aspectos de este entorno de desarrollo de aplicaciones complejas.
- Introducción al *producto de minería de datos* Intelligent Miner for Data: especifica las más relevantes caracteríticas del software y sus principales funciones: estadísticas, preproceso y minería.
- Preparación del *Data Warehouse*: describe las principales etapas para la confección de un Almacén de Datos (Data Warehouse).
- *Extracción de conocimiento* con IBM DB2 Intelligent Miner for Data: especifica los resultados extraídos mediante las técnicas de minería de datos empleadas.
- *Extracción de conocimiento* con Pentaho Business Intelligence: detalla los principales aspectos de esta suite open source business intelligence, como así también los resultados extraídos.
- *Aplicación web multiplataforma*: presenta los principales resultados extraídos con los productos especificados anteriormente.
- *Conclusiones*: presenta las conclusiones a las que se ha llegado al finalizar el presente trabajo y se plantean líneas futuras de acción.

El *DVD*, adjunto al informe final impreso, contiene lo siguiente:

- Instaladores del software utilizado.
- Resúmenes del trabajo realizado.
- Informe final en formato digital.
- Presentación para la defensa final.
- Copia de seguridad de la base de datos de la aplicación.
- Aplicación desarrollada.

Luis Alfonso Cutro
Licenciatura en Sistemas de Información
Universidad Nacional del Nordeste
L.U.: 29027
Corrientes; 23 de Septiembre de 2008

Índice General

Índice de Figuras

Capítulo 1

Introducción a la Minería de Datos

1.1 Los Datos y el Origen de la Información

El *dato* es un hecho que describe un suceso o una entidades.

La importancia de los datos está en su capacidad de asociarse dentro de un contexto para convertirse en *información*.

Por sí mismo los *datos* no tienen capacidad de comunicar un significado y por lo tanto no pueden afectar el comportamiento.

En cambio la información reduce nuestra incertidumbre (sobre algún aspecto de la realidad) y, por tanto, nos permite tomar mejores decisiones.

1.2 El Procesamiento de los Datos

Los *datos* necesitan alojarse en un lugar físico (memoria) para su posterior procesamiento o ejecución. Hasta el momento se ha supuesto que los datos no son tan voluminosos y por lo tanto caben en *memoria*.

Sin embargo, existen problemas en donde el volumen de datos es tan grande que es imposible mantenerlos en memoria. Entonces, los datos se almacenan

en un conjunto de archivos, los que forman una *base de datos*.

Día a día se multiplica la cantidad de datos almacenados, sin embargo, contrariamente a lo que pudiera esperar, esta explosión de datos no supone un aumento de nuestro conocimiento, puesto que resulta imposible procesarlos con los métodos clásicos.

Es así que hoy las organizaciones tienen gran cantidad de datos almacenados y organizados, pero a los cuales no los pueden analizar eficientemente en su totalidad.

Con algunas sentencias de *SQL* se puede realizar un primer análisis, pero la mayoría de las veces, se requiere la utilización de técnicas más avanzadas.

El *descubrimiento de conocimiento* en bases de datos apunta a procesar automáticamente grandes cantidades de datos para encontrar conocimiento útil en ellos.

1.3 Descubrimiento de Conocimiento en Bases de Datos (KDD)

El KDD (Knowledge Discovery from Databases) es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia, comprensibles a partir de los datos. [20]

El objetivo fundamental del *KDD (Knowledge Discovery from Databases)*, es encontrar conocimiento útil, válido, relevante y nuevo sobre una determinada actividad mediante algoritmos, dadas las crecientes órdenes de magnitud en los datos (ver fig. 1.1 de la pág. 3).

Al mismo tiempo hay un profundo interés por presentar los resultados de manera visual o al menos de manera que su interpretación sea muy clara.

El resultado de la exploración deberá ser interesante y su calidad no debe ser afectada por ruido en los datos.

1.4 Estructuración de los Datos

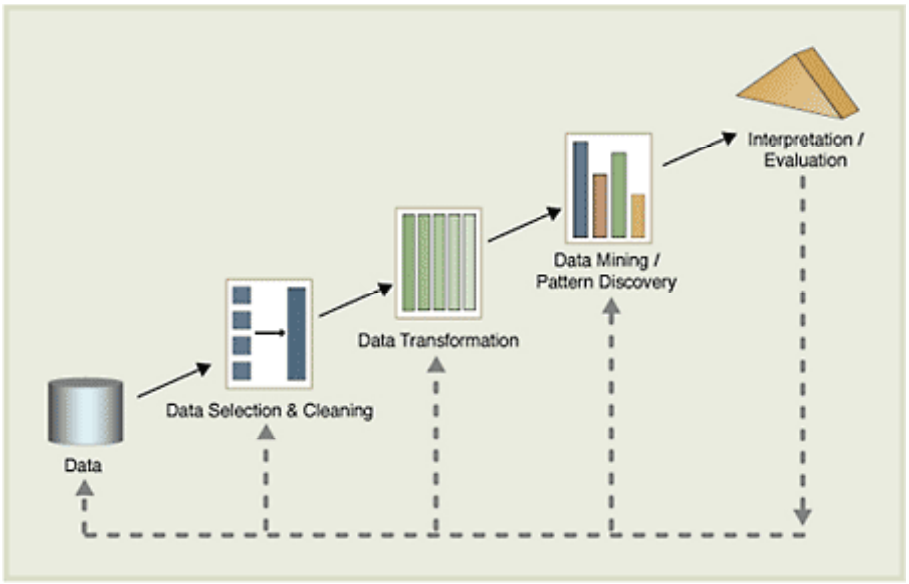

Figura 1.1: Proceso del KDD(Knowledge Discovery from Databases).

Para poder analizar los datos con fiabilidad es necesario que exista una cierta estructuración y coherencia entre los mismos [21].

Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra en *bases de datos* y otras fuentes muy diversas, y a su vez estas pueden ser tanto internas como externas.

Surge aquí la necesidad de conjugar los distintos ficheros y bases de datos de manera que se pueda utilizarlos para extraer conclusiones.

Solucionados los inconvenientes de heterogeneidad de las fuentes, surgen otros problemas relacionados a la estandarizacion de los datos:

- Diferentes tipos de datos representando el mismo concepto (ejemplo: la representación de fecha, donde al año se lo puede guardar con dos o cuatro dígitos).
- Diferentes claves para representar el mismo elemento (ejemplo: un mismo cliente puede ser representado por un código de cliente o por un NIF).

- Diferentes niveles de precisión al representar un dato (ejemplo: los números reales no siempre se almacenan de la misma forma, y es posible que generen algún problema.

Como se ve, la estructuracion de los datos no es sencilla y esto se agrava cuando los diferentes ficheros se encuentran en sistemas informáticos y soportes diferentes.

Por ello la calidad de los resultados está directamente relacionada con la correcta comprensión y posterior estructuración de los datos almacenados.

Lo razonable sería recoger los datos (información histórica) en un sistema separado y específico. Nace el *Data-Warehousing*: *Almacenes o Bodegas de Datos*, con la necesidad de unificar los distintos ficheros y bases de datos para poder comprenderlos. Por ello, se necesita de tecnologías que sirvan de guía para comprender el contenido de las Bases de Datos.

1.5 Data Warehouse (DW), Bodegón de Datos o Almacén de Datos

Básicamente se la puede describir como una *combinación hardware, software especializado y datos provenientes de distintas fuentes que sirve a la administración para la toma de decisiones* [22].

Es un sistemas de información orientado a la toma de decisiones empresariales que almacenando de manera integrada la información relevante del negocio, permite la realización de consultas complejas con tiempos de respuesta cortos.

El *Data Warehouse* es un almacén estructurado de la información clave de nuestro negocio, que integra datos provenientes de todos los departamentos, sistemas, etc., y que nos permite analizar el funcionamiento de nuestra compañía y tomar de decisiones sobre su gestión.

Es un almacén destinado específicamente para mantener datos organizados.

1.5.1 Características del DW

Un Data Warehouse es una colección de datos orientados a temas integrados, no volátiles y variantes en el tiempo, organizados para soportar necesidades empresariales [21]

Por ello es que un *Data Warehouse* se caracteriza por ser *Integrado, Temático, Histórico y No volátil.*

Integrado, es decir que al fluir del entorno operacional al entorno de almacén de datos, los datos asumen una codificación consistente.

Temático, debido a que almacena información resumida que se estructura en función de temas empresariales u organizacionales.

Histórico, dado que contiene suficiente espacio para almacenar datos que posean una antigüedad de diez años o mayor aun .

No volátil, es decir los datos no se modifican o cambian bajo ningún concepto una vez introducidos en el almacén de datos, únicamente puede ser cargados o leídos.

1.5.2 Beneficios del DW

Las claves que provee el Data Warehouse son, por un lado la creación de una arquitectura de datos única para todas las aplicaciones, como se vemos en la fig. 1.2 de la pág. 6 y también la resolución de problemas de integridad y calidad de datos.

Permitiendo así a los Administradores de Bases de Datos que redacten informes o analicen estas grandes cantidades de información, para así poder tomar decisiones según los resultados del análisis [23].

1.5.3 Construcción del DW

Un *Data Warehouse* se genera a partir de otras bases de datos, su construcción y desarrollo requiere integrar varios componentes de tecnología y la habilidad para hacerlos funcionar todos juntos [8].

El objetivo fundamental es transformar datos en conocimiento.

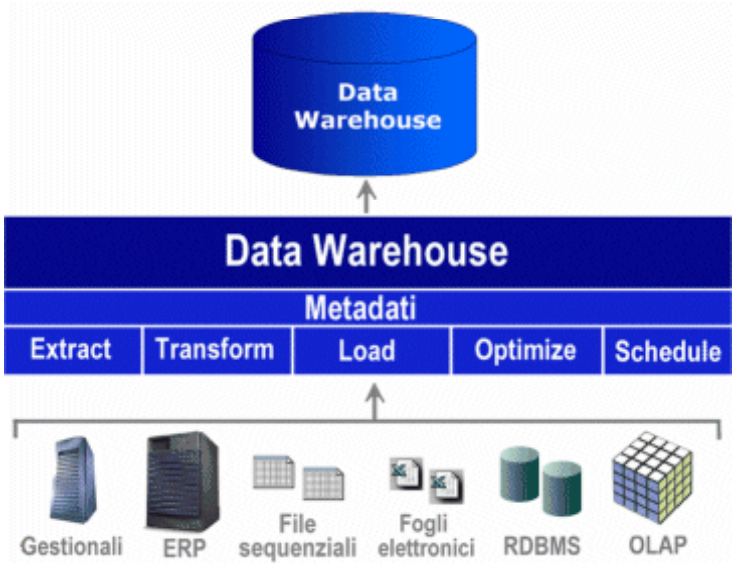

Figura 1.2: Principales Aplicaciones del Data Warehouse.

Para ello es necesario ensamblar datos existentes siguiendo instrucciones precisas para obtener un óptimo resultado.

Para su *construcción* se debe considerar en primer lugar el hardware necesario, dado que a mayor tamaño del almacén, mayor deberá ser la capacidad de almacenamiento y el procesamiento. Luego el software y los datos que se utilizarán.

Las Principales etapas de construcción del *Almacén de Datos* son:

- *Extracción*: Se crea los archivos de la Base de Datos para transacciones y se guardan en el servidor que mantendrá el Almacén de Datos (se extrae la información operacional).
- *Depuración*: Se unifica la información de los datos de manera que se pueda insertar en el Almacén de Datos (se transforma la información a un formatos consistentes).
- *Carga*: Se transfiere los archivos depurados a la base de datos que servirá como almacén de datos.
- *Comparación*: Se comparan los datos del almacén con los originales.

De todas maneras, el éxito de *Data Warehouse* no está en su construcción, sino en saber utilizarlo para mejorar procesos empresariales, operaciones y decisiones.

1.5.4 Información Oculta en los DW

Si se almacena la información mas relevante de nuestro negocio en un sistema que acumula y acumula datos sin parar, un análisis razonable nos puede permitir descubrir tendencias, localizar grupos de datos con comportamiento homogéneo, establecer relaciones, etc [2].

Esta información está oculta en los datos y será necesario utilizar todas las técnicas a nuestro alcance para obtenerla. El objetivo que nos planteamos es localizar relaciones entre atributos de nuestro *Data Warehouse.*

1.5.5 DW Como Soporte de Decisión Para los Negocios

Los negocios necesitan aprovechar las posibilidades que les ofrece la actual tecnología para permanecer competitivos y rentables.

El conocimiento del mercado y de los clientes se ha convertido en un factor de supervivencia para las empresas, y el *Data Warehouse* se perfila como la tecnología para lograr manejarlo.

Las organizaciones necesitan información renovada acerca de las tendencias presentes para mantener su competitividad. Precisan saber qué es lo que está pasando por las mentes de sus clientes.

Asimismo, necesitan determinar los requerimientos corporativas y traducirlos en consultas que puedan ser respondidas a través del *Data Warehouse.*

Para ello, el *Data Warehouse* conserva información histórica y actual sobre un negocio, y permite recuperar datos que, bajo la forma de informes, facilitan el descubrimiento y las comprensión de patrones de comportamiento y tendencias de las cuales resultan conclusiones o recomendaciones para los futuros cursos de acción.

Sintetiza algunos datos muy importantes, otorgando al usuario nuevo conocimiento comercial.

1.6 Inteligencia de Negocios

Hace referencia a un conjunto de productos y servicios para acceder a los datos, analizarlos y convertirlos en información.

Es un paraguas bajo el que se incluye un conjunto de conceptos y metodologías cuya misión consiste en mejorar el proceso de toma de decisiones en los negocios basándose en hechos y sistemas que trabajan con hechos. [Howard Dresner ,Gartner Group, 1989].

La *Inteligencia de Negocios* es una manera de manejar la información histórica de una empresa a través de la construcción de un *Data Warehouse,* y explotarla con fines de análisis para una mejor toma de decisiones [14].

A través de la creación de modelos de información multidimensionales una organización puede beneficiarse al conocer de manera óptima cómo su negocio se ha comportado a lo largo del tiempo, cómo se comporta en el presente y cómo se estima se comportará en el futuro [12].

Algunos de los *beneficios* que obtienen las organizaciones al implementar este sistemas son:

- Capacidad de análisis.
- Reducción de costos.
- Reducción de tiempos de proceso.
- Búsqueda de patrones desconocidos que sólo aparecen al momento en que los datos son analizados.
- Generación de pronósticos, presupuestación y planeación.

La inteligencia en el negocio electrónico, incluye actividades como el procesamiento analítico en línea (*OLAP*) y aprovechamiento de datos, también llamada extracción de datos o *Minería de Datos* (ver fig. 1.3 de la pág. 9).

1.7 Minería de Datos

La *Minería de Datos* es la etapa de descubrimiento en el proceso de *KDD (Knowledge Discovery from Databases)*: *"paso consistente en el uso de algorit-*

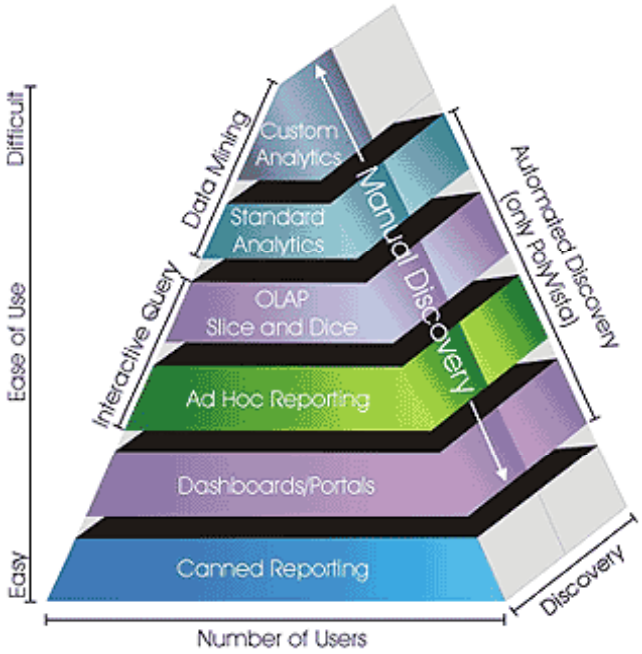

Figura 1.3: Inteligencia de Negocios BI.

mos concretos que generan una enumeración de patrones a partir de los datos preprocesados" [7].

Para conseguirlo hace uso de diferentes tecnologías que resuelven problemas típico de agrupamiento automático, clasificación y asociación de atributos, etc.

La *Minería de Datos* es, en principio, una fase dentro de un proceso global denominado *Descubrimiento de Conocimiento en Bases de Datos,* aunque finalmente haya adquirido el significado de todo el proceso en lugar de la fase de extracción de conocimiento [9].

Es un mecanismo de explotación, consistente en la búsqueda de información valiosa en grandes volúmenes de datos.Está muy ligada a las *Bodegas de Datos* ya que las mismas proporcionan la información histórica con la cual los algoritmos de minería tiene la información necesaria para la toma de decisiones [8].

El *Data Mining (DM)* es un conjunto de técnica de análisis de datos que permiten:

- Extraer **Patrones, Tendencias y Regularidades** para describir y comprender mejor los datos.
- Extraer **Patrones y Tendencias** para predecir comportamientos futuros.

Debido al gran volumen de datos este análisis:

- Ya no puede ser manual (ni incluso facilitado por herramientas de *Almacén de Datos*).
- Ha de ser (semi-) automático.

En los sistemas estándar de gestión de bases de datos las consultas se resuelven accediendo a distintos conjuntos de datos almacenados.

Los sistemas de *Data Mining (DM)* infieren conocimiento de las *bases de datos* en forma de **Estructuras y Patrones**. Este conocimiento supone un nuevo conjunto de información en base a la cual se responden las consultas.

1.7.1 Evolución e Historia de la Minería de Datos

La idea de *Minería de Datos* no es nueva. Ya desde los años sesenta los estadísticos manejaban términos como *Data Fishing*, *Data Mining (DM)* o *Data Archaeology* con la idea de encontrar correlaciones sin una hipótesis previa en *bases de datos* con ruido.

A principios de los años ochenta, *Rakesh Agrawal, GioWiederhold, Robert Blum y Gregory Piatetsky-Shapiro* entre otros, empezaron a consolidar los términos de *Minería de Datos* y *KDD*.

Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios.

La evolución de sus herramientas en el transcurso del tiempo puede dividirse en cuatro etapas principales:

- *Colección de Datos (1960).*
- *Acceso de Datos (1980).*
- *Almacén de Datos y Apoyo a las Decisiones (principios de la década de 1990).*
- *Minería de Datos Inteligente.(finales de la década de 1990).*

1.7.2 Aplicación de la Minería de Datos

En Internet

- *E-bussines*: Perfiles de clientes, publicidad dirigida, fraude.
- *Buscadores Inteligentes*: Generación de jerarquías, bases de conocimiento web.
- *Gestión del Tráfico de la Red*: Control de eficiencia y errores.

El Mundo de los Negocios

- *Banca:* Grupos de clientes, préstamos, oferta de productos.

- *Compañías de Seguros*: Detección de fraude, administración de recursos.
- *Marketing*: Publicidad dirigida, estudios de competencia.

En Mundo de la Ciencias

- *Meteorología*:Teleconexiones (asociaciones espaciales), predicción.
- *Física*: Altas energías, datos de colisiones de partículas (búsqueda de patrones).
- *Bio-Informática:* Búsqueda de patrones en ADN, proyectos científicos como genoma humano, datos geofísicos, altas energías, etc.

1.7.3 Ejemplos de las Aplicaciones de la Minería de Datos

En el Área de la Meteorología

Teleconexiones: Son predicción de asociaciones espaciales sobre una determinada Área Geográfica (ver fig. 1.5 de la pág. 13).

Existen bases de datos con simulaciones de los campos atmosféricos en rejillas dadas (ver fig. 1.4 de la pág. 13).

Se dispone de gran cantidad de información en observatorios locales: precipitaciónes,

temperaturas, vientos, etc. (ver fig. 1.6 de la pág. 14).

En el Ámbito de la Web

- *Reglas de Asociación*:

El 60% de las personas que esquían viajan frecuentemente a Europa.

- *Clasificación*:

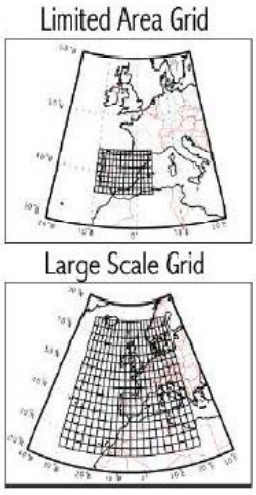

Figura 1.4: Areas de los Campos Atmosféricos.

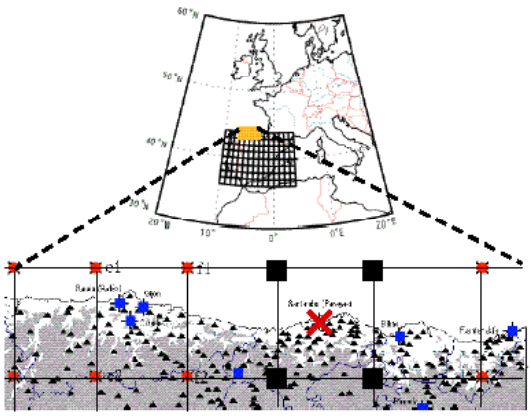

Figura 1.5: Analisis sobre una determinada Área Geográfica.

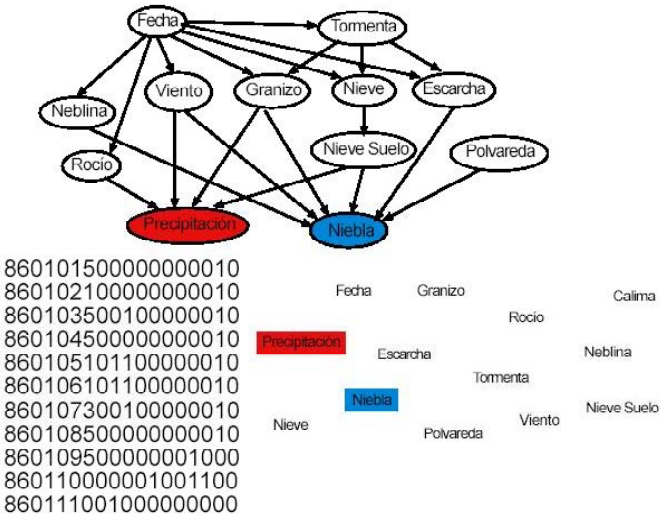

Figura 1.6: Información obtenida en los observatorios.

Personas menores de 40 años y salario superior a $2000 compran on-line frecuentemente.

- *Clustering*:

Los usuarios A y B tienen gustos parecidos (acceden URLs similares).

- *Detección de "Outliers":*

El usuario A navega en Internet más del doble del tiempo promedio. [9]

1.8 Sistemas OLAP (On-Line Analytical Processing)

El procesamiento analítico en línea se define como el análisis rápido de información compartida [1].

Aparece en contraposición al concepto tradicional *OLTP (On-Line Transactional Processing)*, que designa el procesamiento operacional de los datos,

orientado a conseguir la máxima eficacia y rapidez en las transacciones individuales de los datos.

Es una aplicación de bases de datos orientada a array que permite visualizar, manipular y analizar *bases de datos* multidimensionales.

Permite a los usuarios analizar datos corporativos críticos para descubrir los factores decisivos que influyen en el negocio. Realizan todas las tareas analíticas y de reporte incluyendo informes de medidas de rendimiento del negocios que resaltan indicadores de rendimiento clave.

1.8.1 Las Herramientas del OLAP

Están basadas generalmente en sistemas o interfaces Multidimensionales.

Utilizan operadores específicos (además de los clásicos):

- Drill.
- Roll.
- Pivot.
- Slice.
- Dice.

El resultado se presenta de una manera Matricial o Hibrida.

1.8.2 Principales Beneficios del OLAP

Permite a los usuarios de entender no solo lo que está pasando, sino cuándo, por qué y cómo.

Resuelve todas las necesidades de análisis con una herramienta de velocidad electrónica.

Proporciona capacidades de análisis para todos los tipos de usuario así como para clientes y proveedores.

1.8.3 Diferencias Entre OLAP y Minería de Datos

Las Herramientas del OLAP

Proporcionan facilidades para "Manejar" y "Transformar" los *datos*.

Producen otros *datos* (más agregados, combinados).

Ayudan a analizar los *datos* por que producen diferentes vistas de los mismos.

Las Herramientas de Minería de Datos

Son muy variadas permites extraer *Patrones, Modelos, Relaciones, Regularidades, Tendencias, etc.*

Producen *Reglas* o *Patrones* (*Conocimiento*).

Capítulo 2

Introducción al DB2

2.1 Introducción a las Bases de Datos

Antes de las *Bases de Datos* se utilizaban archivos secuenciales para almacenar datos. Estos daban un acceso muy rápido pero sólo de forma secuencial en donde para acceder a una posición, se debía recorrer el archivo entero. Más tarde aparecieron los archivos indexados, donde el acceso ya podía ser aleatorio (acceder de una vez a la posición deseada del mismo).

El sistema de archivos era el sistema más común de almacenamiento de *datos*. Para compartirlos entre varias máquinas surgió el *NFS2 (Network file system)*, y más tarde para evitar fallos en los sistemas de archivo aparecieron los sistemas *RAID3 (Redundant Array Of Independent / Inexpensive Disks)*.

Pero los programas y *datos* cada vez eran más complejos y grandes. Se requería de un almacenamiento que garantizara un cierto número de condiciones y que permitiera operaciones complejas sin que se violaran estas restricciones. Además cada usuario que accediera a los *datos* debía tener su trabajo protegido de las operaciones que hicieran el resto de usuarios.

Respondiendo a estas necesidades, surgieron las *Bases de Datos Jerárquicas,* donde los *datos* se situaban siguiendo una jerarquía.

Las *Bases de Datos Jerárquicas* tenían el problema que los accesos a los datos eran *unidireccionales*, y era más complicado hacer el camino inverso (pero posible, aunque el tiempo de cálculo era mayor).

Por ejemplo: Era fácil saber qué cuentas tenía un cliente, pero no tanto saber de qué cliente era una cierta cuenta.

Para dar absoluta libertad a las relaciones entre tablas surgieron las *Bases de Datos Relacionales (Relational Data Base Management System):*

2.2 Definición de Bases de Datos

Se define una Base de Datos como una serie de datos organizados y relacionados entre sí, y un conjunto de programas que permitan a los usuarios acceder y modificar esos datos. [18].

De forma sencilla podemos indicar que una *Base de Datos* no es más que un conjunto de *información relacionada* que se encuentra *agrupada o estructurada.*

Es un conjunto *exhaustivo, no redundante de datos estructurados, organizados* independientemente de su utilización y su implementación en máquina, accesibles en tiempo real y compatibles con usuarios concurrentes con necesidad de información diferente y no predecible en tiempo; donde la información se encuentra almacenada en una memoria auxiliar que permite el acceso directo a un conjunto de programas que manipulan esos datos [19].

Una *Base de Datos* es un conjunto de *datos* de operación almacenados y utilizados por los sistemas de aplicación de una empresa, y al mencionar empresa, se lo hace en sentido genérico y amplio, pero lo importante es que necesita de datos de operación referente a su funcionamiento.

Por ejemplo: Un Banco requiere *datos* de sus Clientes, una Mutual de sus Afiliados, un Hospital de sus Pacientes, una Facultad de sus Alumnos y Profesores.

La idea general es que estamos tratando con una colección de *datos* que cumplen las siguientes propiedades:

- Están estructurados independientemente de las aplicaciones y del soporte de almacenamiento que los contiene.
- Presentan la menor redundancia posible.

- Son compartidos por varios usuarios y/o aplicaciones.

2.3 Principales Diferencias con los Archivos Convencionales

El *Archivo* por sí mismo no constituye una *Base de Datos*, sino más bien la forma en que está organizada la *información* es la que da origen a la *Base de Datos*.

Las *Bases de Datos* manuales, pueden ser difíciles de gestionar y modificar. **Por ejemplo**: En una guía de teléfonos no es posible encontrar el número de un individuo si no sabemos su apellido, aunque conozcamos su domicilio.

Del mismo modo, en un *Archivo* de pacientes en el que la información esté desordenada por el nombre de los mismos, será una tarea bastante engorrosa encontrar todos los pacientes que viven en una zona determinada.

No podemos comparar directamente *Base de Datos* con *Archivos*, porque para ello es necesario tener más de un *Archivo*, pero si esto es así entraríamos en los problemas de: *redundancia de datos, inconsistencia de datos, heterogeneidad de formatos de datos*, no podemos compartir datos de las distintas aplicaciones, no manejamos la *seguridad* de todos los *Archivos* y por último ante pequeñas modificaciones en la estructura de los datos se requiere de muchas horas de programación para adecuar las mismas.

Los problemas expuestos anteriormente se pueden resolver creando un *Sistemas de Gestión de Bases de Datos (SGBD), DBMS (Data Base Management System). .*

2.4 Orígenes y Antecedentes de las Bases de Datos

El término *Base de Datos* fue acuñado por primera vez en 1963, en un simposio celebrado en California.

En la década del 70

Edgar Frank Codd definió el *modelo relacional* y publicó una serie de reglas para la evaluación de administradores de sistemas de datos relacionales y así

nacieron las bases de datos relacionales.

A partir de los aportes de *Codd* el multimillonario *Larry Ellison* desarrolló la base de datos *Oracle*, la cual es un sistema de administración de *Base de Datos*, que se destaca por sus *transacciones, estabilidad, escalabilidad y multiplataforma.*

Inicialmente no se usó el *Modelo Relacional* debido a que tenía inconvenientes por el rendimiento, ya que no podían ser competitivas con las bases de datos *Jerárquicas* y de *Red*. Ésta tendencia cambio por un proyecto de IBM el cual desarrolló técnicas para la construcción de un sistema de bases de datos relacionales eficientes, llamado *System R.*

En la década del 80

Las *Bases de Datos Relacionales* con su sistema de *Tablas, Filas y Columnas*, pudieron competir con las *Bases de Datos Jerárquicas* y de *Red*, ya que su nivel de programación era bajo y su uso muy sencillo.

En esta década el *Modelo Relacional* ha conseguido posicionarse en el mercado de las *Bases de Datos*. Y también en este tiempo se iniciaron grandes investigaciones, como las *Sistemas de Gestión de Bases de Datos Orientadas a Objetos SGBDOO (System Management Object Oriented Databases).* .

Principios década de los 90

Para la toma de decisiones se crea el *lenguaje SQL (Structured Query Language)* , que es un lenguaje programado para consultas. El programa de alto nivel *SQL* es un *lenguaje de consulta* estructurado que analiza grandes cantidades de información, el cual permite especificar diversos tipos de operaciones frente a la misma información, a diferencia de las *bases de datos* de los 80 que eran diseñadas para las aplicaciones de procesamiento de transacciones. Los grandes distribuidores de bases de datos incursionaron con la venta de bases de datos orientadas a objetos.

Finales de la década de los 90

El boom de esta década fue la aparición de la *WWW "Word Wide Web"* ya que por este medio se facilitaba la consulta de las bases de datos. Actualmente tienen una amplia capacidad de almacenamiento de información, también una de las ventajas es el servicio de siete días a la semana las veinticuatro horas del día, sin interrupciones a menos que haya planificaciones de mantenimiento de las plataformas o el software.

2.5 Modelo de Base de Datos

Además de la clasificación por la función de las *Bases de Datos*, éstas también se pueden clasificar de acuerdo a su *Modelo de Administración de Datos*.

Un *Modelo de Datos* es básicamente una "descripción" de algo conocido como contenedor de *datos* (algo en donde se guarda la información), así como de los métodos para almacenar y recuperar información de esos contenedores. Los *Modelos de Datos* no son cosas físicas; son abstracciones que permiten la implementación de un sistema eficiente de *Bases de Datos*, por lo general se refieren a algoritmos, y conceptos matemáticos.

2.6 Organización de Sistema de Gestión de Bases de Datos (SGBD)

Los *Modelos* más comunes de organización de *Bases de Datos* son:

- *Jerárquico.*
- *En Red.*
- *Relacional.*
- *Orientado a Objetos.*

2.6.1 Bases de Datos Jerárquicas

Estructura los campos en nodos en una estructura jerárquica. Los nodos son puntos conectados entre sí formando una especie de árbol invertido. Cada entrada tiene un nodo padre, que puede tener varios nodos hijos; esto suele denominarse relación uno a muchos. Los nodos inferiores se subordinan a los que se hallan a su nivel inmediato superior.

Un nodo que no tiene padre es llamado raíz, en tanto que los que no tienen hijos son conocidos como hojas. Cuando se desea hallar un campo en particular, se empieza por el tope, con un nodo padre, descendiendo por el árbol en dirección a un nodo hijo.

Por Ejemplo: Un Sistema de Reservaciones de una Línea Aérea (ver fig. 2.1 de la pág. 22).

El *Nodo Padre* es la **Ciudad de Salida** en este caso es (Caracas), *Nodos Hijos* representando las **Ciudades Destino** que tiene a su vez *Nodos Hijos*, que son el **Número de Vuelo**. El Número de Vuelo tendrá también *Nodos Hijos*, que son los **Pasajeros**.

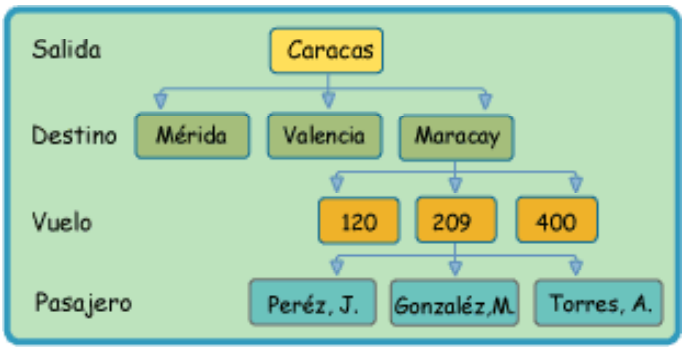

Figura 2.1: Modelo de Bases de Datos Jerárquica

Limitaciones de las Base de Datos Jerárquicas

- Al borrar un nodo padre, desaparecen también sus nodos subordinados.
- Sólo podrá añadirse un nodo hijo, si existe el nodo padre.
- Pero lo más significativo es la rigidez de su estructura: sólo un padre por hijo y ausencia de relaciones entre los nodos hijos.

2.6.2 Bases de Datos en Red

Se trata también de una organización jerárquica de nodos, pero un nodo hijo puede tener más de un solo nodo padre (relación muchos a muchos). Existen los punteros, que son conexiones adicionales entre nodos padres y nodos hijos, que permiten acceder a un nodo por vías distintas accediendo al mismo en dirección descendente por las diversas ramas.

Representa una mejora al modelo jerárquico.

Por ejemplo: Los vendedores destacados para distribuir determinados productos en algunas ciudades pueden ilustrar este modelo (ver fig. 2.2 de la pág. 23).

Cada *Producto* puede ser distribuido por más de un *Vendedor*, así mismo cada *Vendedor* puede encargarse de diferentes *Ciudades*.

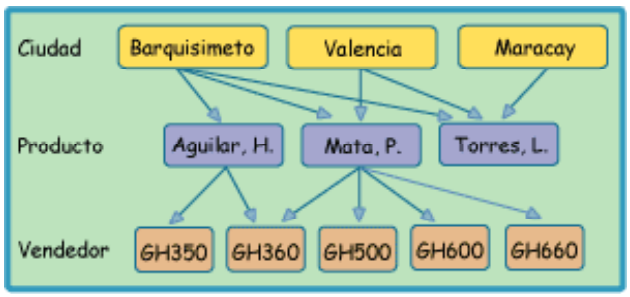

Figura 2.2: Modelo de Bases de Datos en Red

2.6.3 Bases de Datos Relacionales

Esta organización ofrece la mayor flexibilidad ya que los datos se almacenan en *Tablas* diferentes, conformadas así mismo por *Filas y Columnas*. Una tabla se denomina relación. En una *Tabla* las *Filas* contienen los *Registros*. Las *Columnas* representan los *Campos*. Las *Tablas* relacionadas poseen un campo común, el *Campo Clave*, mediante el cual la información almacenada en una tabla puede enlazarse con la información almacenada en otra.

El acceso a los datos se realiza mediante consultas escritas en *SQL (Structured Query Language)*. La *Organización de Bases de Datos Relacional* es la más difundida en la actualidad debido a su sencillez para realizar operaciones de adición, eliminación y modificación en contraste con la mayor rigidez de las *Organizaciones Jerárquicas y de Red.*

Por ejemplo: En un pequeño negocio, se puede contar con una *Tabla de Clientes* y *Tabla de Pedidos* (ver fig. 2.3 de la pág. 24).

Las órdenes que pertenecen a un determinado cliente son identificadas colocando el campo de identificación del cliente en la orden (campo clave de la tabla de clientes), lo cual permite enlazar las dos tablas.

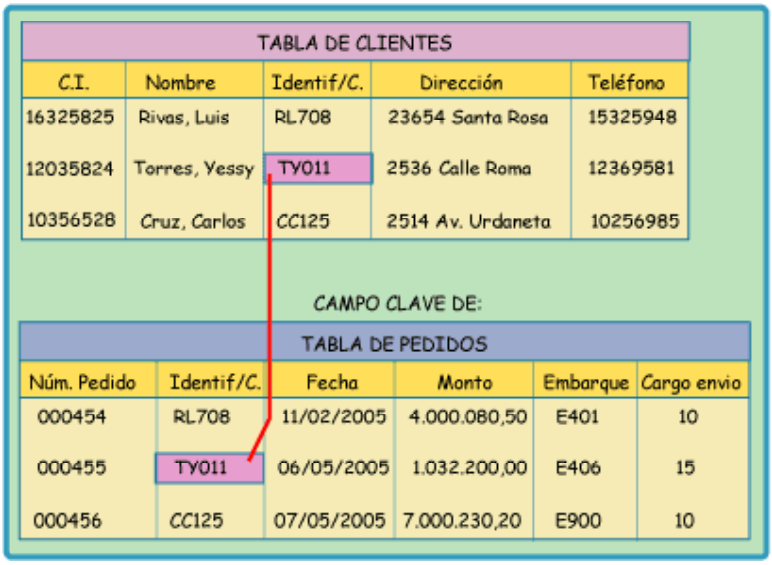

Figura 2.3: Modelo de Bases de Datos Relacional

Limitaciones de las Base de Datos Relacionales

- Estructuras muy simples *(1FN)*.
- Poca riqueza semántica.
- No soporta tipos definidos por el ususarios (solo *Dominios*).
- No soporta *Recursividad*.
- Falta de *Procesamiento/Disparadores*.
- No admite *Herencia*.

2.6.4 Bases de Datos Orientadas a Objetos (BDOO)

Que es la O.O. (Object-Oriented)

El *Análisis Orientado a Objetos (O.O.)* no es un enfoque que modela la realidad. En lugar de esto, modela la forma en que las personas comprenden la realidad.

Un *Objeto* es una representación detallada, concreta y particular de un "algo". Tal representación determina su *Identidad, su Estado y su Comportamiento* particular en un momento dado.

- *Identidad:* Le permite a un *Objeto* ser distinguido de entre otros y esto se da gracias al nombre que cada *Objeto* posee.
- *Estado:* El estado de un *Objeto* es el conjunto de valores concretos que lo caracterizan en un momento dado, como peso, color, precio, etc.
- *Comportamiento:* Define un conjunto de funciones que el objeto es capaz de llevar a cabo. Tales funciones pueden estar relacionadas entre sí, modificar el estado del objeto o invocar funcionalidades de otros objetos, entre muchas otras cosas más.

Una *Clase* se define como la generalización de un objeto en particular. Es decir, una *Clase* representa a una familia de *Objetos* concretos.

De lo anterior, podemos decir que una instancia de una clase es siempre un objeto en particular.

Qué es una Bases de Datos Orientadas a Objetos (B.D.O.O.)

Es una estructura relativamente nueva que ha suscitado gran interés.

El *Modelo de Datos Orientado a Objetos*, es una adaptación para los sistemas de *Bases de Datos del Paradigma de la Programación Orientada a Objetos*. Se basa en el concepto de *Encapsular* elementos de datos, sus características, atributos y el código que opera sobre ellos en elementos complejos llamados *Objetos*.

Los *Objetos* estructurados se agrupan en *Clases*.

Por ejemplo: El conjunto de las clases se estructura en subclases y superclases como se puede ver en la fig. 2.4 de la pág. 26) [5].

Ventajas en BDOOs

- Se destaca su flexibilidad y soporte para el manejo de tipos de datos complejos.

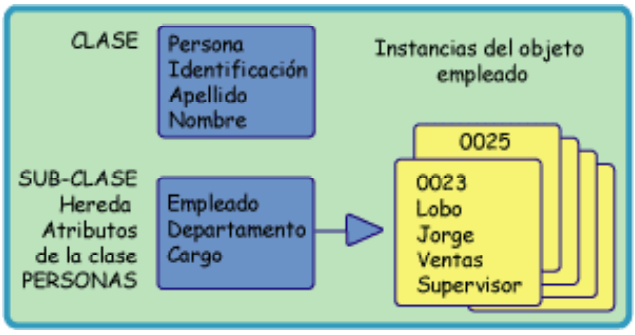

Figura 2.4: Modelo de Bases de Datos Orientada a Objetos

- Manipula datos complejos en forma rápida y ágilmente. La estructura de la Base de Datos está dada por referencias (o apuntadores lógicos) entre Objetos [6].

Posibles Desventajas de la BDOOs

- La inmadurez del mercado de BDOO constituye una posible fuente de problemas por lo que debe analizarse con detalle la presencia en el mercado del proveedor para adoptar una línea de producción sustantiva.
- Es la falta de estándar en la industria Orientado a Objetos [6].

2.7 Introducción a DB2 UDB

DB2 UDB Universal Database es una *Base de Datos Universal.* Es completamente *escalable, veloz y confiable.*

Corre en modo nativo en casi todas las plataformas como ser: *Windows NT, Sun Solaris, HP-UX, AIX U, OS/2 entre otros.*

DB2 es un software de *base de datos relacional.* Es completamente multimedia, disponible para su uso en la *Web*, muy bueno para satisfacer las demandas de las grandes corporaciones y bastante flexible para servir a los

medianos y pequeños negocios. *DB2 UDB* es un sistema manejador de base de datos relacional fuertemente

escalable. Es suficientemente flexible para atender estructuras e inestructuras manejadoras de datos necesarias para usuarios simples de grandes empresas. Es conveniente para una gama amplia de aplicaciones de los cliente, quienes pueden desplegar una variedad de plataformas de hardware y software desde dispositivos manuales a los sistemas multiprocesador paralelos masivos.

2.7.1 Características Generales del DB2 UDB

DB2 UDB es el producto principal de la estrategia de *Data Management de IBM.*

DB2 UDB es un sistema para administración de *Bases de Datos Relacionales (RDBMS).* Es multiplataforma, especialmente diseñada para ambientes distribuidos, permitiendo que los usuarios locales compartan información con los recursos centrales. Es el sistema de gestión de datos que entrega una plataforma de base de datos flexible y rentable para construir un sistema robusto para aplicaciones de gestión.

DB2 UDB libera los recursos con amplio apoyo al *open source* (fuente abierta) y plataformas de desarrollo populares como *J2EE y Microsoft .NET.*

Integridad

El *DB2 UDB* incluye características de *Integridad*, asegurando la protección de los *datos* aún en caso de que los sistemas sufran un colapso, y de *Seguridad* permitiendo realizar respaldos en línea con distintos grados de granularidad, sin que esto afecte la disponibilidad de acceso a los *datos* por parte de los usuarios.

Múltiples Usos

Provee la capacidad de hacer frente a múltiples necesidades, desde *Procesamiento Transaccional de Misión Crítica (OLTP),* hasta análisis exhaustivo de los datos para el soporte a la toma de decisiones *(OLAP).*

Escalabilidad

Sus características distintivas de *Escalabilidad* le permiten almacenar información en un amplio rango de equipos, desde un PC portátil hasta un complejo ambiente de mainframes procesando en paralelo.

Web Enabled Para e-Business

Incluye tecnología basada en *Web* que permite generar aplicaciones en las Intranets y responder a las oportunidades de negocios disponibles en *Internet*.

Facilidad de Instalación y Uso

La primera versión de *DB2 para NT* fue reconocida en el mercado como una base de datos muy poderosa, pero difícil de instalar y usar.

En esta versión *(DB2 UDB), IBM* agregó muchas herramientas gráficas para facilitar el uso para los usuarios, como también para los administradores y desarrolladores. Dicha versión incluye guías para operaciones como

instalación, configuración de performance, setup, etc. Además, se agregaron herramientas para facilitar las tareas de integración con otras bases de datos, tecnologías de networking y desarrollo de aplicaciones.

Universalidad

DB2 UDB es, además, la única base de datos realmente universal; es multiplataforma (16 plataformas - de las cuales 10 no son de IBM), brinda soporte a un amplio rango de clientes, soporta el acceso de los datos desde Internet y permite almacenar todo tipo de datos:

- Texto, Audio, Imágenes y Video *(AIV Extender)* (ver fig. 2.5 de la pág.29) .
- Documentos XML *(XML Extender)* (ver fig. 2.6 de la pág.29).

Figura 2.5: AIV Extender

Figura 2.6: XML Extender

Ejemplos de los Formatos de Datos Soportados Por DB2 UDB

- **Video**: playback, streaming, etc.
- **Imágenes**: almacenamiento y búsqueda por patrones de colores y texturas (ver fig. 2.7 de la pág.30).
- **Audio**: maneja diferentes formatos de audio.

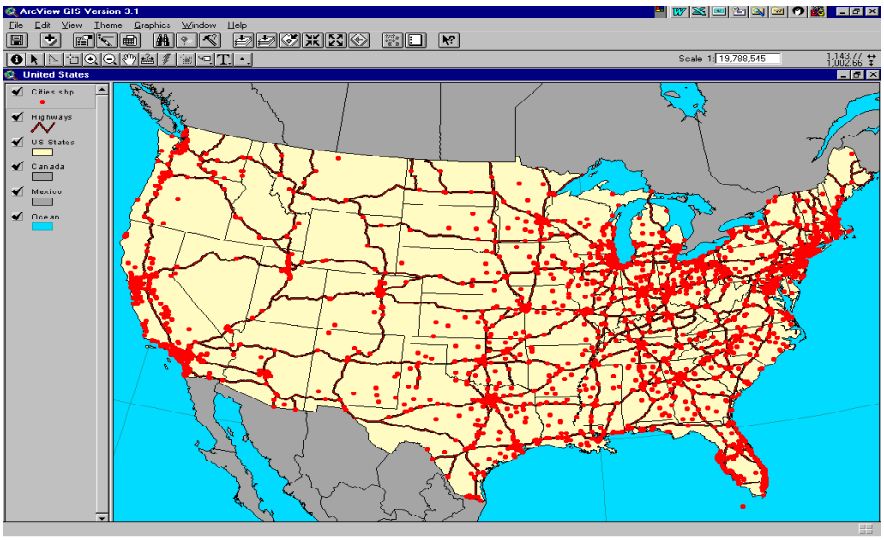

Figura 2.7: Almacenamiento de Imagenes en DB2

Permitiendo realizar :

- **Búsquedas Textuales**: Net Search Extender, Text Extender.
- **Información Espacial**: Spatial Extender, Geodetic Extender.

El *Spatial Extender DB2 y el Geodetic Extender DB2* utilizan diversas tecnologías de la base de datos. El *Spatial Extender* utiliza un mapa plano (o planar), basado en coordenadas proyectadas. Sin embargo, ninguna proyección del mapa puede representar fielmente la tierra entera porque cada mapa tiene bordes; mientras que, la tierra no tiene bordes.

El *Geodetic Extender* utiliza un elipsoide como su modelo para tratar la tierra como un globo íntegrol que no tiene ninguna distorsión en los postes o bordes en el 180^o del meridiano.

2.7.2 Funciones Complementarias del DB2 UDB

Conectividad

Las herramientas de *conectividad* permiten acceder a los *datos* más allá de donde ellos se encuentren. El slogan cualquier cliente, a cualquier servidor, en cualquier red está completamente sustentado por la funcionalidad que sus herramientas ofrecen. *DB2* permite acceder a los *datos* de *DB2* en *mainframe* o *AS/400*, desde *Windows NT*, *Windows 95/98*, *OS/2* o cualquiera de los *Unix* soportados. Además, el producto *Datajoiner* posibilita acceder de forma única y transparente a los datos residentes en *Oracle, Sybase, Informix, Microsoft SQL Server, IMS, VSAM* y otros.

Data Warehousing

El *DB2 UDB* provee la infraestructura necesaria para soportar el proceso de toma de decisiones en cualquier tamaño y tipo de organización. Está dirigido a resolver la problemática a nivel departamental *(Data Marts)*, ya que un único producto provee la capacidad para acceder a datos en *Oracle, Sybase, Informix, Microsoft SQL Server, VSAM o IMS,* además de la familia *DB2.*

Permite de forma totalmente gráfica acceder, transformar y distribuir los datos automáticamente y sin programar una línea de codigo (ver fig. 2.8 de la pág.31).

Data Mining

Las empresas suelen generar grandes cantidades de información sobre sus procesos productivos, desempeño operacional, mercados y clientes. Pero el éxito de los negocios depende por lo general de la habilidad para ver nuevas tendencias o cambios en las tendencias.

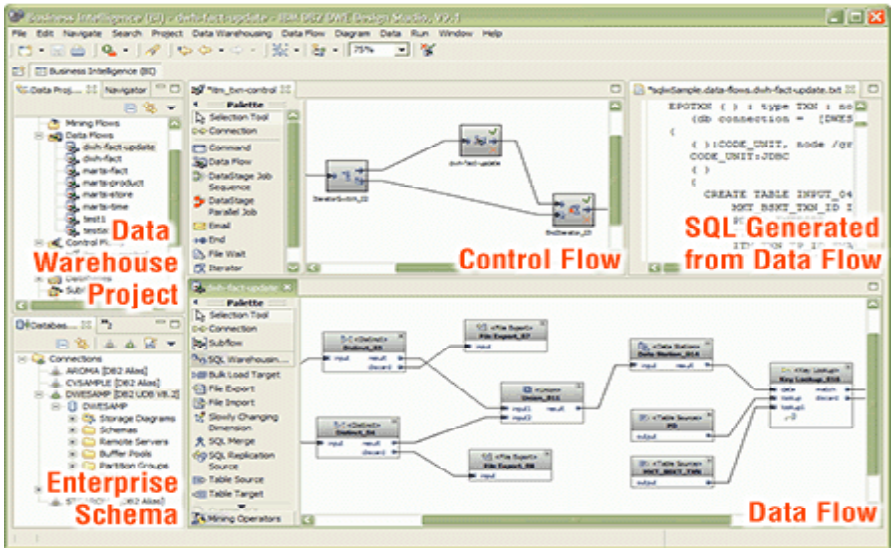

Figura 2.8: DB2 Data Warehouse Edition Design Studio

Las aplicaciones de *Data Mining* pueden identificar tendencias y comportamientos, no sólo para extraer información, sino también para descubrir las relaciones en bases de datos que pueden identificar comportamientos que no son muy evidentes (ver fig. 2.9 de la pág.32).
DB2 UDB posibilita el análisis orientado al descubrimiento de información escondida en los datos, realizando modelización predictiva, segmentación de la base de datos, análisis de vínculos, o detección de desviaciones.

Incluye las siguientes técnicas:

- *Clustering (segmentación).*
- *Clasificación.*
- *Predicción.*
- *Descubrimiento Asociativo.*
- *Descubrimiento Secuencial de Patrones.*
- *Descubrimiento Secuencias Temporales.*

Todas las técnicas mencionadas permiten realizar:

Figura 2.9: *Data Mining*

- *Segmentación de Clientes.*
- *Detección de Fraudes.*
- *Retención de Clientes.*
- *Ventas Cruzadas.*
- *Etc.*

2.8 Business Intelligence Para DB2 UDB

Las ediciones del *DB2 Data Warehouse* proporcionan gran funcionalidad de *BI (Business Intelligence)* dentro de las bases de datos.

Estas nuevas ediciones combinan la fuerza del *DB2 UDB* a la infraestructura esencial de *Business Intelligence*.

La tecnología basada en las ediciones del *DB2 UDB Data Warehouse*, permite integrar la información en:

- **Tiempo Real.**
- **Percepción.**
- **Toma de Decisiones.**

El *DB2 Data Warehouse* hacen más fácil, la implementacíon de la plataforma completa de *BI (Business Intelligence)* del *DB2* , realizada por los clientes. Proporcionan gran funcionalidad de BI dentro de la base de datos.

La Edición Corporativa de DB2 Data Warehouse representa el marco más reciente de la evolución de DB2.

2.8.1 Funcionalidad de Business Intelligence

La funcionalidades del Business Intelligence incluyen:

- Servicios robustos de *ETML (Extract, Transform, Load and Move)* con agentes distribuidos para maximizar el desempeño.
- Aplicaciones eficaces de búsqueda de datos para modelado y visualización de rutinas y resultados de búsqueda, así como para la integración de aplicaciones analíticas
- Optimizaciones integradas a *OLAP* para acelerar el desarrollo e implementación de aplicaciones analíticos
- Soporte para configuraciones de servidores en cluster, *MPP (Massively Parallel Processing)* en una arquitectura true shared-nothing.
- Funcionalidad de administración de consultas y recursos para controlar, administrar y monitorear el ambiente de carga, de consultas y de actividades.
- Y además, todos los recursos de desempeño y funcionalidad de *Business Intelligence* en el *DB2 UDB Enterprise Server Edition* y más...(ver fig. 2.10 de la pág.35)

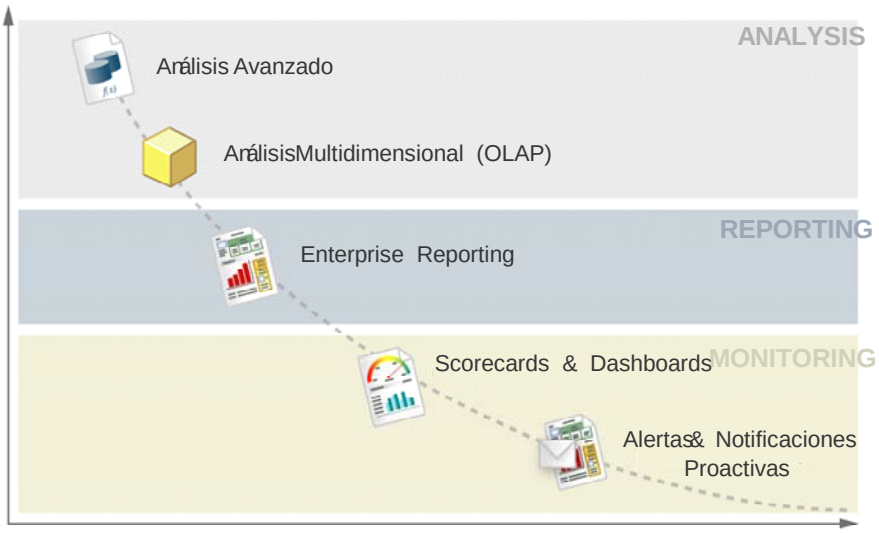

Figura 2.10: Herramientas del BI *(Business Intelligence)*

2.9 DB2 Data Warehouse

Son sistemas que contienen datos de operaciones que se ejecutan en las transacciones diarias de una empresa. Estos contienen información que es útil para los analistas comerciales. Por ejemplo: Los analistas pueden utilizar información sobre qué productos se han vendido, en qué regiones y en qué época del año para buscar anomalías o para proyectar ventas futuras (ver fig. 2.11 de la pág.35).

2.9.1 Esquema Conceptual de un DB2 Data Warehouse

El *Data Warehouse* se define en el *Centro de Depósito de Datos* del *DB2 UDB* para automatizar los procesos necesarios para poblar y mantener el depósito de datos.

Antes de definir el depósito se reúne información acerca de los datos operativos que se van a utilizar como entrada para el depósito y de los requisitos para los datos de depósito.

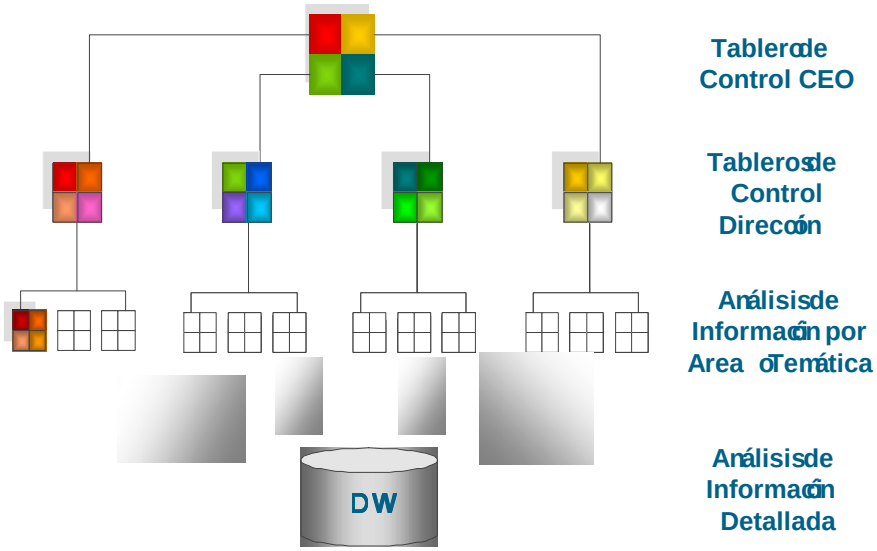

Figura 2.11: Jerarquías de la Informacíon

El *Administrador de la Base de Datos* es el responsable de los datos operativos, es una buena fuente de información acerca de los datos operativos.

Los usuarios de negocios que tomarán decisiones comerciales basadas en los datos del depósito son una buena fuente de información acerca de los requisitos del depósito.

Y finalmente se crea una *Base de Datos* que contendrá las tablas de destino del depósito, que consisten en datos de fuente de depósito limpios y transformados (ver fig. 2.12 de la pág. 36).

EL *DB2 UDB* incluye funciones y funcionalidades que transforman al *DW* en una plataforma que permite distribuir y manejar información multidimensional a través de la empresa. Estas convierten el *Warehouse Relacional* en una plataforma para el análisis *OLAP* de alta performance que permite el despliegue de los datos contenidos en cubos multidimensionales a lo largo de la empresa.(ver fig. 2.13 de la pág.37).

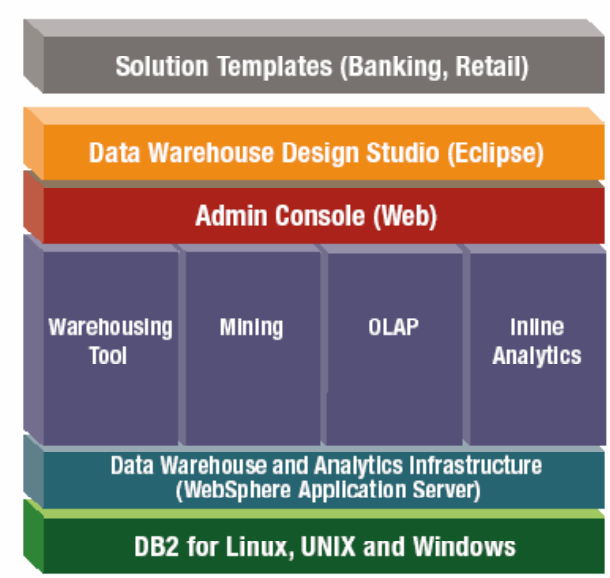

Figura 2.12: Infraestructura completa para DW

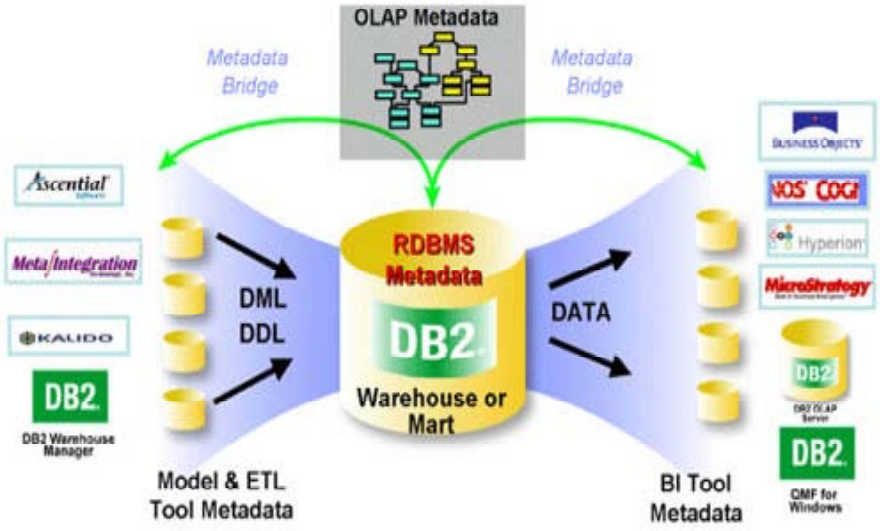

Figura 2.13: DB2 Data Warehouse

2.9.2 Principales Problemas del DB2 Data Warehouse

Se producen diversos problemas si los analistas acceden a los datos de operaciones directamente:

- Puede que no tengan conocimientos suficientes para consultar la base de datos de operaciones. Por ejemplo: La consulta de bases de datos IMS requiere un programa de aplicación que utiliza un tipo especializado de lenguaje de manipulación de datos. En general, los programadores que tienen los conocimientos para consultar la base de datos de operaciones tienen un trabajo a tiempo completo efectuando el mantenimiento de la base de datos y sus aplicaciones.
- El rendimiento es esencial para muchas bases de datos de operaciones, como por ejemplo las bases de datos de un banco. El sistema no puede manejar la realización de las consultas pertinentes por parte de los usuarios.
- Los datos de operaciones no acostumbran a estar en el formato óptimo para que los utilicen los analistas comerciales. Por ejemplo, los datos de ventas que se resumen por producto, región y temporada son mucho más útiles para los analistas que los datos sin clasificar.

Capítulo 3

Introduccion al WebSphere Studio

3.1 Introducción y Conceptos

WebSphere Studio Application Developer es un productos se ha desarrollado basado en el *Workbench* (banco de trabajo) de *Eclipse* [17].

La plataforma del *Workbench* de *Eclipse* fue diseñada por *IBM* y lanzado a la comunidad de *open-source* (código abierto).

Este *Workbench* se ha diseñado para proveer la máxima flexibilidad en el desarrollo de las herramientas y las nuevas tecnologías que pueden emerger en el futuro.

Los ambientes de desarrollo realizados para el *Workbench* deben apoyar a el modelo de desarrollo *role-based* (basado en roles).

La familia del *WebSphere Studio Application Developer* se basa en un ambiente integrado de desarrollo *(IDE)*, donde este permite: Desarrollar, Probar, Eliminar errores y desplegar su usos. Donde también proporciona la ayuda para cada fase del desarrollo del ciclo vida.

Los líderes de la industria de software como: *IBM, Borland, Merant, QNX Software Systems, Rational Software, RedHat, SuSE, TogetherSoft y WebGain* formaron inicialmente la *eclipse.org* que actualmente administra los directores

del *Eclipse open source project.*

Eclipse es una plataforma abierta para la integración de herramienta construida por una comunidad abierta de los abastecedores de la herramienta.

Está plataforma proporciona herramienta con la última flexibilidad y control sobre su tecnología del software.

Eclipse se ha diseñado desde la necesidad de Construir, Integrar los desarrollos útiles del uso de las tecnologías.

El valor más importante que tiene esta plataforma es: el rápido desarrollo de herramienta siendo esta una de las características basadas en un modelo *plug-in* (con enchufe) (ver fig. 3.1 de la pág. 40).

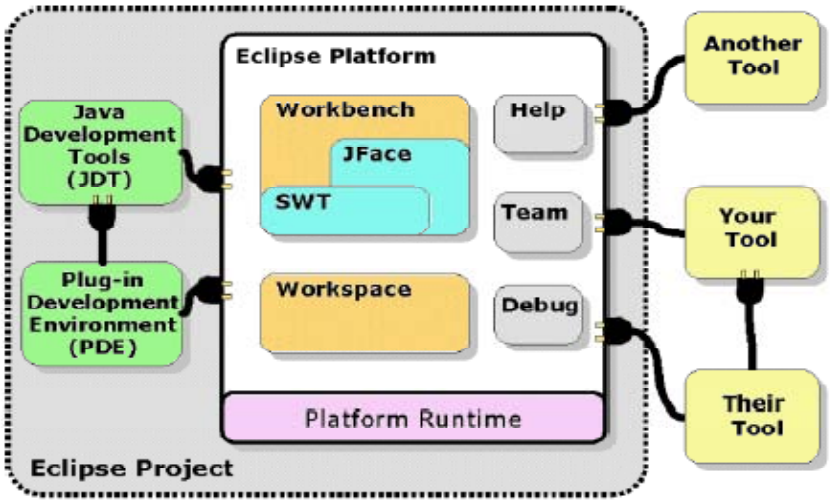

Figura 3.1: Descripción de la plataforma del Eclipse.

3.2 Productos WebSphere Studio

WebSphere Studio (ver fig. 3.2 de la pág. 41) es actualmente conocida como una familia de productos de software propietario de *IBM*, aunque el término se refiere de manera popular a uno de sus productos específicos: *WebSphere Application Server (WAS)* [3].

Figura 3.2: IBM *WebSphere Studio*

Todos los productos del *WebSphere Studio* fuerón construidos sobre el *Workbench* de *Eclipse* como un sistema de *plug-ins* conforme al estándar *APIs* del *Workbenchs.*

La familia del *WebSphere Studio* tiene actualmente los siguientes miembros (ver fig. 3.3 de la pág. 42):

- *WebSphere Studio Site Developer Advanced .*
- *WebSphere Studio Application Developer .*
- *WebSphere Studio Application Developer Integration Edition .*
- *WebSphere Studio Enterprise Developer .*

Estos productos proporcionan la ayuda para el desarrollo, la prueba, y el despliegue end-to-end del *Web* y de los usos de *J2EE (Java 2 Enterprise Edition).*

Cada producto de la familia *WebSphere Studio* presenta el mismo entorno de desarrollo integrado *(IDE)* y una base común de herramientas, por ejemplo para el desarrollo *Java y Web* (ver fig. 3.4 de la pág. 42).

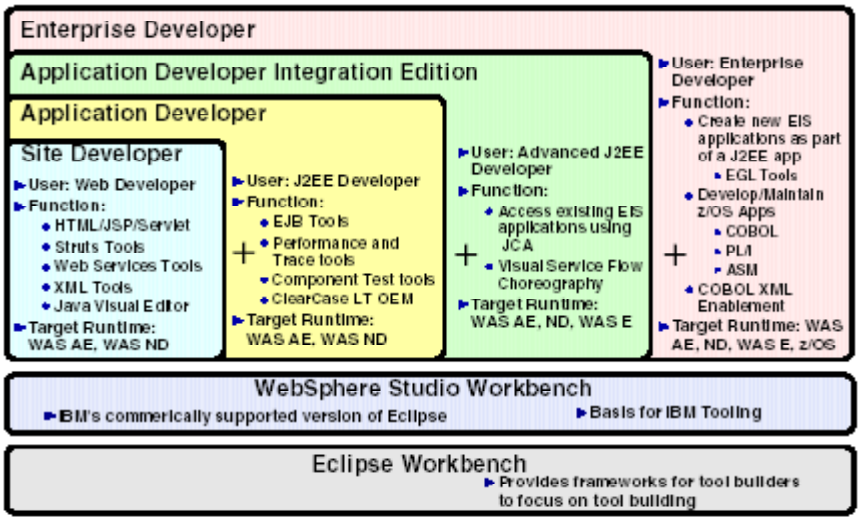

Figura 3.3: La familia del WebSphere Studio.

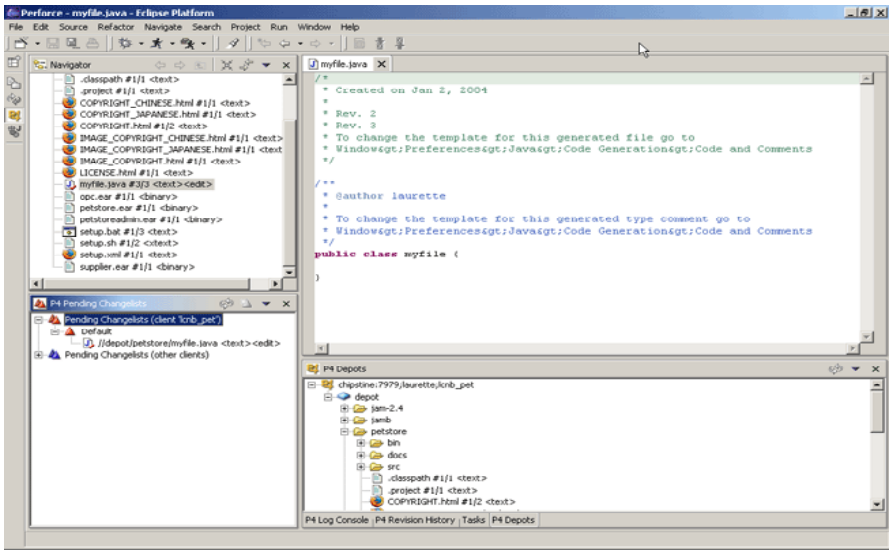

Figura 3.4: *WebSphere Studio* posee un único entorno

WebSphere Studio es único entorno de desarrollo completo, diseñado para satisfacer todas las necesidades de desarrollo, desde interfaces *Web* a aplicaciones del lado del servidor en desarrollo individual a desarrollos avanzados en equipo, desde el desarrollo *Java* a la integración de aplicaciones. Además proporciona un conjunto de herramientas para facilitar el desarrollo de aplicaciones. Posee un entorno visual para la distribución de los elementos de una página *Web* usando *Java Server Pages (JSPs), HTML y Java Script,* ayudando a un rápido desarrollo aplicaciones de *e-business (comercio electrónico)* con contenido dinámico.

Es ideal para el desarrollo de aplicaciones *multiplataforma*, comenzando desde pequeños sitios *Web* hasta megasitios. Proporciona código preconstruido y pretesteado. Permitiendo administrar cargas pico en los servidores *Web*.

3.2.1 WebSphere Studio Site Developer

Site Developer es un *IDE* provisto para los desarrolladores *Web* y a los que administran y manejan site complejos [22].

Es un ambiente fácil de utilizar que reduce al mínimo el tiempo y el esfuerzo requerido para crear, maneja, y elimina errores del sitio *Web* multi-plataforma. Se diseña deacuerdo las especificaciones de *J2SE , J2EE* y soporta *JSPs, servlets, HTML, Javascript, y DHTML.* Además incluye herramientas para desarrollar imágenes y GIFs animado.

Site Developer le permite a los desarrolladores Web utilizar sus herramientas para la creación local incorporado la posibilidad de publican los proyectos remotamente.

Empleando *Site Developer* se podrá desarrollar las aplicaciones Web que utilizan las siguientes tecnologías:

- *JSPs:* Es una manera simple, rápida, y firme de ampliar la funcionalidad del servidor web y de crear el contenido dinámico de la Web.
- *Servlets:* Es el código del servidor que se ejecuta dentro de la aplicacion del servidor Web.
- *Servicios de la Web:* Son aplicaciones independientes, modulares que pueden ser representadas o publicadas sobre el *Internet* o dentro de

Intranets.

3.2.2 WebSphere Studio Application Developer

Application Developer fué diseñado para los desarrolladores profesionales de *Java* y de los utilizan el *J2EE*, y quiénes requieren integrar *Java , Web y XML*, con la ayuda de servicios de la *Web*.

Incluye todas las características del *Site Developer*, y además se agregan las herramientas para el desarrollo de aplicaciones EJB, así como funcionamiento instrumentos copiadores que registran tanto para ejecución local como para remota.

Los desarrolladores pueden construir y probar rápidamente la lógica de negocio y realizar las presentaciónes con instrumentos creados dentro de la Web por herramientas del *Application Developer IDE* antes que despliegue en un servidor.

Utilizando el desempeño de las herramientas copiadoras y trazadoras, es posible descubrir los embotellamientos del funcionamiento de las aplicación de forma temprana en el ciclo de desarrollo.

Además, el ambiente de prueba incorporado por el *WebSphere Application Server* posee instrumentos avanzados para la ayuda de la generación de código que acortan el ciclo de prueba.

3.2.3 WebSphere Studio Application Developer Integration Edition

Integration Edition incluye toda la funcionalidad en el *Application Developer*, más:

- Poderosas herramientas gráficas para ayudar rápidamente y fácilmente la construccion adaptadores para integrar *J2EE* con el *back-end* del sistemas, ayudando a ahorrar tanto en tiempo como en dinero por reutilizando recursos existentes.
- Las herramientas visuales *flow-based* aumentan la productividad, permitiéndonos visualmente definir la secuencia y el flujo de información

entre artefactos de aplicación como adaptadores, Enterprise JavaBeans componentes y servicios *Web*.

3.2.4 WebSphere Enterprise Developer

Enterprise Developer incluye toda la funcionalidad *WebSphere Studio Application Developer Integration Edition* entre otros más:

- Ambientes transaccionales integrados tales como *CICS e IMS*.
- Desarrollar y mantener las aplicaciones *z/OS*.
- Soportá *Java, COBOL, PL/I, y EGL* (enterprise generation language).
- Puede implementar estructuras basadas en aplicaciones *MVS* utilizando conectores y *EGL*.

Otra tecnología que se que se integra en el *Enterprise Developer*:

- *WebSphere Studio Asset Analyzer (WSAA):* Identifica procesos en uso para conectar puntos, y proporcionar asi la capacidad de generar componentes del código existente

3.3 Entorno de Desarrollo de WSAD

WebSphere Studio Workbench, es una herramienta de integración abierta y extensible sobre la que es posible construir diferentes herramientas de terceros *(plug-ins)*(ver fig. 3.5 de la pág. 46). El *Workbench* está basado en la plataforma *open-source Eclipse* , y constituye la base de la siguiente generación de herramientas de desarrollo *IBM*.

WebSphere Studio Enterprise Developer es el entorno que acabará sustituyendo a *VisualAge Generator*.

Tanto si partimos de entornos de desarrollo *IBM (VisualAge for Java)* o de entornos de otros fabricantes *(WebGain VisualCafé, BEA WebLogic)*.

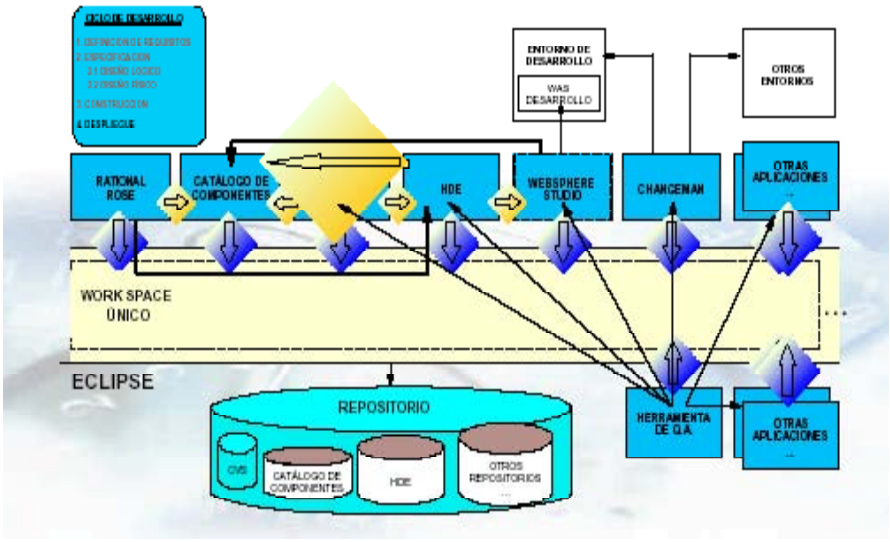

Figura 3.5: WSAD posee un entorno completo integrado

3.4 Ventajas de Migrar a WSAD

La ventaja fundamental consiste en la integración de todos los entornos de desarrollo *Java, Web* en una única plataforma de desarrollo.

3.4.1 J2EE

- Herramientas de importación/exportación, generación de código, edición de deployment descriptors estandars, extensiones y bindings (mapeos) específicos para *WebSphere Application Server (WAS).*
- Herramienta de mapeo *EJB-RDB* soportanto tanto top-down, como bottom-up y meet-in-the-middle.
- Herramientas de edición gráfica de esquemas de bases de datos.
- Herramientas para la creación, edición y validación de ficheros *EAR.*
- Editores para deployment descriptors (ejb-jar.xml y application.xml).

3.4.2 Desarrollo Java

- Nuevo *Editor Visual Java* para *GUIs (Swing y AWT).*
- Nueva generación de *JavaDoc.*
- Soporte *JDK 1.3.*
- Capacidad de utilizar diferentes *JREs.*
- Compilación incremental automática.
- Posibilidad de ejecutar código incluso con errores.
- Protección contra crashs y auto-recovery.
- Error Reporting y corrección.
- Editor Java con asistente contextual.
- Herramientas de refactoring de código.
- Búsquedas inteligentes y herramientas para comparar código y "merge".
- Scrapbook para evaluación rápida de código.

3.4.3 Web Services

- Nuevo soporte *UDDI Version 2.*
- Soporte *UDDI* privado.
- Nuevo soporte de *WSIL.*
- Posibilidad de crear un web service a partir de un fichero *ISD.*
- Visualización de *UDDI business entry* para localización de web services existentes.
- Creación de web services a partir de código existente *(JavaBeans, RLSs, DB2 XML Extender calls, procedimientos almacenados DB2 y queries SQL).*
- Crear wrappers *SOAP y HTTP GET/POST* de código existente.

- Generación de proxies desde el *Web Services Client/Wizard* para tratar mensajes *SOAP*.
- Generación de una aplicación de ejemplo, a partir de la cual crear el resto.
- Realizar el test de un web service local o remoto.
- Deployment de un web service sobre el entorno de test de tanto *WebSphere Application Server* como Tomcat.
- Publicar web services en un *UDDI business registry*.
- Nuevos menús pop-up para la creación y consumo de web services, además de los típicos wizards.

3.4.4 XML

- Entorno totalmente visual.
- Editor de *XML* con posibilidades de validación de documentos.
- Editor de *DTD* con posibilidades de validación de documentos.
- Editor de *XML* schemas.
- Editor de *XSL*.
- Debugger de *XSL* y herramienta de transformación para aplicar *XSL a XML*.
- Editor de mapping *XML - XML*.
- Wizard de creación de *XML* a partir de *queries SQL*.
- Editor de mapping *RDB - XML*.

3.4.5 Desarrollo Web

- Nuevo soporte para *XHTML y Struts*.
- Nuevo entorno visual de construcción de aplicaciones basado en struts.
- Editor visual de *HTML y JSPs*.

- Edición y validación de *JavaScript.*
- Soporte de *JSP* Custom tags (taglibs) 1.2.
- Edición de imágenes y animaciones.
- Edición de *CSS.*
- Importación via *HTTP/FTP.*
- Exportación vía *FTP* a un servidor.
- Visualización de links, broken links, etc.
- Wizards para la creación de servlets.
- Wizards para la creación de proyectos *J2EE.*
- Wizards para la creación de aplicaciones web.

3.4.6 Testing y Deployment

- Incrementa la productividad de forma muy importante.
- Entorno ligero de carga rápida.
- Permite pruebas unitarias locales.
- Permite debugger de código en el servidor a través del debugger integrado.
- Permite configurar deiferentes aplicaciones web.
- *TCP/IP* monitoring server.
- Permite instalar los siguientes entornos, tanto locales como remotos: *(WebSphere Application Server AEs Version 4.0.3 and Version 5, WebSphere Application Server - Express Version 5, Apache Tomcat).*

3.4.7 Tracing, Monitoring y Performance

- Performance Analyzer muestra los tiempos de ejecución y ayuda a detectar memory leaks.
- Muestra información de los objetos existentes.
- Tiene capacidades de *"Pattern extraction".*
- Es posible monitorizar varios procesos simultaneamente, incluso corriendo en diferentes máquinas.
- Codificación por colores de las clases.
- Presentación de los resultados en modo gráfico y estadístico.
- Soporte de profiling a nivel de objetos.
- Análisis de los logs de *WebSphere Application Server* e interacción con la bases de datos de problemas.
- Edición de items en la base de datos de problemas.

3.4.8 Debugger

- Muy similar al existente en *VisualAge for Java.*
- Permite realizar debug tanto a código local como a código residente en el servidor.

Capítulo 4

Introducción a Intelligent Miner for Data

4.1 Introducción a la Minería de Datos

La *Minería de Datos* es el proceso de descubrir nuevas y ú tiles correlaciones, patrones y tendencias dentro de grandes cantidades de datos almacenadas en repositorios, utilizando tecnología para el reconocimiento de patrones así como técnicas matemáticas y estadísticas.

Minería de Datos es el análisis de conjuntos de datos (comunmente grandes) de observaciones para encontrar relaciones inesperadas y presentar los datos en formas que sean tanto entendibles como útiles para el dueno de la información [4].

La Minería de Datos es un campo interdisciplinario (ver fig. 4.1 de la pág. 52) *que conjunta diferentes técnicas desde inteligencia artificial, reconocimiento de patrones, estadística, bases de datos y visualización para realizar la extracción de información dentro de grandes cantidades de datos* [11].

4.1.1 Etapas del Proceso de Minería de Datos

Es un proceso que permite descubrir información novedosa y válida, partiendo de grandes almacenes de datos. Donde este proceso implica:

Figura 4.1: La Minería de Datos es un campo multidisciplinario

- *Seleccionar y transformar datos de entrada.*
- *Ejecutar una función de minería.*
- *Interpretar los resultados obtenidos.*

Puede ser un proceso iterativo, siempre en busca de la obtención de la mejor calidad en el resultado.

No se la debe confundir con un gran software. Es un proceso que utiliza diferentes aplicaciones software en las diferentes etapas.

Las etapas del proceso de *Minería de Datos* abarca las siguientes:

- *Selección* de los datos de entrada.
- *Transformación* de los datos de entrada.
- *Minería de datos.*
- *Interpretación de los resultados* obtenidos.(ver fig. 4.2 de la pág. 53).

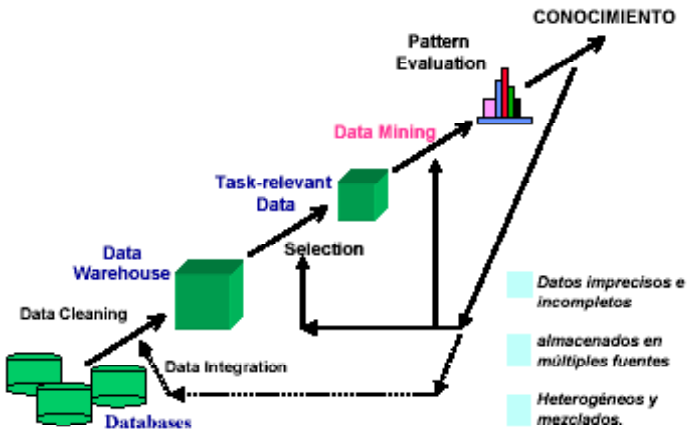

Figura 4.2: Los procesos que abarca la *Minería de Datos*

4.2 Introducción al IBM Intelligent Miner for Data

Intelligent Miner comunica las funciones de minería con las de preproceso en el servidor, así como las herramientas de administracion con las de visualizacion en el cliente. Se pueden tener componentes de cliente y servidor en la misma máquina [14].

El componente cliente incluye una interfaz de usuario desde la cual se pueden invocar funciones de un servidor de Intelligent Miner. Los resultados se devuelven al cliente, en el que se pueden visualizar y analizar.

El software de servidor está disponible para los sistemas *AIX, OS/390, iSeries, Solaris Operating Environment y Windows*, el software de servidor soporta la minería en paralelo con varios procesadores [12].

El *IBM Intelligent Miner for Data* (ver fig. 4.3 de la pág. 55) es un software que comprende un conjunto de funciones: *Estadísticas, Preproceso y Minería* que se utilizan para analizar grandes volúmenes de datos.

Es conveniente tener conocimientos previos de Bases de Datos y de Estadística.

Ofrece ayuda en todas las etapas del proceso de *Minería de Datos.*

4.2.1 Componentes IBM Intelligent Miner for Data

Los componentes que integran *Intelligent Miner* son:

- **Interfaz de usuario**: Programa que permite definir las funciones de mineria de datos en un entorno gráfico. se pueden definir las preferencias de la interfaz de usuario, que están almacenadas en el cliente (ver fig. 4.4 de la pág. 56).

- **API de capa de entorno:** Conjunto de funciones API que controlan la ejecución de procesos y resultados de minería. Las secuencias de funciones y operaciones de minería se pueden definir y ejecutar mediante la interfaz de usuario a través de la API de capa de entorno. La API de capa de entorno está disponible en todos los sistemas operativos servidores.

Figura 4.3: IBM Intelligent Miner for Data Version 8.1

- **Visualizador:** Herramienta que visualiza el resultado generado por una función estadística o de minería. *Intelligent Miner* ofrece un amplio conjunto de herramientas de visualizacíon (ver fig. 4.5 de la pág. 57).
- **Acceso a datos**: Acceso a datos de archivos planos, tabla y vistas de bases de datos.
- **Biblioteca de proceso:** Biblioteca que proporciona acceso a funciones de bases de datos.
- **Bases de minería:** Colección de objetos de minería de datos que se utilizan para un objetivo de minería o un problema de gestión. Las bases de minería se almacenan en el servidor, que permite el acceso desde distintos clientes.
- **Kernels de minería:** Algoritmos que comienzan a operar cuando se ejecuta una minería de datos o una función estadística.
- **Resultados de minería, API de resultado y herramientas para exportación**: datos extraídos por la ejecución de minería o la función estadística.

Estos componentes permiten visualizar los resultados en el cliente. Los resultados se pueden exportar para algún proceso posterior o para utilizarlos con herramientas de visualización.

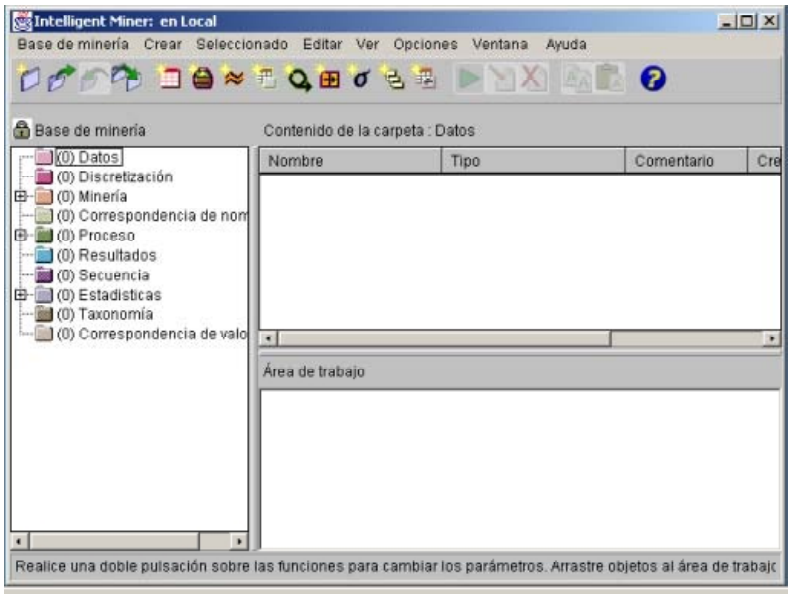

Figura 4.4: Interfaz del usuario, Intelligent Miner for Data

4.3 Instalación e Inicio del Intelligent Miner

4.3.1 Instalación del Servidor para Windows

- **Requisitos de Hardware**: El servidor *Intelligent Miner* para *Windows* se ejecuta en sistemas con procesadores a 300 MHz o superiores. Para ejecutar *IBM DB2 Intelligent Miner for Data* en windows, debe instalar uno de los clientes soportados, en la misma máquina o en una máquina remota. El espacio de almacenamiento necesario varía según la cantidad de datos procesados por ejecución. El mínimo es de 128 MB, pero se recomienda utilizar 512 MB de RAM. El espacio de disco necesario para una demostración del producto depende del tipo de partición del disco duro.

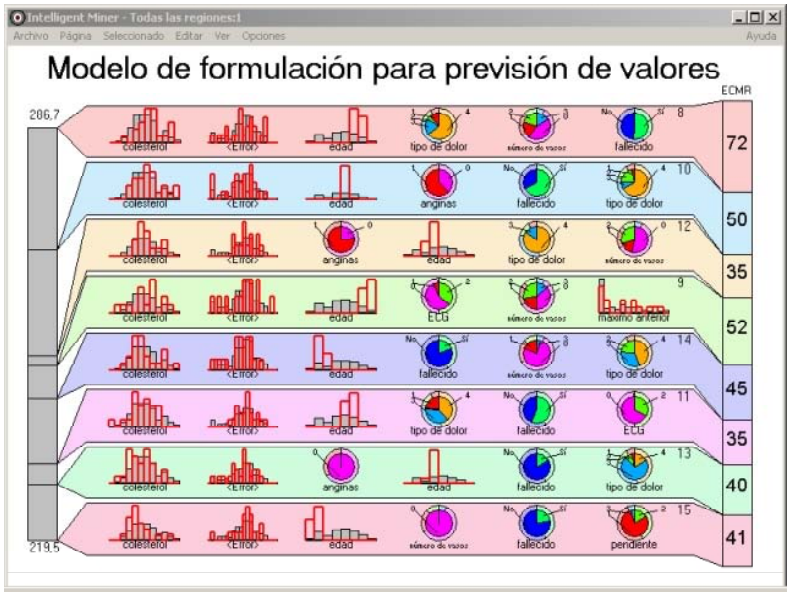

Figura 4.5: Herramientas de Visualización

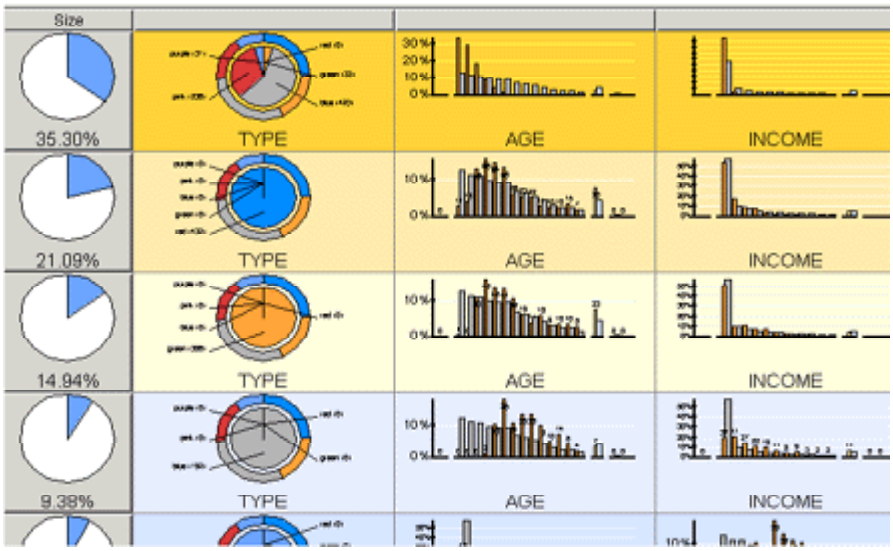

Figura 4.6: Herramientas de Visualización (otra vista)

- **Requisitos de Software**: El servidor *Windows* requiere tener incorporado *Microsoft Windows NT, 2000, XP,* además un servidor *DB2 Universal Database.*

El servidor *Intelligent Miner de Windows* se inicia como un servidor windows nativo denominado *IBM Intelligent Miner.* El servidor *IBM Intelligent Miner* se inicia automáticamente al iniciar el sistema.

4.3.2 Instalación del Cliente Windows

- **Requisitos de Hardware**: El cliente *Intelligent Miner para Windows* se ejecuta en una estación de trabajo con un procesador a 300 Mhz o superior. El espacio de almacenamiento necesario varía según la cantidades de datos procesados por ejecución. El mínimo es de 128 MB, pero se recomienda utilizar 512 MB de RAM. El espacio de disco necesario para una demostración del producto depende del tipo de partición del disco duro.

- **Requisitos de Software**: El cliente *Windows* requiere tener incorporado *Microsoft Windows NT, 2000, XP.*

4.3.3 Conceptos Básicos del Intelligent Miner

En general, la *minería de datos* en *Intelligent Miner* se consigue por medio de la creación de objetos interrelacionados. Estos objetos se muestran como íconos y representan el conjunto de atributos o valores que definen los datos o funciones. se crean objetos de formulación para llevar a cabo una tarea en concreto.

Intelligent Miner crea otros objetos, como objetos de resultado que contienen los elementos encontrados en una ejecucion de minería.

Los objetos de un proyecto de *minería de datos* concreto se guardan como un grupo denominado base de minería. Se puede crear una base de minería para cada objetivo o proyecto de minería.

Cuando se trabaja con *Intelligent Miner*, una de las tareas fundamentales consiste en crear objetos de formulación.

4.3.4 Funciones de Minería del Intelligent Miner

Función Asociaciones

El propósito de esta técnica es encontrar elementos de una transacción que impliquen la presencia de otros elementos en la misma transacción.

Suponiendo que se tiene una base de datos con operaciones de compra, y que cada transacción consiste en un conjunto de elementos que el cliente ha adquirido, la función de minería *Asociaciones* podría detectar relaciones entre los elementos del conjunto.

Función Clustering Demográfico

El objetivo de descubrir clusters es agrupar registros que tengan características similares.

Intelligent Miner busca en la base de minería las características que se dan con más frecuencia y agrupa los registros relacionados de acuerdo con ello. El resultado de la función de clustering tiene el número de clusters detectados y las caracteristicas que los constituyen. Además, el resultado muestra la forma en que las características están distribuidas en los clusters.

Suponiendo que se tiene una base de un supermercado que incluye la identificación de los clientes e información acerca de la fecha y la hora de las compras. La función de minería clustering podría agrupar en clusters para permitir la identificación de diferentes tipos de compradores.

El *Clustering Demográfico* proporciona la agrupación de clusters rápida y deforma natural de bases de datos de gran tamaño. Determina automáticamente el número de clusters que se generarán. Las semejanzas entre registros se denominan comparando los valores de los campos. Los clusters se definen para maximizar el *criterio de Condorcet.* Donde el *criterio de Condorcet* es la suma de todas las semejanzas de registros de pares dentro del mismo cluster menos la suma de todas las semejanzas de registros de pares en diferentes cluster.

Función Clustering Neuronal

El objetivo de descubrir cluster es agrupar registros que tengan caracteristicas similares.

Intelligent Miner busca en la base de minaría las características que se dan con más frecuencia y agrupa los registros relacionados de acuerdo con ello. El resultado de la función clustering muestra el número de clusters detectados y las características que los constituyen. Además, el resultado muestra la forma en que las carasterísticas que los constituyen. Además, el resultado muestra la forma en que las características están distribuidas en los clusters.

El *Clustering Neuronal* utiliza una Red neuronal de mapa de características de Kohonen. Los mapas de mapa de características de Kohonen utilizan un proceso denominado organización automática para agrupar los registros de entrada similares. El usuario especifica el numéro de clusters y el número máximo de pasadas sobre los datos. Estos parámetros controlan el tiempo de proceso y el grado de granularidad que se utiliza al asignar los registros de datos a los clusters.

La función principal del *Clustering Neuronal* es buscar un centro para cada cluster. Este centro se denomina también prototipo de cluster. Para cada registro de los datos de entrada, la función de minería *Clustering Neuronal* calcula el prototipo de cluster más cercano al registro.

Con cada pasada sobre los datos de entrada, los centros se ajustan de forma que se logra una calidad mejor en el modelo de clustering global. El indicador de proceso muestra la mejor en la calidad en cada pasada durante la ejecución de la función de minería.

Función Patrones Secuenciales

El objetivo de esta técnica es encontrar todas las apariciones de subsecuencias semejantes en una base de datos de secuencias.

Por ejemplo, suponiendo que se tiene una base de datos de un comerciante que desea optimizar sus compras y el sistemas de almacenamiento de; al realizar una ejecución de minería en estas base de datos se obtendrá los nombres de parejas de secuencias con el grado de semejanza y el número de subsecuencias.

Esta técnica tambien se puede utilizar para identificar empresas con pa-

trones de crecimiento similares, determinar productos con patrones de ventas similares o determinar acciones con movimientos de precios similares. Otro uso puede ser la detección de ondas sísmicas que no sean similares o la localización de irregularidades geológicas.

Función Clasificación en Árbol

Se hacen predicciones de las clasificaciones para crear modelos basados en datos conocidos. Estos modelos se pueden utilizar para analizar la razón por la cual se ha hecho una clasificación o para calcular la clasificación de nuevos datos.

Los datos históricos se componen frecuentemente de un conjunto de valores y de una clasificación de estos valores. Si se analizan los datos que ya se han clasificado se descubrirán las características que han contribuido a realizar la clasificación anterior. El modelo de clasificación resultante se podrá utilizar luego para predecir las clases de registros que contienen nuevos valores de atributos.

Se puede utilizar estas técnicas para aprobar o denegar reclamaciones de seguros, detectar fraudes en las trajetas de crédito, identificar defectos en imágenes de componentes manufacturados y diagnosticar condiciones de error. También se puede aplicar para determinar objetivos de márketing, en el diagnóstico médico para determinar la eficacia de los tratamientos médicos, para la reposición de inventarios o en la planificación de la ubicación de una tienda.

El algoritmo de inducción con árbol ofrece una descripción de fácil comprensión sobre la distribución subyacente de los datos. Este algoritmo realiza un ajuste proporcional con respecto al número de ejemplos de preparación y al número atributos que se encuentran en bases de datos extensas. Es conveniente utilizar este técnica para conocer mejor la estructura de la base de datos o para estructurar las bases de datos que no estén clasificadas.

Función Clasificación Neuronal

Se hacen predicciones de las clasificaciones para crear modelos basados en datos conocidos. Estos modelos se pueden utilizar para analizar la razón por la cual se ha hecho una clasificación o para calcular la clasificación de nuevos datos.

La función *Clasificación Neuronal* emplea una red neuronal de retropropagación para clasificar los datos. La clasificación se basa en el valor de clase y las relaciones de los atibutos descubiertos mediante un proceso de minería realizado en unos datos clasificados anteriormente. El aprendizaje de red significa desarrollar un modelo que represente dichas relaciones, Una red que ha realizado un aprendizaje es una salida de la ejecucíon de minería. El análisis de sensibilidad, otro tipo de salida, se utiliza para comprender la contribución relativa de los campos deatributos en la decisión de clasificación.

Una red nueronal con aprendizaje puede generalizar a partir de su experiencia pasada, y calcular una clasificación razonable incluso tomando como punto de partida combinaciones de atributos que no haya visto nunca.

Función Prediccíon FBR

La finalidad de la predicción de valores es descubrir la dependencia y la variación de un valor de un campo en relación con los valores de otros campos que se encuentren es el mismo registro. Se genera un modelo que puede predecir un valor para ese campo particular en un registro nuevo con el mismo formato, en base a otros valores de campo.

Por ejemplo, un comerciante desea utilizar datos históricos para calcular los ingresos por ventas que puede suponer un cliente nuevo. Una ejecución de minería sobre esos datos históricos crea un modelo. Este modelo se puede utilizar para predecir los ingresos que supondrán las ventas realizadas a un cliente nuevo en base a los datos de éste. El modelo también pude mostrar que las campañas de incentivos dirigidas a algunos clientes mejoran las ventas.

Se puede utilizar el método de función de base radial (FBR) para ajustar datos que son fución de diversas variables. El algoritmo bá sico puede formar un modelo para predecir el valor de un campo determinado partiendo de los valores de otros atributos. Una función base-radial require varios centros de ajuste. Donde un centro de ajuste es un vector del espacio de atributos. En cada uno de estos centros, se define una función de base. La función de base es una función no lineal de distancia desde el centro de ajuste. Por este motivo, las funciones de base se denominan Funcion de base radial: tienen el mismo valor en cualquier punto con la misma distancia o radio desde el centro de ajuste.

Función Prediccíon Neuronal

La finalidad de la predicción de valores es descubrir la dependencia y la variación de un valor de un campo en relacíon a los valores de otros campos que se encuentren en el mismo registro. Se genera un modelo que puede predecir un valor para ese campo particular en un registro nuevo con el mismo formato, en base a otros valores de campo.

La función de minería *Prediccíon Neuronal* crea un modelo que se utiliza para predecir nuevos valores para regresión y pronóstico de series temporales.

Utiliza una red neuronal de retropropagación para predecir valores. La predicción se basa en el valor de predicción y en las relaciones entre los atributos descubiertas al explorar un conjunto de datos de preparación que contienen tanto la variable independiente como las dependientes. Al desarrollo de un modelo que represente estas relaciones se le denomina aprendizaje o preparación de la red neuronal.

Además de la predicción de valores estándar, también denominada regresión, la función Predicción Neuronal ofrece soporte a la predicción de series temporales al permitir que el usuario especifique un horizonte de previsión y un tamaño de ventana de entrada. Estos dos parámetros se utilizan para dar formato a los registros de preparación

internamente para que la red neurinal tome un conjunto de "m" registros consecutivos (el tamaño de la ventana) y prediga el valor dependiente de "n" registros (el hotizonte) en el futuro.

4.3.5 Funciones Estadísticas del Intelligent Miner

Las funciones estadísticas de Intelligent Miner ofrecen diversos métodos de estadísticas y de pronostico para dar apoyo a sus decisiones empresariales.

Se puede utilizar las funciones estadísticas para obtener más información sobre los datos, lo que permitirá tomar decisiones más acertadas cuando se apliquen los procesos de minería a los datos. Las funciones estadísticas se aplican a los datos de entrada y producen datos de salida y resultados.

Las funciones estadísticas de Intelligent Miner aplican distintos cálculos y teorías estadísticas a los datos de entrada para descubrir en ellos patrones ocultos. Dichas funciones se pueden utilizar en los pasos de transformación y

minería del proceso de minería de datos.

Se puede utilizar la función estadística de regresión lineal para predecir valores mediante un modelo de ajuste lineal. Además se puede utilizar el análisis de componentes de principios para var los atributos más dominantes en sus datos.

4.3.6 Funciones de Preproceso del Intelligent Miner

Las funciones de preproceso se utilizan para transformar los datos antes, durante y después de la ejecuciones de minería. Las funciones se ejecutan sobre datos de entrada, y cada función produce datos de salida, excepto enel caso de las funciones Ejecutar *SQL* y Borrar fuentes de datos.

Los datos de entrada consisten en tablas o vistas de bases de datos en un servidor. Las funciones de preproceso nunca modifican los datos de entrada.

Los datos de salida se pueden escrivir en tablas o vistas de bases de datos, excepto la función Copiar registros en archivo, que sólo produce archivos. A fin de evitar la duplicación de los datos, los datos de salida acostumbran a constituir vistas. Si se desea reiteración de datos se pueden utilizar tablas.

En el paso de transformación del proceso de minería de datos, se puede utilizar las funciones de preproceso de Intelligent Miner para preparar los datos para la minería. Podrían excluirse campos o registros de los datos de entrada que no sean relevantes para la finalidad de la minería de datos o realizar operaciones matemáticas sobre los campos de los datos de entrada antes de llevar a cabo la minería de los datos.

4.3.7 Visualización de Resultados

Cuando se finaliza una ejecución de minería, se puede abrir una ventana de resultados que proporcione una visión general inicial. Modificando la representación de los resultados pueden verse aspectos concretos detalladamente.

La mayoría de visores de resultados ofrecen la posibilidad de imprimirlos.

En general, aparacerá el panal estándar de impresión del sistema operativo del cliente cuando seleccione la opción de impresión.

En síntesis, *IBM DB2 Intelligent Miner for Data Versión 8.1* brinda una amplia gama de herramientas que posibilitan el análisis de grandes bases de datos. También ofrece herramientas de visualización para interpretar los resultados de minería.

4.4 Ejemplo Práctico de Visualizador de Asociación

El modelo de asociación se basa en un ejemplo práctico desarrollado, partiendo de la creación del modelo de minería de datos utilizando para ello *Intelligent Miner Modeling* (ver fig.7.15 de la pág.
299).

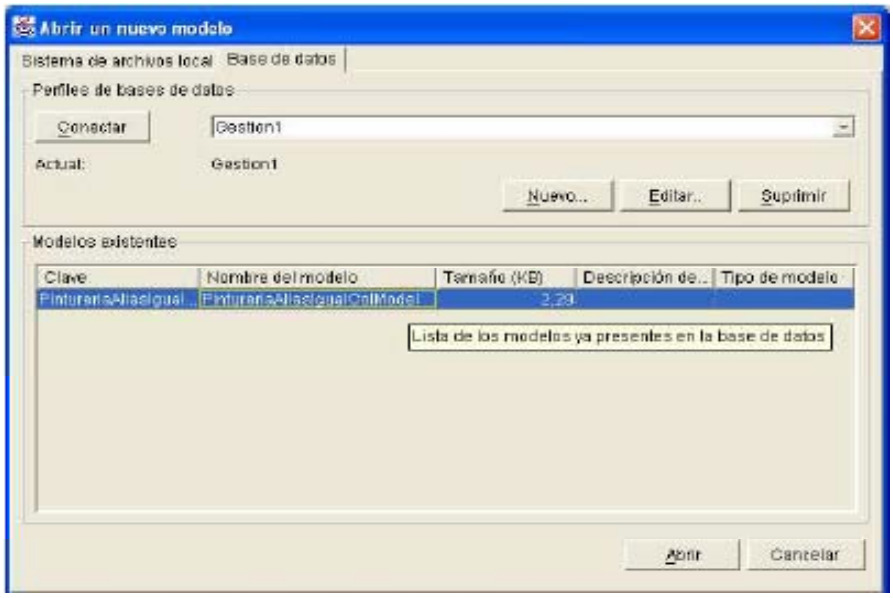

Figura 4.7: Se puede apreciar la ventana del *Intelligent Miner Modeling.*

4.4.1 Vista Reglas

El visualizador de asociación muestra las reglas de asociaciones y los conjuntos de ítems con varios valores de campos, tales como el soporte o la confianza.

Pueden mostrarse las reglas tanto en forma textual como tabular (ver fig.4.8 de la pág. 66).

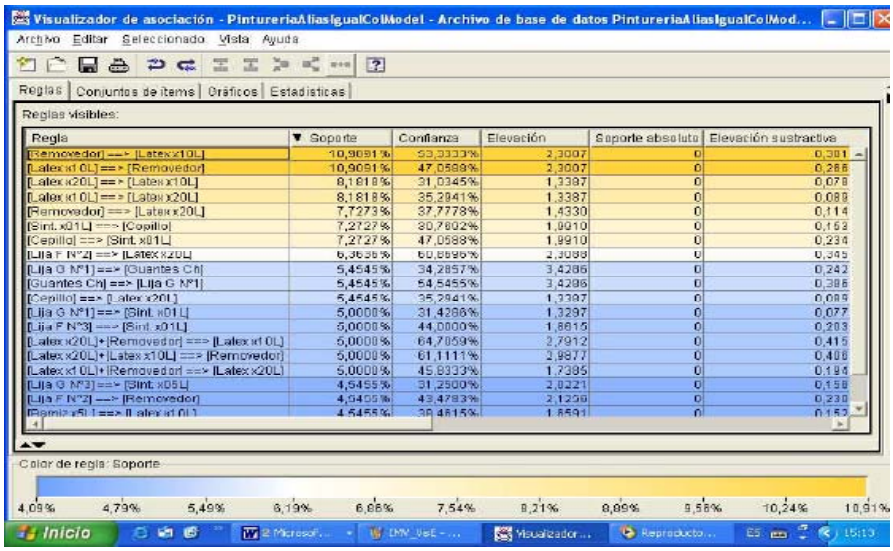

Figura 4.8: Apreciamós así reglas de asociaciones, conjuntos de ítems

Asimismo el usuario puede establecer no sólo los colores sino también la ubicación de los valores de campo, etc.

Una *norma de asociación* consta de:

- Dos conjuntos afines de elementos: el cuerpo de la norma y la cabecera de la norma.
- El soporte de la norma, que es un valor estadístico en forma de porcentaje.
- La fiabilidad de la norma, que es asimismo un valor estadístico en forma de porcentaje.

Por ejemplo, del modelo de la pinturería puede apreciarse que:

Látex x20L [Removedor]

- **Soporte** = 5 %.

- **Fiabilidad** = 64,7 %.

En este caso:

Látex x20L [Removedor] es el *Cuerpo de la norma*

Látex x10L es la *Cabecera de la norma*

El conjunto de elementos [Látex x20L][Removedor][Látex x10L] estaba presente en un 5% de las transacciones de compra consideradas. Este es el valor de soporte.

En las transacciones donde aparecían juntos los elementos [Látex x20L] [Removedor], también estaba presente el elemento [Látex x10L] en un 64,7% de los casos. Este es el valor de fiabilidad.

4.4.2 Vista Conjuntos de Ítems

Muestra los conjuntos de ítems que se incluyen en una regla de asociación (ver fig.4.9 de la pág. 68).

Puede apreciarse la siguiente información:

- *Conjunto de ítems.*
- *Soporte.*
- En reglas como *Cuerpo.*
- En reglas como *Cavezera.*

4.4.3 Vista Gráficos

Los conjuntos de ítems se visualizan como nodos y las reglas de asociaciones como flechas. Las flechas conducen desde los conjuntos de ítems del cuerpo de la regla a los conjuntos de la cabecera de la regla.

El color de los nodos y el color de las flechas representa el valor de un parámetro en particular como, por ejemplo, Soporte o En reglas como cuerpo (ver fig.4.10 de la pág.68).

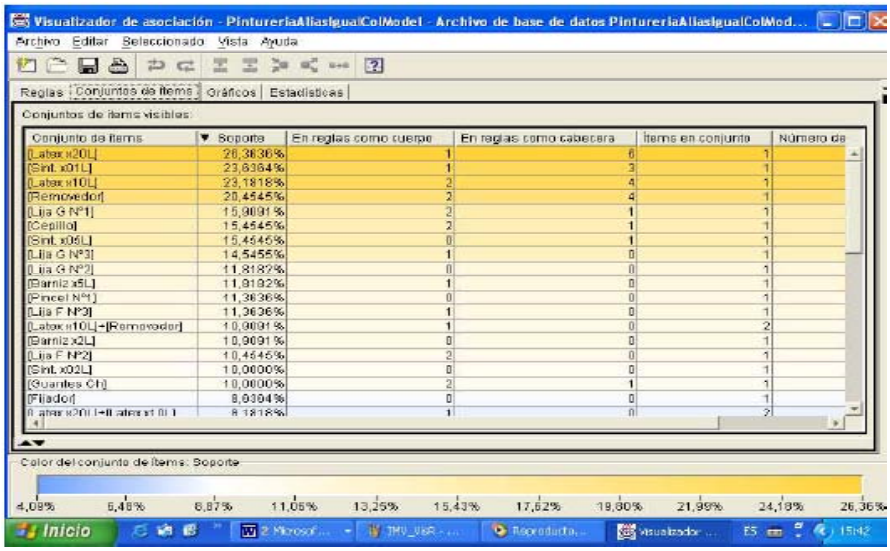

Figura 4.9: Apreciando el Conjunto de Items, Soporte y En reglas.

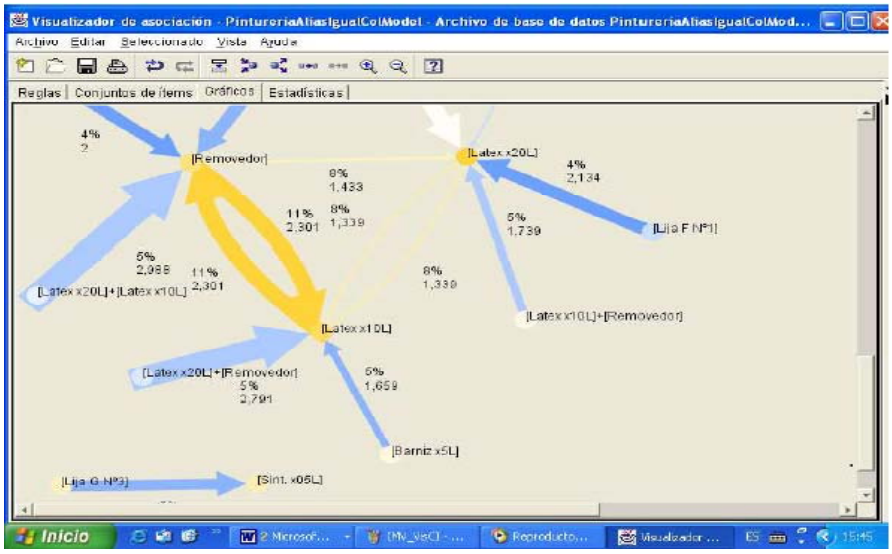

Figura 4.10: Visualizamos así los nodos y las reglas de asociaciones como flechas.

4.4.4 Vista Estadísticas

Incluye las secciones que pueden apreciarse en la siguiente imagen.

La Sección Estadísticas visibles le muestra la cantidad de reglas y conjuntos de reglas del modelo que son visibles en el Visualizador de asociación.

Si se han ocultado reglas o conjuntos de ítems, se visualizará la cantidad de reglas o conjuntos de ítems visibles. Si no se han ocultado reglas ni conjuntos de ítems, se mostrará la cantidad total de reglas y conjuntos de ítems que incluye el modelo (ver fig.4.11 de la pág.69).

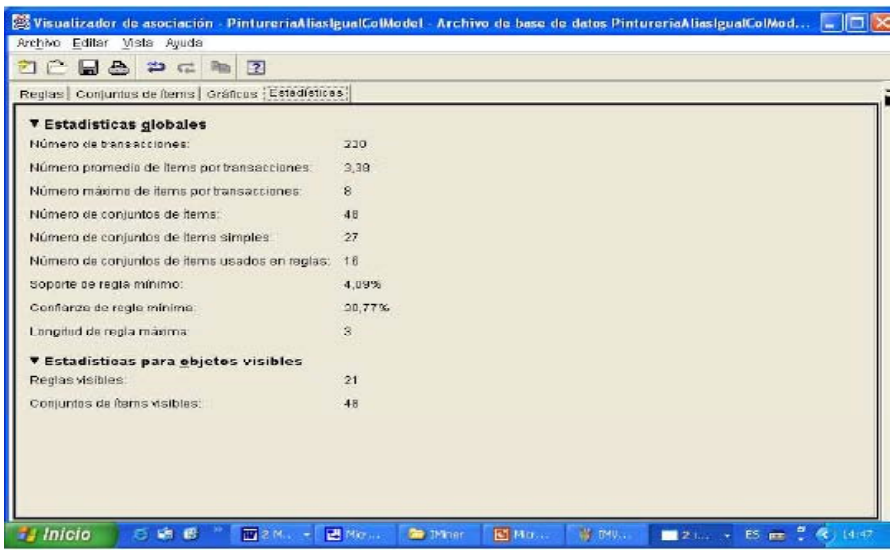

Figura 4.11: Visualizamos en esta caso los valores Estadísticos Globales y Objetos.

Capítulo 5

Preparación del Data Warehouse

5.1 Introducción

En este capítulo se describirán las principales *etapas* para le creación del *Data Warehousing*. Éstas ya se han visto en el Capítulo Nº1 *"Introducción a la Minería de Datos"*, las cuales son [1]:

La fases para la construcción de un *Data Warehousing* son:

- *Fase de Extracción:* Los constructores crean los archivos de la base de datos para transacciones y los guardan en el servidor que mantendrá el almacén de datos.

- *Fase de Depuración:* Se unifica la información de los datos de forma que permita la inserción en el almacén de datos.

- *Fase de Carga:* Se transfiere los archivos depurados a la base de datos que servirá como almacén de datos.

5.2 Intalación del Ambiente Operacional

5.2.1 Selección y Exploración de la Fuente de Datos

Inicialmente se dispone de ocho bases de datos en formato de *Microsoft Access.* Las mismas contienen informacíon de la nueva *EPH (Encuesta Permanente de Hogares)* cuya muestra incluye 25.000 familias de 31 conglomerados urbanos de la *República Argentina* con una frecuencia trimestral.

Cada *base de datos* es llamará base usuaria, compuesta por dos tablas:

- **Hogar** (Ejemplo: *USH_ T107*), base usuaria *Hogar* del primer trimestres del 2007.
- **Persona** (Ejemplo: *USP_ T407*), base usuaria *Persona* del cuarto trimestres del 2007.

Para el desarrollo operacional se han considerado únicamente datos de la base de datos *Persona.* Las diferentes tablas contienen en promedio entre 470.031 y 590.000 registros, con un tamaño de almacenamiento aproximado de 32.000Kb.

Los archivos se visualizan sin inconvenientes con *Microsoft Access.*

5.2.2 Trabajando en Microsoft Access

Al exportar los datos fuentes se utilizan diversos formatos:

- Texto delimitado por coma (*USP_ T107.csv*).
- Texto sin ningún tipo de delimitador: Archivos Planos (*USP_ T107.txt*).

Para este último *Microsoft Access* tiene la opción llamada *Asistente para la vinculación de texto*, que permite transformar el texto base en una tabla relacional con sus corespondientes campos perfectamente definidos.

Para realizar la exportación se recomienda no trabajar con tablas vinculadas, dado que en éstas no se permite realizar modificaciones.

Más allá de que *Intelligent Miner* puede utilizar como datos de entrada archivos planos, es posible el uso del sistema administrador de bases de datos *DB2 UDB Universal Database*, para entre otras cosas, aprovechar las ventajas de las funciones de preproceso de *Intelligent Miner,* las que trabajan con datos almacenados en servidores de bases de datos (no con archivos planos).

Los pasos para llevar a cabo la Exportación son los siguientes:

- Abrir el archivo (*USP_ T107.dbf)* con *Microsoft Access.*
- Seleccionar la tabla (*USP_ T107*) y precionando el botón derecho del mouse y se escoje la opción de *Exportar.*
- Se abrirá una ventana que permitirá selecionar el nombre como así también el tipo de formato del archivo.
- Se inicializa automáticamente el Asistente para la exportacíon, que es el que permite manipular los distintos tipos de delimitadores de caracteres.
- Una vez seleccionados los tipos de delimitadores con los que se separán los campos, se tendrá como resultado final un archivo listo para ser Importado o Cargado por cualquier Base de Datos.

5.2.3 Trabajando con DB2 UDB Universal Database

Para el desarrollo de esta tarea no hace falta estar al tanto por completo del *DB2 UDB Universal Database*, ya que en todo momento se utilizan asistentes. De todas maneras, para una mayor comprensión, se recomienda consultar el *Capitulo N°2 "Introduccíon al DB2 UDB Universal Database".*

Los pasos que se llevan a cabo son:

- Creación de la base de datos denominada *EPH (Encuesta Permanente Hogares).*
- Creación de la tabla *USP,* en la cual se realizarán la carga de archivos planos, exportados con *Microsoft Access.*
- Visualización del muestreo del contenido.

Creación de la Base de Datos

Seleccionar la opción Crear, utilizando el asistente haciendo click con el botón derecho sobre la carpeta bases de datos (ver fig. 5.1 de la pág. 74).

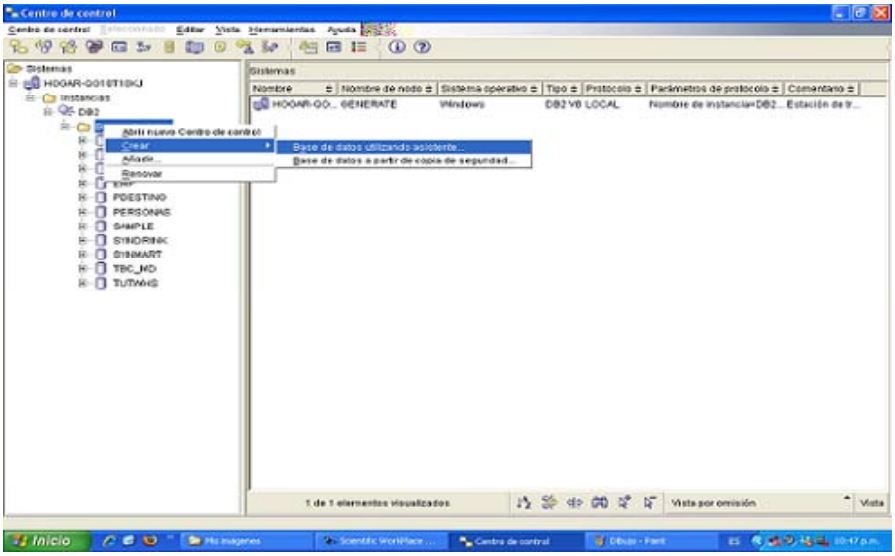

Figura 5.1: Creacíon de la base de datos utilizando el Asistente.

Una vez finalizada la creación de la base de datos, se pasa a la realización de las tablas. Para ello, se debe hacer click con el botón derecho sobre la carpeta *Tablas* de la base de datos *PERSONAS* y seleccionar la opcion *Crear* (ver fig. 5.2 de la pág. 75).

Luego se carga el Asistente, donde se tendrá que defenir los siguientes pasos:

- Especificar el Esquema y el Nombre de la nueva tabla (ver fig. 5.3 de la pág. 76).
- Cambiar las definiciónes para cada columna. Presionar el botón **Añadir** para ir insertando las columnas de la tabla; se deben elegir tipo y características de los datos, como también si estos alojan nulo (ver fig. 5.4 de la pág. 76).

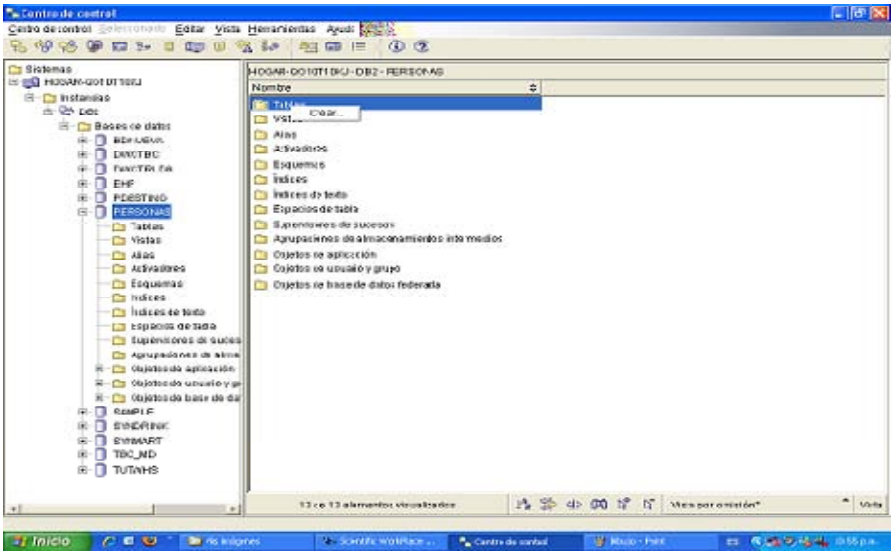

Figura 5.2: Selección de la opción crear tablas.

- Definir la clave para su utilización. Para ello se deberá presionar el botón **Añadir** primaria, luego el asistente mostrará las columnas candidatas, se seleccionará las adecuadas y presionar la opción >. Luego se presiona **Finalizar** (ver fig. 5.5 de la pág. 77).

5.2.4 Cargando Datos Fuentes a DB2 UDB Universal Database

Al hacer click con el botón derecho del mouse sobre la tabla *UTP_T107*, en la opción **Cargar**, se inicia al Asistente de Carga de Datos.

Como se visualiza en la fig. 5.6 de la pág. 78, en la pestaña **Especificar archivos de entrada y salida**, botón **Opciones DEL**→ Delimitador de Columna (COLDEL) se debe especificar el delimitador que utiliza el archivo plano, en este caso, el punto y coma(;).

Luego se debe especificar el archivo de entrada (*USP_T107.txt*), y el archivo para almacenar los mensajes de progreso (*mensajes.txt*). Es conveniente que estos archivos estén ubicados en el mismo disco donde se encuentra ins-

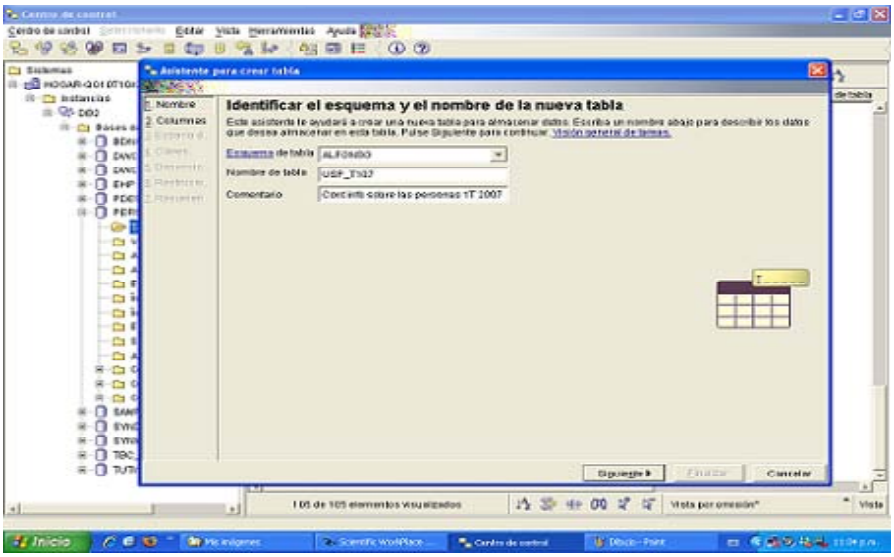

Figura 5.3: Identificación del esquema y del nombre de la nueva tabla.

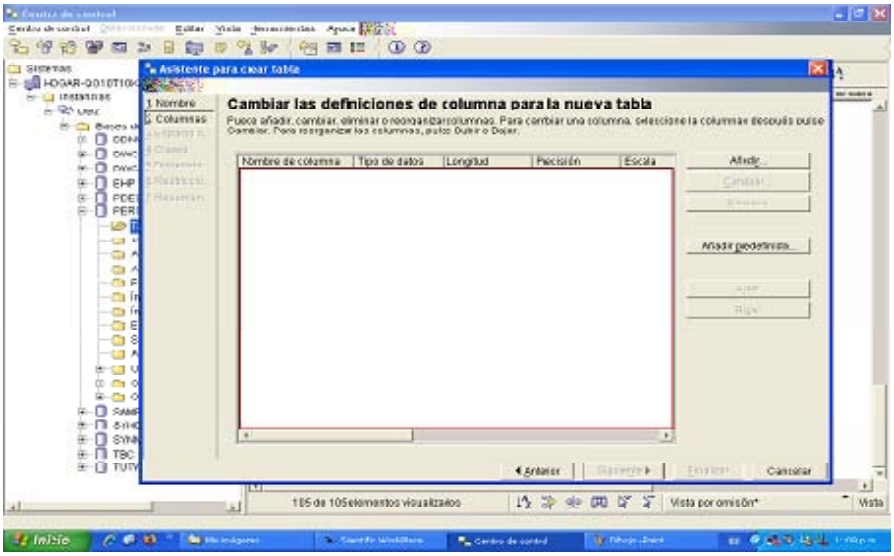

Figura 5.4: Cambiar las definiciones de columna para la nueva tabla.

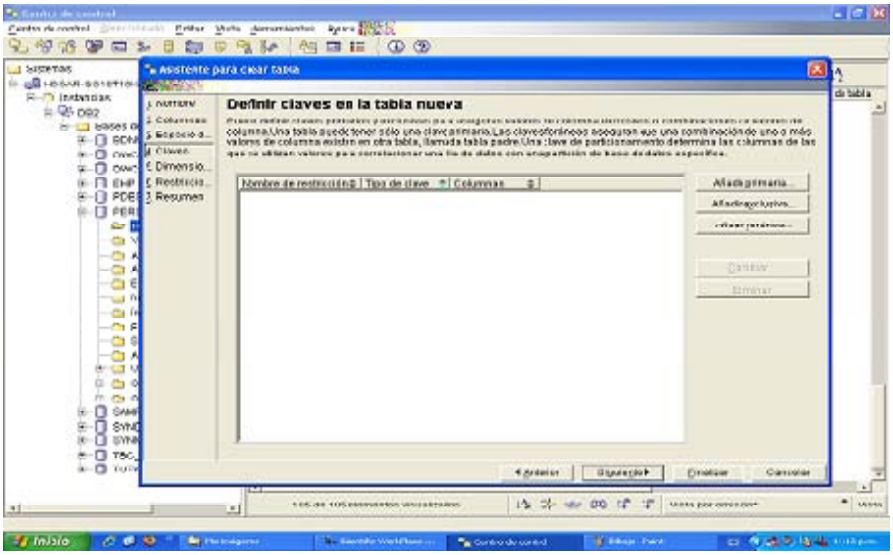

Figura 5.5: Definición de las claves primarias en la nueva tabla.

talado *DB2 UDB Universal Database*, dado que disminuyen los tiempos de carga.

Para obtener información acerca de los registros cargados se debe leer el archivo *(mensaje.txt)* especificado en el asistente de carga de datos, el cual contiene datos similares a los que se visualizan (ver fig. 5.7 de la pág. 78).

Al hacer click con el botón derecho sobre la tabla creada, opción *Muestreo del contenido*, se puede visualizar datos de la tabla *USH_ T107*, similar al que se puede observar en la fig. 5.8 de la pág. 79.

5.2.5 Comprensión de Datos

Luego de haber exportado los datos y controlado la correcta interpretación de los mismo por el *DB2 UDB Universal Database,* se observa que existe un total de *47.030 registros.*

Variables que contiene esta tabla:

- *Identificación.*

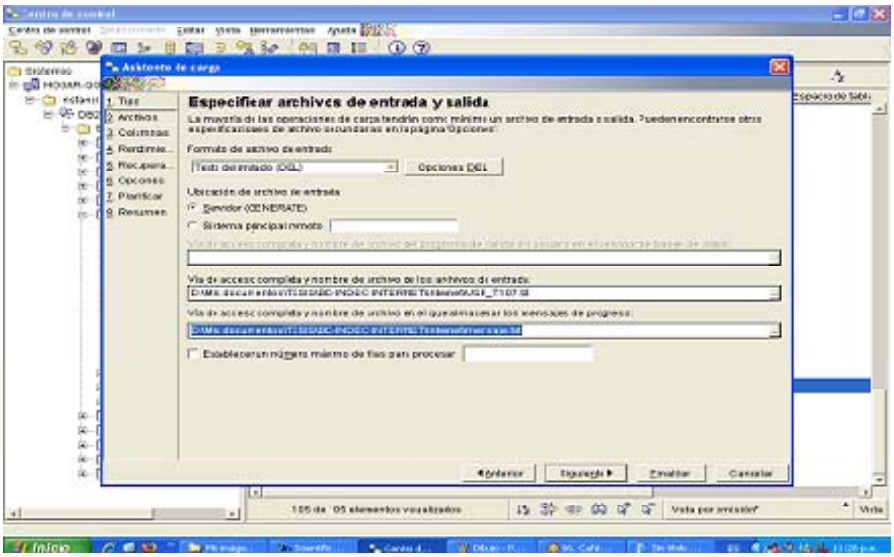

Figura 5.6: Especificación del archivo de Entrada/Salida en el asistente de carga de datos.

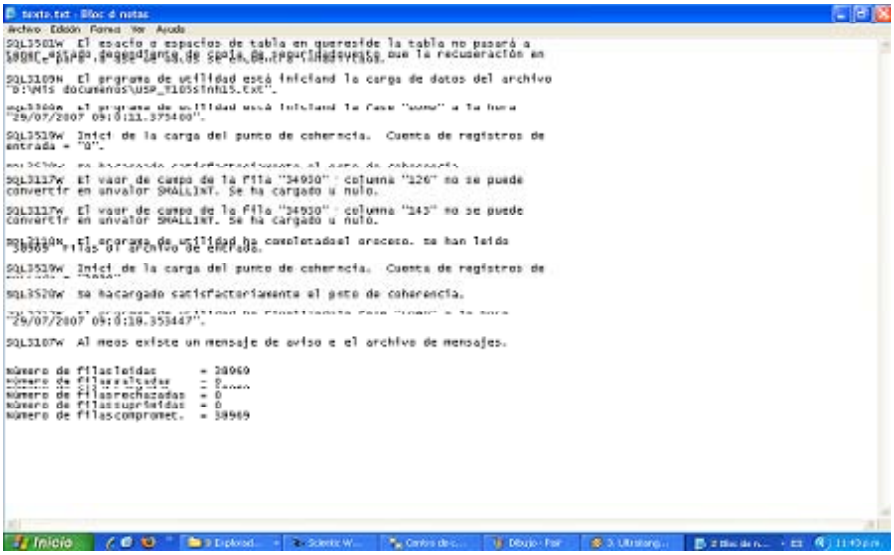

Figura 5.7: Muestreo del contenido del archivo de mensaje de progreso.

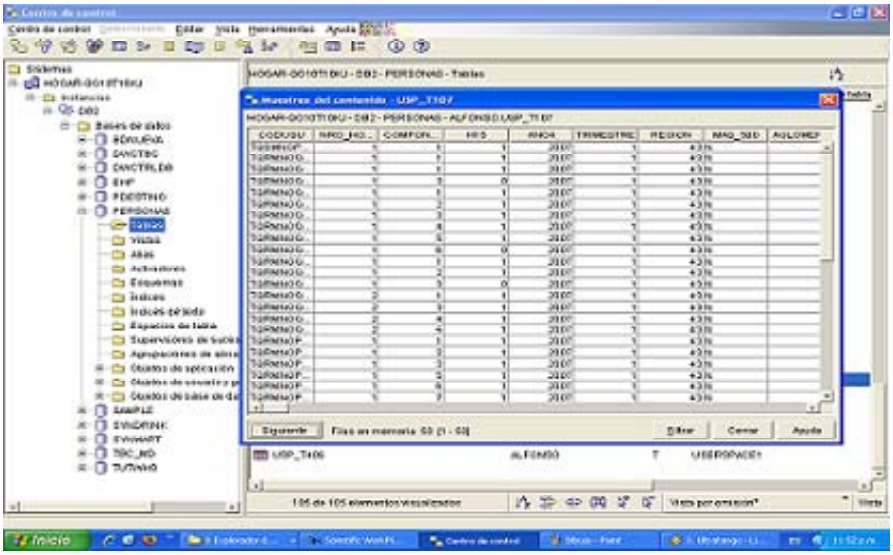

Figura 5.8: Muestreo del contenido de la tabla USP_T107 en el DB2.

- *Características de los miembros del hogar.*
- *Ocupados que trabajaron en la semana de referencia.*
- *Ocupados en general.*
- *Ocupación principal.*
- *Ocupación principal de los trabajadores independientes.*
- *Ingresos de la ocupación principal de los Trabajadores Independientes.*
- *Ocupación principal de los Asalariados (Excepto Servicio Doméstico).*
- *Ocupación principal de los Asalariados (Incluido Servicio Doméstico).*
- *Ingresos de la ocupación principal de los Asalariados.*
- *Movimientos Interurbanos (sólo para Ocupados).*
- *Desocupados.*
- *Desocupados con empleo anterior: (finalizada hace 3 años o menos).*
- *Ingresos de la ocupación principal.*

- *Ingresos de otras ocupaciones.*
- *Ingresos Total Individual.*
- *Ingresos No Laborales.*
- *Ingresos Total Familiar.*
- *Ingresos Per Cápita Familiar.*
- *Plan Jefas y Jefes de Hogar.*

Hasta aquí se ha finalizado la fase de la *Instalación del Ambiente Operacional*. Esta es de suma importancia ya que determina que las fases sucesivas sean capaces de extraer conocimientos válidos y útiles a partir de la información original.

Se deben observar si los datos con los que se cuenta son suficientes para hallar conocimiento, es decir, si son realmente útiles. Se entiende con el concepto de suficientes no el numero de registros, en cuanto a cantidad, si no la riqueza o importancia de los atributos a tener en cuenta.

Algunas veces, estos datos no pueden proveer la respuesta que se está buscando, por ello la importancia de prestar total atención a este punto.

Otro factor que es de suma importancia es el buen desarrollo del *Destino de Depósito*, lo que se verá a continuación.

5.3 Instalación del Ambiente Datamart

En esta fase se definirán todas las tablas correspondientes a las *dimensiones* y a la *tabla de hecho* de nuestro análisis del *Data Warehouse*.

Para mayor comprensión, se recomienda consultar *Capitulo Nº1 "Introducción a la Minería de Datos"*, precisamente la sección *Características del Data Warehouse*.

5.3.1 Selección y Exploración de la Destino de Depósito

Luego de un arduo estudio sobre la problemática hacia donde se enfoca la *EPH (Encuesta Permanente de Hogares)*, como así también la comprensión del alcance de las variables a considerar, se ha logrado determinar las siguientes *dimensiones*:

- **Nivel Educativo** (ver fig. 5.9 de la pág. 82).
- **Población de Asalariados** (ver fig. 5.10 de la pág. 83).
- **Población de Independientes** (ver fig. 5.11 de la pág. 84).
- **Población Desocupada** (ver fig. 5.15 de la pág. 87).
- **Población Desocupada c/Empleo anterior** (ver fig. 5.13 de la pág. 86).
- **Población c/Plan Jefes y Jefas de Hogar** (ver fig. 5.12 de la pág. 85).
- **Población Ocupados** (ver fig. 5.14 de la pág. 87).
- **Ocupación Principal** (ver fig. 5.16 de la pág. 88).

Siendo la tabla de *Hecho*:

- **Individuos**(ver fig. 5.17 de la pág. 88).

Formando así el esquema en estrella correspondiente (ver fig. 5.18 de la pág. 89).

Una vez definida todas la *dimensiones* se deberá exportar estas estructuras a el *DB2 UDB Universal Database.* Para llevar a cabo esto se debe trabajar de la misma forma que en el apartado anterior *"Trabajando con DB2 UDB Universal Database".*

Los pasos son:

- Creación de la base de datos denominada *PDESTINO.*
- Creación de una tabla por cada *dimensión.*

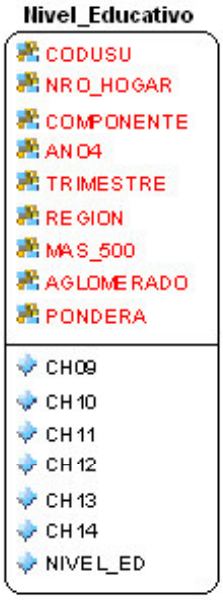

Figura 5.9: Visualización de la dimensión Nivel Educativo.

Figura 5.10: Visualización de la dimensión Población de Asalariados

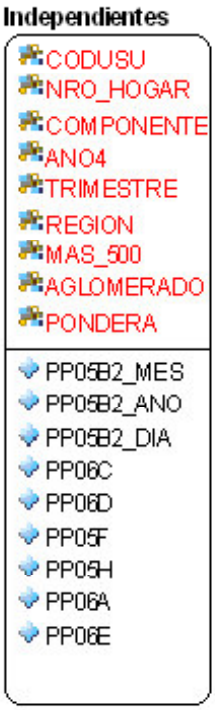

Figura 5.11: Visualización de la dimensión Independientes.

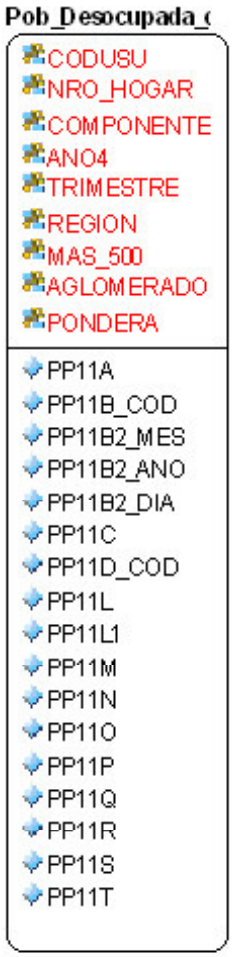

Figura 5.12: Visualización de la dimensión Población Desocupada con Empleo Anterior.

Figura 5.13: Visualización de la dimensión Población c/Plan Jefes y Jefas de Hogar.

Figura 5.14: Visualización de la dimensión Población Ocupados.

Figura 5.15: Visualización de la dimensión Población Desocupada.

Figura 5.16: Visualización de la dimensión Ocupación Principal.

Figura 5.17: Visualización de la dimensión Individuos (HECHO).

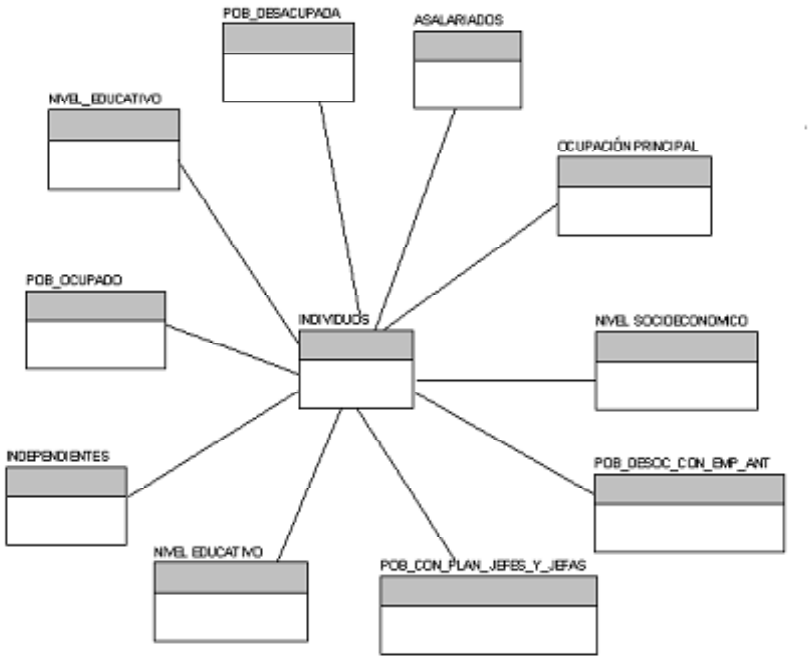

Figura 5.18: Visualización de la estructura del esquema en estrella.

- Creación de una tabla que corresponda a la tabla de *hecho.*

El primer paso utilizando el *DB2 UDB Universal Database* es creación de la base de datos. En este caso se llamará *PDESTINO,* hacia donde se exportarán todas las tablas de *dimensiones* junto con la de *hecho* (ver fig. 5.19 de la pág. 90).

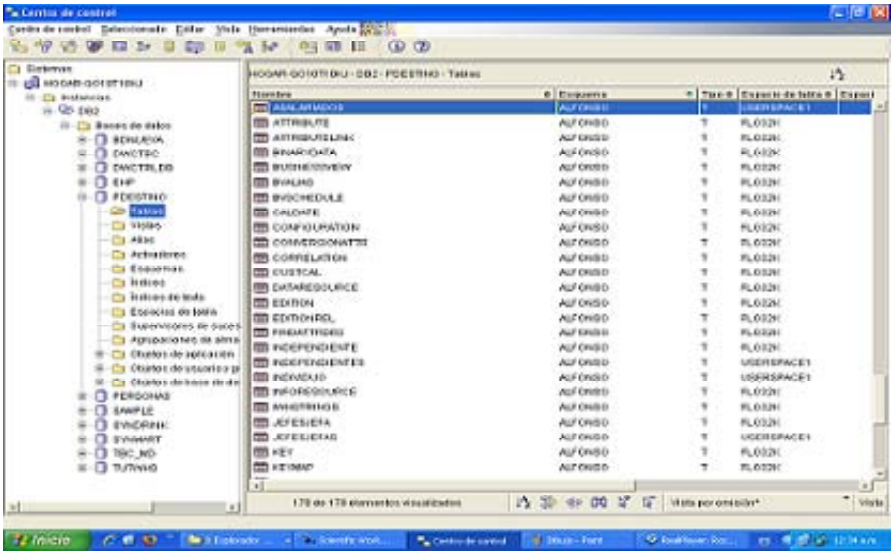

Figura 5.19: Creación de la base de datos denominada *PDESTINO.*

5.4 Introducción al Centro de Depósito de Datos

Una vez finalizado con el *Ambiente Operacional* y el *Ambiente Datamart* se pasa al

Centro de depósito de datos del *DB2 UDB Universal Database* (ver fig. 5.20 de la pág. 91) [8].

Cuando se escoge esa opción, el *DB2 UDB Universal Database* solicita que se ingrese un *ID* y la *Contraseña* del usuario, para que la base de datos pueda conectarse al *Centro de depósito de datos.* Luego se debe presionar el botón *Bien* (ver fig. 5.21 de la pág.91).

Figura 5.20: Visualización del icono Centro de depósito de datos.

A continuación aparece la ventana del *Centro de depósito de datos* (ver fig. 5.22 de la pág.92).

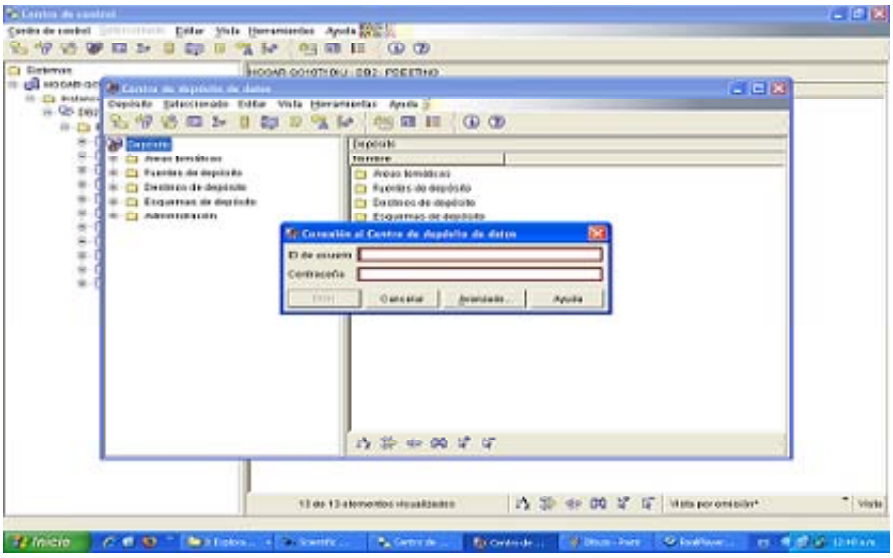

Figura 5.21: Iniciando la conexión al centro de depósito de datos.

El en ambiente de trabajo del *Centro de depósito de datos*, se deberá definir:

- *Áreas temáticas.*
- *Fuentes de depósitos.*
- *Destino de depósitos.*
- *Esquemas de depósitos.*
- *Administración.*

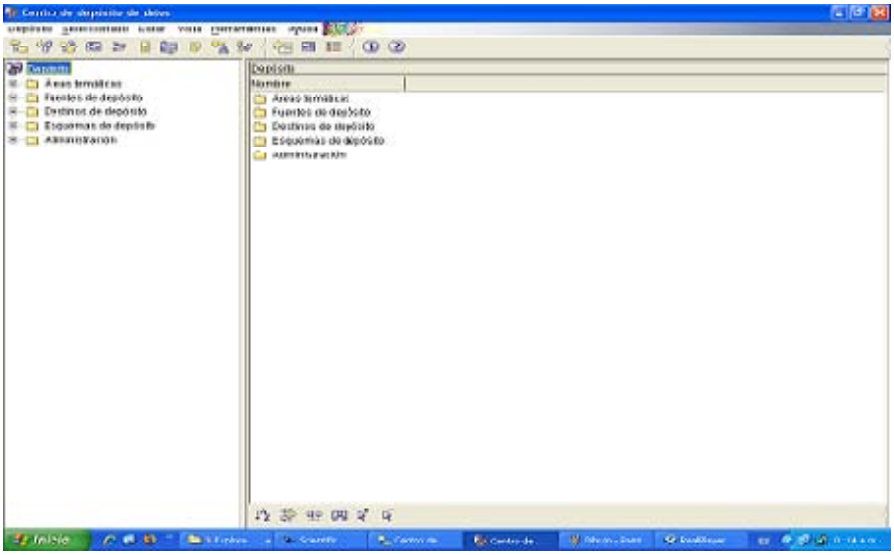

Figura 5.22: Visualización del Centro de depósito de datos.

5.4.1 Definición de una Área Temática

Un *Área temática* identifica y agrupa procesos relativos a un área lógica de la organización.

Por ejemplo, si se está creando un depósito de datos de *Marketing y Ventas*, se definirá una *Área temática* Ventas y otra Marketing. Luego se añadirán los procesos relativos a las ventas debajo del *Área temática Ventas*. Del mismo modo, se añadirán las definiciones relativas a los datos de Marketing debajo del *Área temática Marketing*.

Definición del Área Temática Encuesta Permanente de Hogares

En el árbol de la izquierda de la ventana del *Centro de depósito de datos* se debe seleccionar el nodo *Áreas temáticas* y luego pulsar *Definir*.

Se abrirá el cuaderno de *Propiedades del área temática* (ver fig. 5.23 de la pág.93).

Donde se cargarán los siguientes campos:

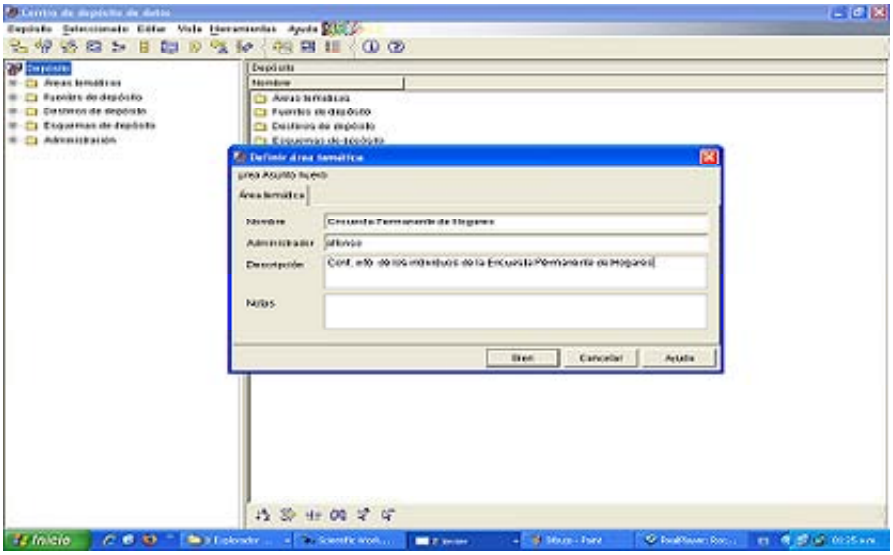

Figura 5.23: Definición del Area Temática (Encuesta Permanente de Hogares).

- **Nombre**, nombre comercial del área temática, para este caso: *Encuesta Permanente de Hogares.*
- **Descripción**, sinopsis del área temática: *Cont. información sobre los Individuos de la Encuesta Permanente de Hogares.*

También se puede utilizar el campo **Notas** para proporcionar información adicional sobre el área temática.

Pulsando en el botón *Bien*, se permitirá crear el área temática en el árbol del Centro de depósito de datos (ver fig. 5.24 de la pág.94).

5.4.2 Definición de las Fuentes de Depósito

El *Centro de depósito de datos* utiliza las especificaciones de las *Fuentes de Depósito* para acceder a los datos y seleccionarlos.

El *DB2 UDB Universal Database* permite que estas puedan ser:

- **Fuentes relacionales:** Correspondiente a la tabla fuente *USP_T107*

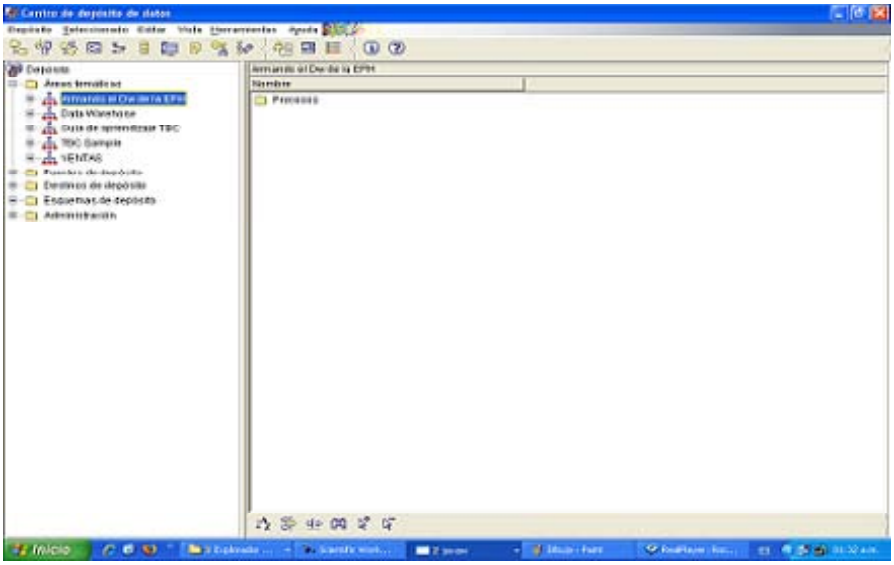

Figura 5.24: Definición culminada de área temática.

de la base de datos *Personas*.

- **Fuentes de archivos:** Correspondiente a los archivos de texto plano (*USP_ T107.txt)* sin delimitadores o delimitados por coma, (*USP_ T107.csv*).

Como se había mencionado en el apartado *Instalación del Ambiente Operacional*, si se están utilizando bases de datos fuente que son remotas para el servidor de depósito, se debe registrar las bases de datos en la estación de trabajo que contiene el servidor de depósito.

Definición de una Fuente de Depósito Relacional

Se definirá una *Fuente de Depósito Relacional* llamada *Fuente de Deposito Relacional de la EPH*. La nombrada anteriormente corresponde a la tabla relacional *USP_ T107* que se proporciona en la base de datos *Personas*.

Procedimientos:

Se debe pulsar con el botón derecho sobre la carpeta *Fuentes de depósito* y selecionar *Definir Familia de DB2*. Luego se abrirá el cuaderno *Definir fuente*

de depósito (ver fig. 5.30 de la pág.99).

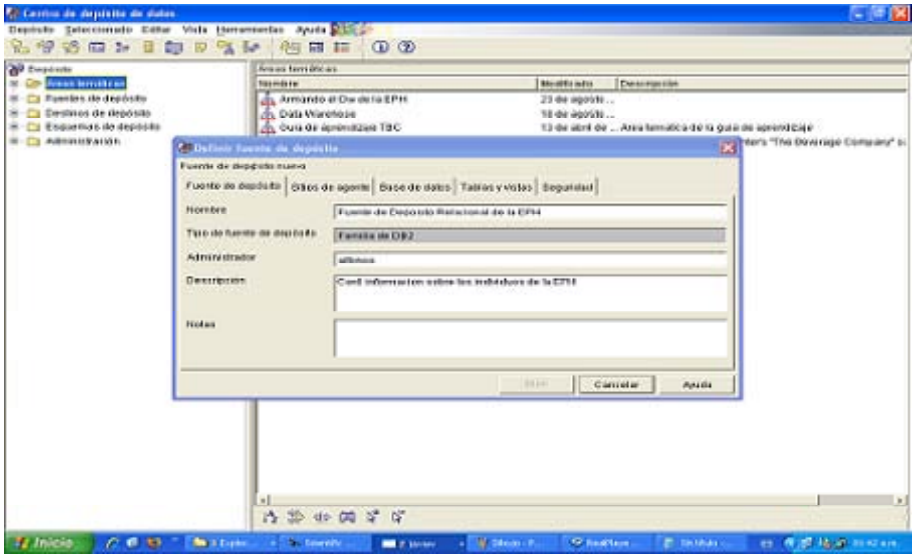

Figura 5.25: Definicion de la fuente de depósito (Fuente de Deposito Relacional de la EPH).

Donde se introduce la siguiente información en los campos:

- **Nombre** de fuente de depósito, se escribe el nombre comercial de la fuente de depósito: *Fuente de Deposito Relacional de la EPH.* Se utilizará este nombre para hacer referencia a la fuente del depósito en el *Centro de depósito de datos.*
- **Administrador**, se escribe un nombre de contacto para la fuente de depósito: *alfonso.*
- **Descripción**, se introduce una breve descripción de los datos: *Cont. informacion en Tablas Relacionales sobre los Individuos de la EPH.* Luego se realiza una pulsación sobre la pestaña *Base de datos.*

Luego se introduce la siguiente información en los campos:

- **Nombre de la base de datos**: *Personas* (base de datos física).

- **ID de usuario**: *alfonso* (id de acceso a la base de datos).

- **Contraseña**: clave de acceso correspondiente al *ID de usuario* que accederá a la base de datos (ver fig. 5.26 de la pág.96).

Se utilizará el *ID de usuario* y la *Contraseña* que se especificó al crear la base de datos de ejemplo en el apartado *"Introduccíon a el Centro de depósito de datos"*.

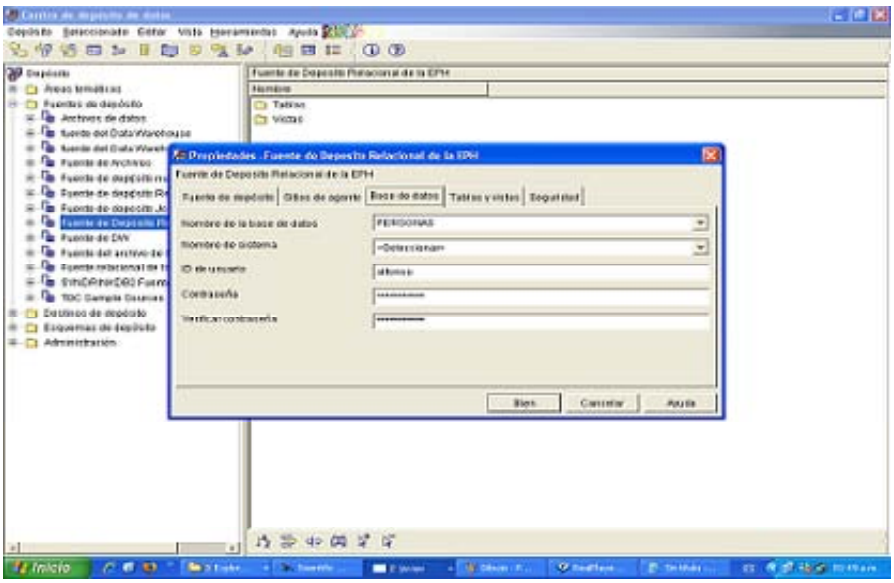

Figura 5.26: Selección de la base de datos para la Fuente de depósito.

El *Centro de depósito de datos* muestra una ventana de progreso. La importación puede tardar unos minutos.

Después de finalizada la importación, el *Centro de depósito de datos* visualiza los objetos importados en el árbol de objetos, Tablas y Vistas disponibles.

Luego se selecciona la tabla *USP_T105*; pulsando > (ver fig. 5.27 de la pág.97).

De esta manera la tabla *USP_T105* se traslada a la lista *Tablas y vistas seleccionadas.* (ver fig. 5.28 de la pág.97). Luego se pulsa el botón **Bien**.

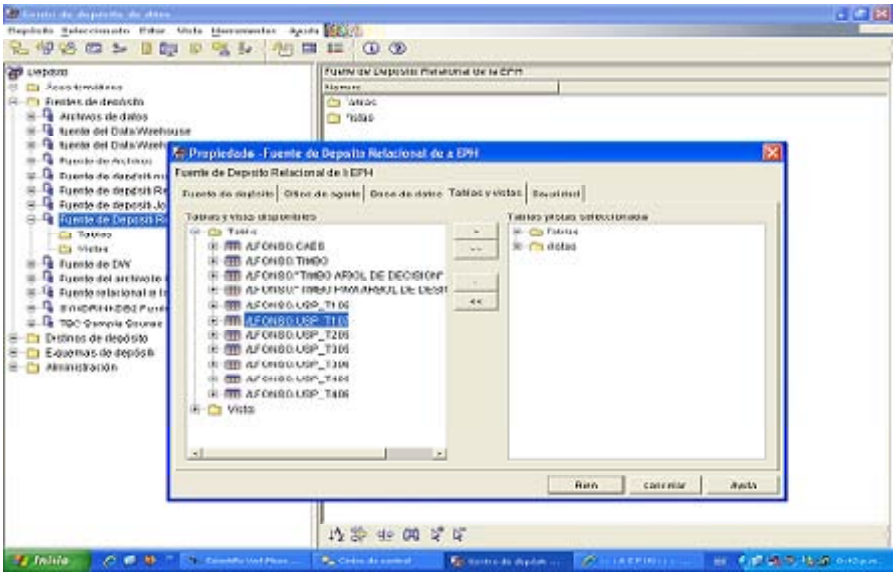

Figura 5.27: Visualización de las Tablas y vistas disponibles.

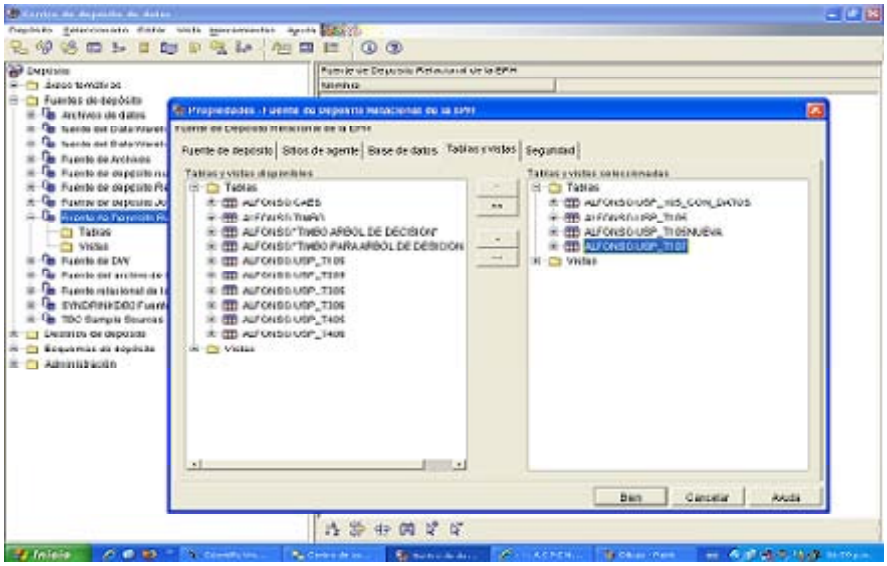

Figura 5.28: Visualizacion de las Tablas y vistas seleccionadas.

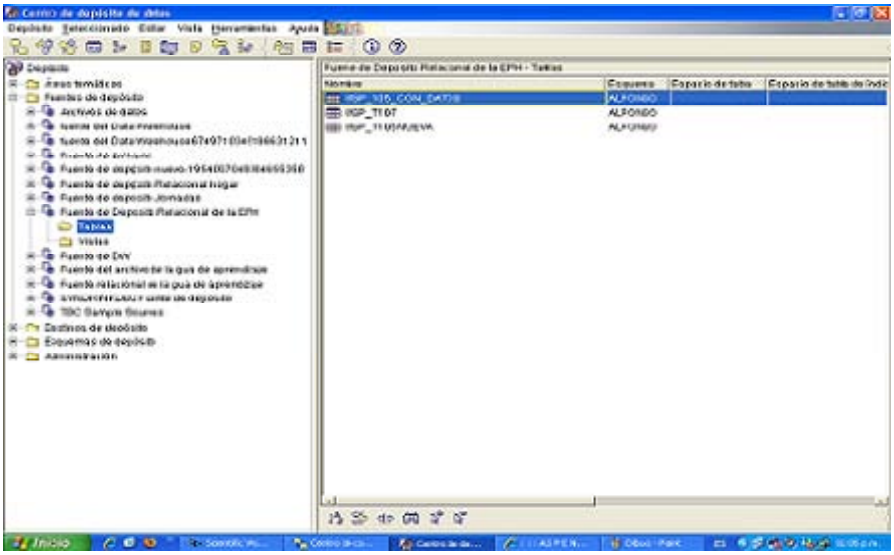

Figura 5.29: Visualización de las Tablas de depósito cargadas a el Centro de depósito de datos.

Se guardarán los cambios y se cerrará el cuaderno *Definir fuentes de depósito* (ver fig. 5.29 de la pág.98).

Debe repetir este proceso hasta que renombre el resto de las columnas de la tabla *USP_T107.*

Luego se deberá pulsar **Bien**. Y se cerrará el *Cuaderno Archivo.*

5.4.3 Definición de Destinos de Depósito

Los *Destinos de depósito* identifican la base de datos y las tablas que el *Centro de depósito de datos* debe utilizar para el depósito. Normalmente, las tablas de destino que se definen en el destino de depósito son las tablas de mediciones y de hechos del esquema en estrella. Sin embargo, el destino de depósito puede incluir también tablas de destino intermedias que se utilizan para la transformación de datos.

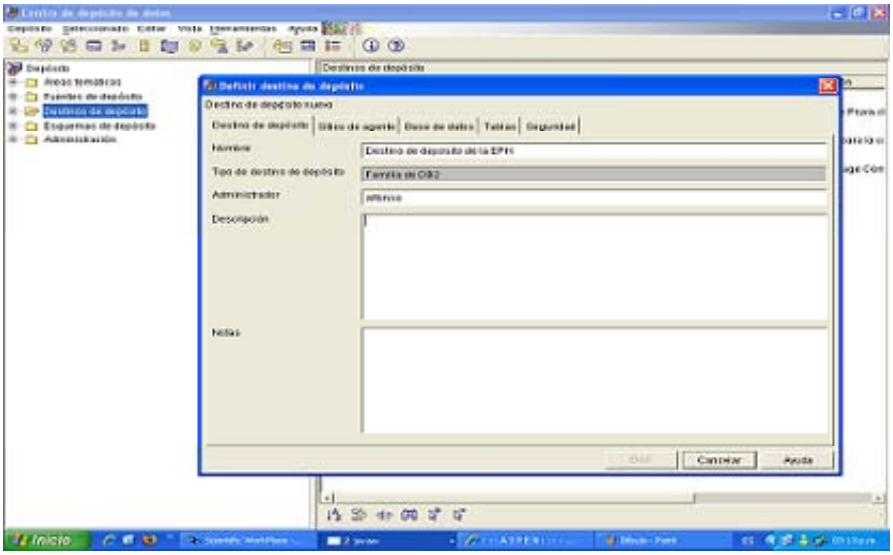

Figura 5.30: Visualización del Cuaderno Destino de depósito.

Definición de un Destino de Depósito

Para definir el *Destino de depósito:*

Se efectúa una pulsación con el botón derecho del ratón en la carpeta **Destinos de depósito,** luego en la opción **Definir** y se desplegará la lista **Tipo de destino de depósito**, se debe seleccionar allí la opción *DB2 UDB Universal Database.* De esta manera se abrirá el *cuaderno Destino de depósito* (ver fig. 5.30 de la pág.99).

A continuación se deberán llenar los correspondientes campos:

- **Nombre:** nombre comercial del destino de depósito: *Destino de deposito de la EPH.*
- **Administrador:** contacto para el destino de depósito.
- **Descripción**: sinopsis de los datos: *Contiene todas las tablas de destino de Depósito.*

Pulzando sobre la pestaña base de datos, se debe llenar los correspondientes

campos:

- **Nombre de base de datos**: *Ppdestino.*
- **Id de usuario**: identificación de acceso a la base de datos.
- **Contraseña**: clave correspondiente al id de usuario.
- **Verificar contraseña**: Repetir la clave.
- Luego aceptar los valores por omisión para el resto de los controles de la página.

Desplegar la pestaña *Tablas* en el cuaderno *Destino de depósito.*

Luego expandir el árbol hasta encontrar la carpeta Tablas y seleccionar todas:

- *Asalariados.*
- *Independientes.*
- *Individuo.*
- *Nivel _ educativo.*
- *Ocupación _ principal.*
- *Pob_ con_ Plan_ Jefes_ y_ Jefas.*
- *Pob_ Desocupada.*
- *Pob_ Desocupada_ con_ empleo_ Anterior.*
- *Pob_ Ocupado.*

Luego pulsar el botón >, y aparecerán listadas en el panel Tablas seleccionadas (ver fig. 5.31 de la pág.101).

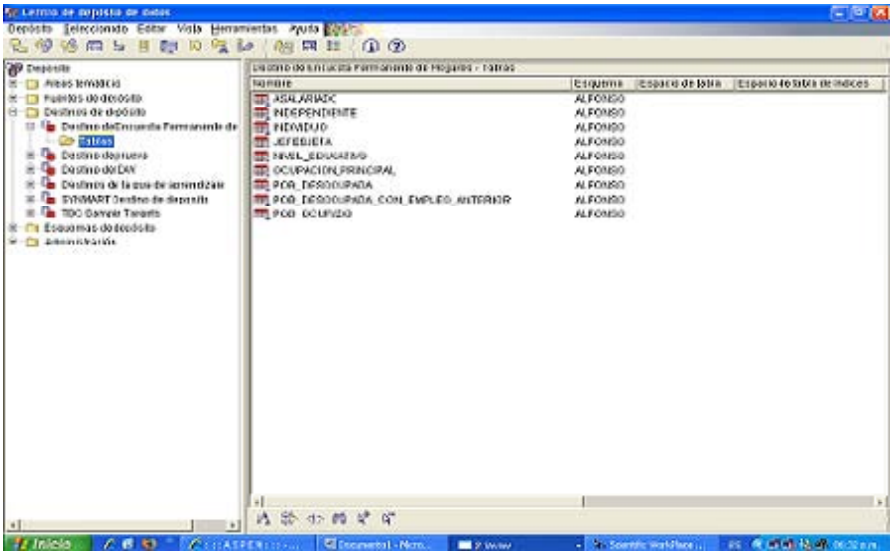

Figura 5.31: Visualización de las Tablas disponibles del cuaderno Destino de depósito.

5.4.4 Definición del Movimiento y Transformación de Datos

En esta sección, se definirá el modo en que el *centro de depósito de datos* debe mover y transformar datos para darles el formato adecuado al depósito de datos. En primer lugar, se definirá un proceso que contenga una serie de pasos que permitan la transformación y movimiento de los mismos. A continuación, se especificarán las tablas fuente que deben transformarse para el depósito. Finalmente, se definirán los pasos de transformación de datos de la siguiente manera:

- Se seleccionan los datos fuente que se unirán a las tablas a través de sentencias de *SQL*. Específicamente, se definirá el proceso llamado *Introducir datos en el DW*, que realiza lo siguiente:

  - Se seleccionan los datos de la tabla *USP_T105NUEVA* y se los transfiere a sus respectivas tablas de destino, logrando el llenado total de las tablas.

Definición de un Proceso

Para la definicíon del objeto del proceso se debe realizar lo siguente:

- Desde la ventana del *centro de depósito de datos*, se expande el árbol *Áreas temáticas* .
- Se selecciona el *Área temática Armando el DW de la EPH*, que se ha definido con anterioridad.
- Se efectúa una pulsación con el botón derecho del ratón en la carpeta *Procesos* y luego se pulsa Definir.
- Se abrirá el cuaderno *Definir proceso*:
  - **Nombre**: se escribe el nombre del proceso: *Introducir datos en el DW.* El mismo puede tener un máximo de 80 caracteres de longitud y es sensible a mayúsculas y minúsculas. El primer carácter del nombre debe ser alfanumérico. No puede utilizar un & como primer carácter. Este campo es obligatorio.
  - **Administrador**: se escribe un nombre de contacto para la definición del proceso.
  - **Descripción**: se escribe la descripción del proceso: este es un proceso que permitirá transportar los datos desde unas entidades fuentes a las entidades depósitos de datos (ver fig. 5.32 de la pág.103).

Luego se deberá pulsar la pestaña *Seguridad.*

En la lista Grupos de seguridad disponibles, se selecciona el *Grupo de depósito de la guía de aprendizaje* pulsando el botón >.

El *Grupo de depósito de la guía de aprendizaje* se visualiza en el panel de Grupos de seguridad seleccionados. Pulsando el botón **Bien**.

Se cerrará el *cuaderno definir proceso.*

Apertura del Proceso

Se abrirá el proceso de modo que se pueda definir gráficamente el flujo de datos del mismo.

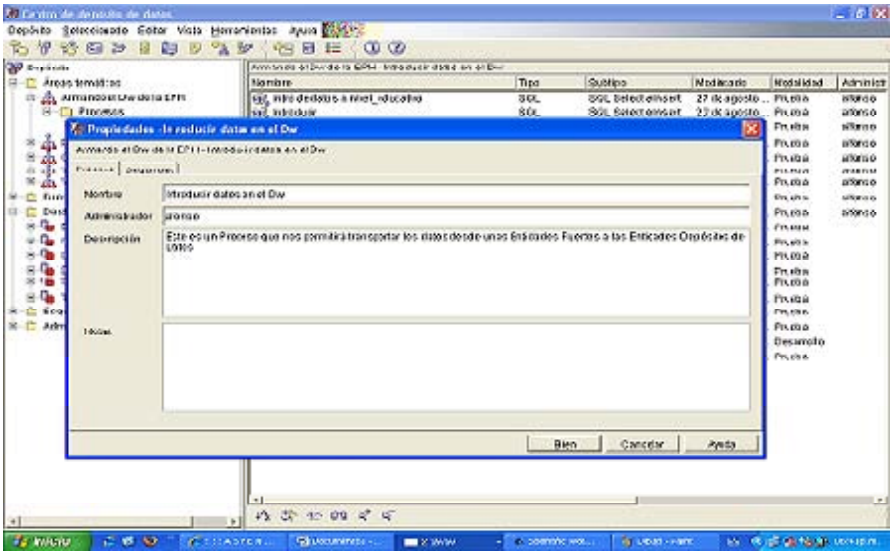

Figura 5.32: Visualización de las propiedades del cuaderno definir proceso.

Para abrir el proceso en este caso, se debe *Introducir datos en el DW*:

- Efectuar una pulsación con el botón derecho del ratón en el proceso *Introducir datos en el DW*.
- Pulsar el botón Abrir, y se abrira el *Modelador de proceso* (ver fig. 5.33 de la pág.104).

Adición de Tablas a un Proceso

Para definir el flujo de datos, es necesario unir cada fuente, transformadas previamente, con las tablas de destino resultantes.

En el proceso *Introducir datos en el DW*, se cargarán los datos de la *Encuesta Permanente de Hogares EPH*, precisamente del primer trimestre del 2005, por lo que es necesario unir la tabla fuente *USP_T105NUEVA* con las tablas de destinos *(Asalariados, Independientes, Individuo,Nivel _ educativo, Ocupación _ principal, Pob_con_Plan_Jefes_y_Jefas, Pob_Desocupada, Pob _Desocupada_con_empleo_Anterior, Pob_Ocupado)*.

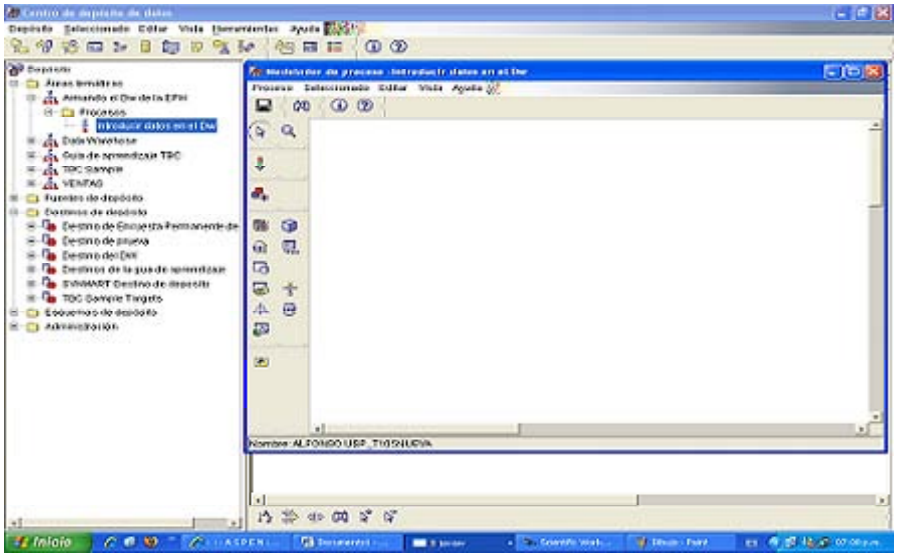

Figura 5.33: Visualización del Modelador de Proceso.

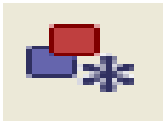

Figura 5.34: Visualización del icono añadir datos.

La tabla *USP_ T105NUEVA* forma parte de la fuente de depósito, que se ha definido en el apartado *Definición de una Fuente de Depósito Relacional* y las definiciónes de las tablas destino de depósito en el apartado *Definición de un Destino de Depósito*.

Para añadir una tabla fuente al proceso se debe realizar lo siguiente:

- Pulsar el icono *Añadir datos* (ver fig. 5.34 de la pág.104).
- Pulsar sobre la cuadrícula en el punto donde se desea colocar la tabla. Se abrirá la ventana *Añadir datos*.
- En la lista Tablas fuente y destino disponibles, expandir el árbol *Fuentes de depósito*. Se visualizará una lista de las *Fuentes de depósito* definidas

en el depósito (ver fig. 5.35 de la pág.105).

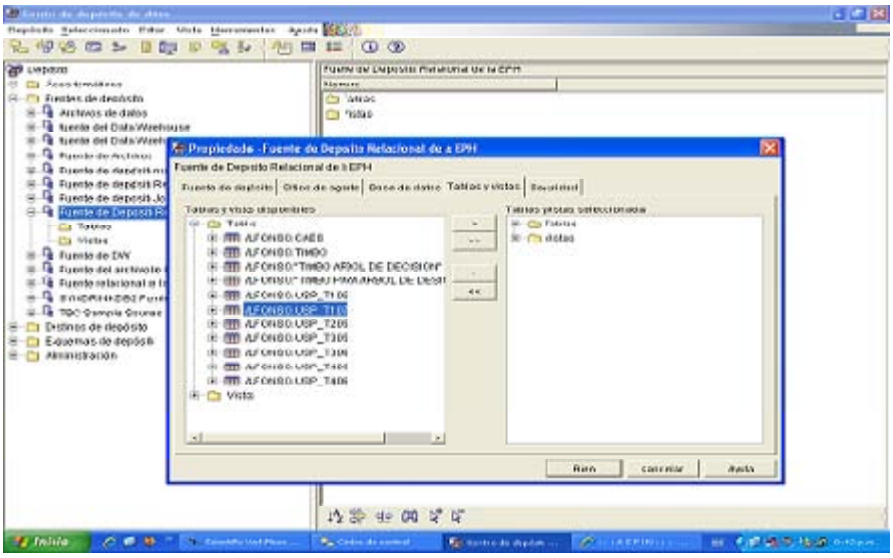

Figura 5.35: Visualizacion de las Tablas fuente disponibles y seleccionadas.

- Expandir el árbol para la fuente de depósito *Fuente de Deposito Relacional de la EPH.*
- Seleccionar la tabla *USP_ T105NUEVA.*
- Pulsar el botón > para añadir la tabla *USP_ T105NUEVA* a la lista Tablas fuente y destino seleccionadas.

Para añadir la tabla de destino:

- Pulsar el icono *Añadir datos* (ver fig. 5.34 de la pág.104):
- Pulsar sobre la cuadrícula en el punto donde se desea colocar la tabla. Se abrirá la ventana Añadir datos.
- En la lista *Tablas fuente y destino disponibles*, expandir el árbol *Destinos de depósito.* Se visualizará una lista de los destinos de depósito definidos en el depósito.

- Desplegar el árbol de destino de depósito *Destino de Encuesta Permanente de Hogares.*
- Desplegar el árbol Tablas. Deberá verse en la lista *tablas de fuente y de destino disponibles* :
  - *Asalariados*
  - *Independientes*
  - *Individuo*
  - *Nivel _ educativo*
  - *Ocupación _ principal*
  - *Pob_ con_ Plan_ Jefes_ y_ Jefas*
  - *Pob_ Desocupada*
  - *Pob_ Desocupada_ con_ empleo_ Anterior*
  - *Pob_ Ocupado* (ver fig. 5.36 de la pág.106).

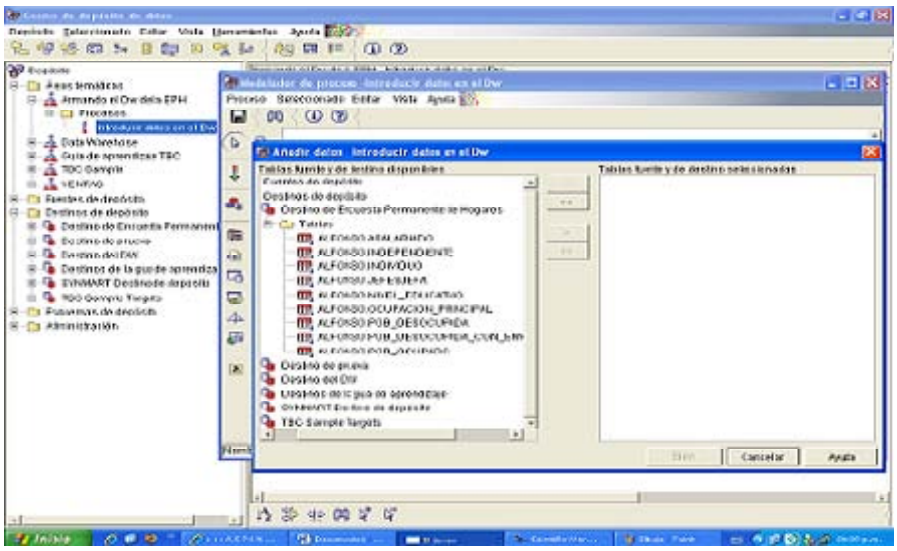

Figura 5.36: Visualización de las tablas de Destino de Depósito.

- Seleccionar la tabla de destino *Nivel _ educativo.*
- Pulsar > para añadir la tabla de destino *Nivel _ educativo* a la lista *Tablas fuente y destino seleccionadas.*

Adición de Pasos al Proceso

Ahora, es necesario añadir los pasos que definen cómo deben transformarse los datos fuente *(Fuente de Destino)* en datos de destino *(Destino de Depósito)*.

En el siguiente apartado, se definirán los pasos *SQL Select e Insert* que permitirán la tranformacion de datos.

Definición del Paso *intro de datos a nivel_ educativo*:

- Desde la paleta, pulsar el icono *SQL* (ver fig. 5.37 de la pág.107).

Figura 5.37: Visualización del icono introducir SQL.

- Pulsar en el punto de la cuadrícula donde desee colocar el *Paso*. Se añadirá a la ventana un ícono para el *Paso*.
- Hacer clic con el botón derecho del ratón sobre el ícono *Paso,* creado previamente.
- Pulsar el botón Propiedades, de esta menera se abrirá el *cuaderno Paso* (ver fig. 5.38 de la pág.108). Luego completar los siguientes datos:
  - **Nombre**: nombre del paso: *intro de datos a nivel_ educativo.*
  - **Administrador**: nombre de contacto para el paso.
  - **Descripción**: breve información indicando lo que realiza dicho paso.
- Luego pulsar el botón **Bien**. Se cerrará el cuaderno *Paso*.
- Pulsar el ícono *Herramientas de enlace* (ver fig. 5.39 de la pág.108).

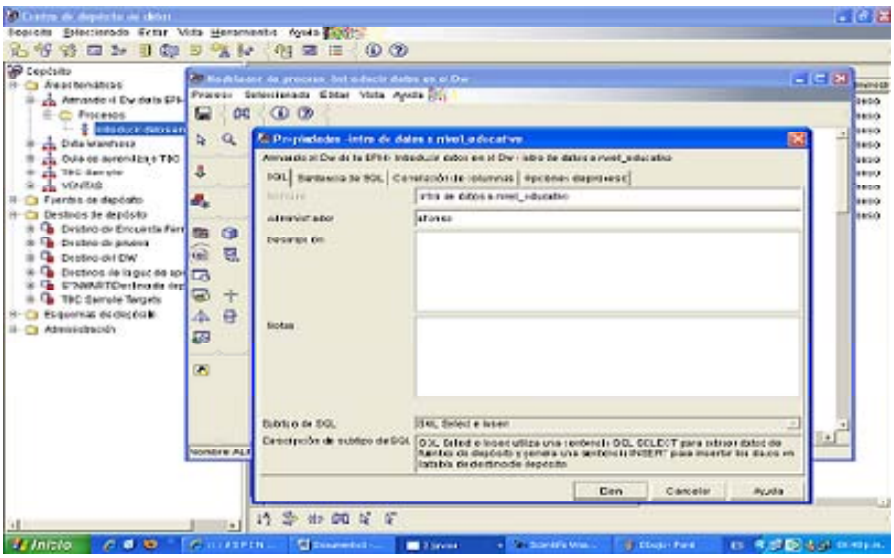

Figura 5.38: Visualización de las propiedades del paso *intro de datos a nivel_educativo*.

Figura 5.39: Visualización del icono Flujo de Datos.

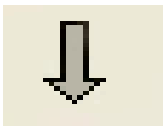

Figura 5.40: Visualización del icono Enlaces de datos.

- Luego pulsar el ícono *Enlace de datos,* (ver fig. 5.40 de la pág.108).

- Pulsar en el medio de la tabla fuente de *USP_ T105NUEVA* y arrastrar el ratón hasta el medio del paso *intro de datos a nivel_ educativo.* El centro de depósito de datos traza una línea que indica que la tabla fuente de *USP_ T105 NUEVA* contiene los datos fuente para el paso.
- Pulsar en el medio del paso *intro de datos a nivel_ educativo* hasta la tabla destino de depósito *NIVEL_ EDUCATIVO.*
- Una vez enlazada una tabla de destino con el paso, el centro de depósito de datos traza una línea que indica que los resultados de la consulta del paso se alojarán en la tabla de destino de depósito (ver fig. 5.41 de la pág.109).

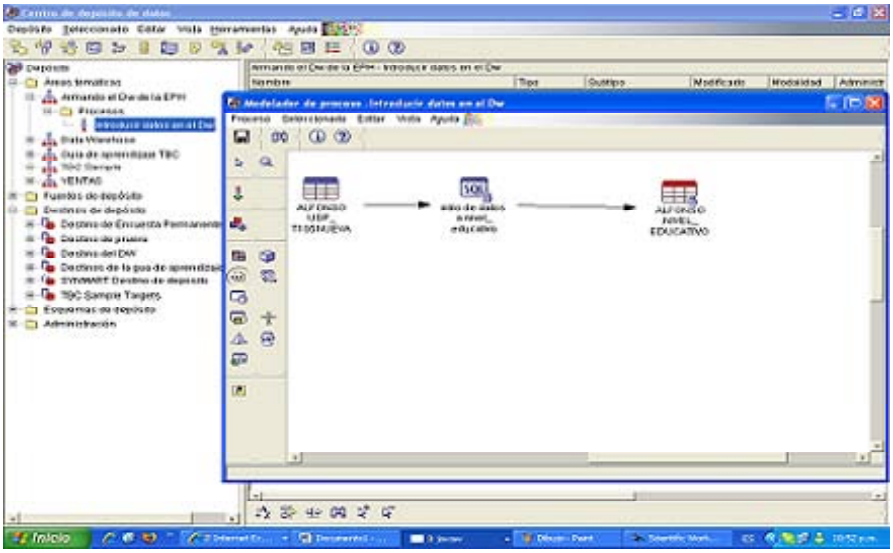

Figura 5.41: Visualizacion del esquema del paso, Introducir datos en el DW.

- Pulsar con el botón derecho del ratón en el paso *intro de datos a nivel_ educativo.*
- Pulsar la opción Propiedades y se abrirá el cuaderno *Paso.*

- Luego pulsar la pestaña *Sentencia de SQL* .
- Pulsar *Crear SQL* para que el Centro de depósito de datos genere automáticamente el código SQL correspondiente (de lo contrario, se podrá escribir la sentencia directamente).
- Se abrirá el *Asistente de SQL.*
- Se seleccionará la pestaña Tablas, se debe verificar si la tabla *USP_ T105NUEVA* está seleccionada.
- Pulsar la pestaña Columnas (ver fig. 5.42 de la pág.110).

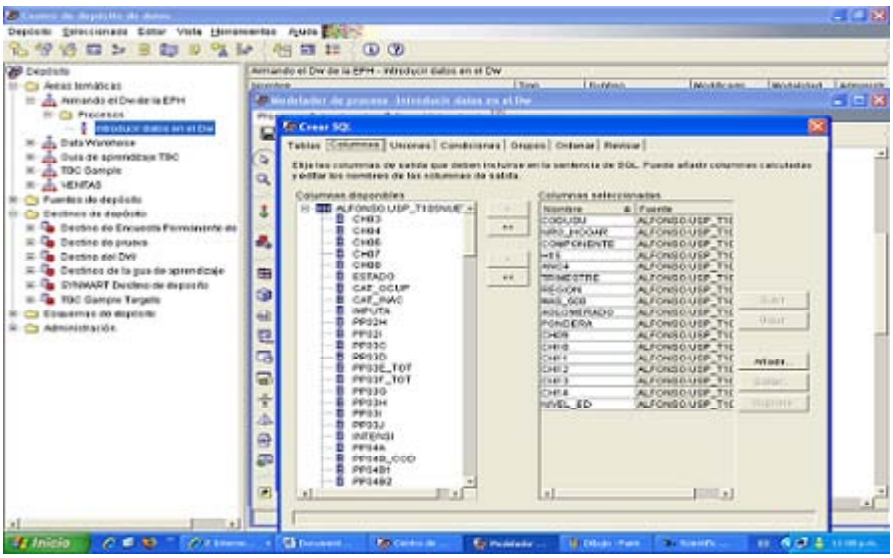

Figura 5.42: Selección de las columnas que deben unirse en la sentencia de SQL.

- Pulsar el botón >> para añadir todas las columnas de la tabla *USP _T105NUEVA.*
- Luego seleccionar la pestaña Revisar, de esta manera se podrá visualizar la consulta *SQL* (ver fig. 5.43 de la pág.111).

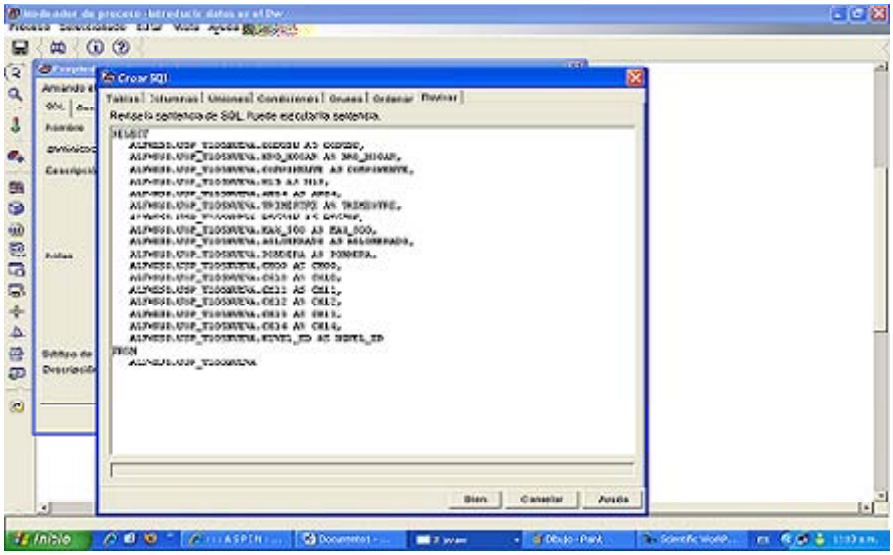

Figura 5.43: Visualización de la sentencia de SQL, con los campos antes seleccionados.

- Al pulsar el botón Bien, se cerrará el *Asistente SQL*.
- Seleccionar la pestaña *Correlación de columnas*, verificando si las columnas fuente se correlacionan con las columnas de destino correctas (ver fig. 5.44 de la pág.112).
- Luego se debe precionar el botón Acciones y la opción correlacionar todos por posición. De esta forma se podrá correlacionar las columnas de ambas tablas (ver fig. 5.45 de la pág.112).
- Seleccionar la pestaña Opciones de Proceso y luego en Agregar en la opción Tipo de llenado.
- Para finalizar se debe seleccionar el botón **Bien**.

Prueba de los Pasos

En este apartado se ejecutará el paso *intro de datos a nivel_ educativo* de manera que se pueda comprobar la correcta definición de los procedimientos realizados anteriormente.

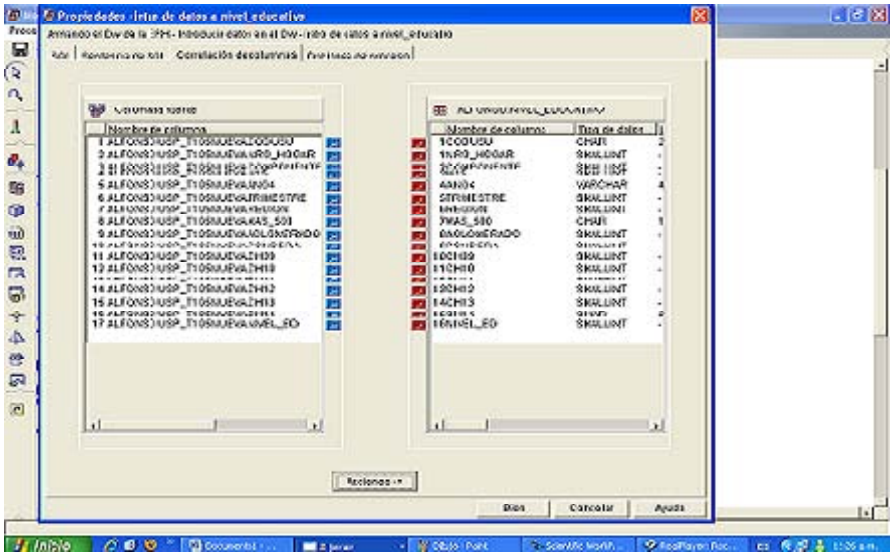

Figura 5.44: Visualización de las columnas fuente que se debe correlacionar con las columnas de destino.

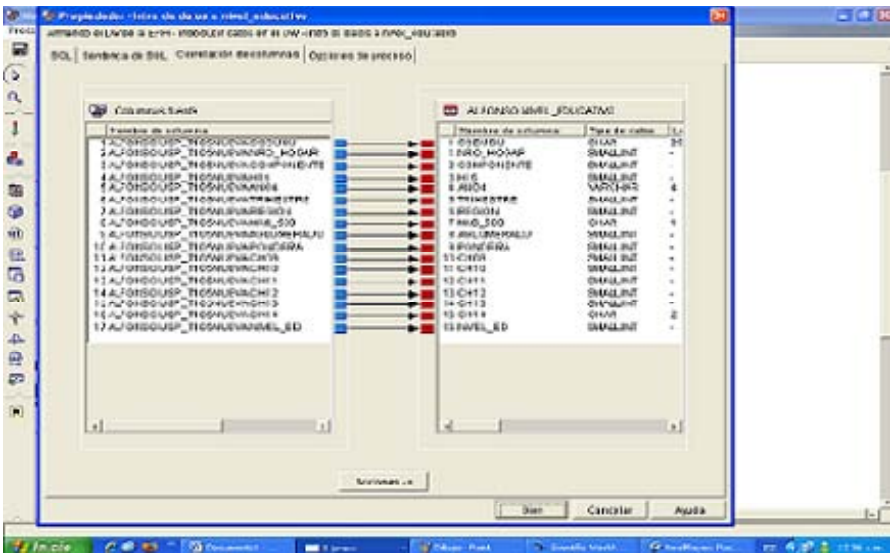

Figura 5.45: Visualización de la acción correlación por posición.

Para promocionar el paso *intro de datos a nivel_ educativo*:

- Desde la ventana Modelo de proceso correspondiente al proceso *Introducir datos en el DW* efectuar una pulsación con el botón derecho sobre el paso *intro de datos a nivel_ educativo.*
- Pulsar en la opción Modalidad y luego en Prueba (ver fig. 5.46 de la pág.113).

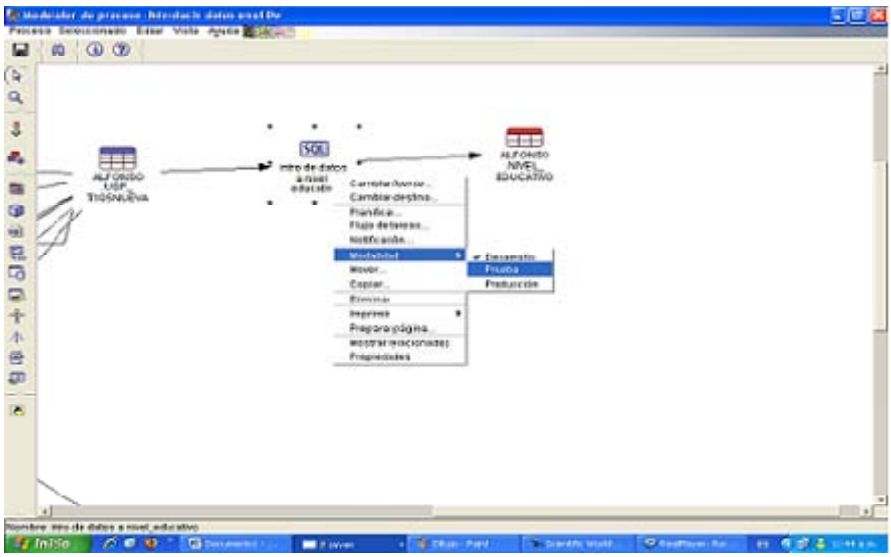

Figura 5.46: Visualización del cambio de Modalidad Desarrollo a la de Producción.

Una ventana de confirmación le solicitará si desea guardar el proceso.

Pulsar Sí o de lo contrario guardar los cambios pulsando en el ícono con forma de diskette de la barra de herramientras (ver fig. 5.47 de la pág.114). Luego se iniciará el centro de depósito de datos para crear la tabla de destino mostrando una ventana de progreso. Antes de iniciar el procedimiento siguiente, esperar a que el centro de depósito de datos finalice el proceso, esto puede tomar varios minutos. Una vez que el centro de depósito de datos finaliza, se visualiza un candado de seguridad indicando que no se podrán realizar modificaciones en el futuro (ver fig. 5.48 de la pág.114).

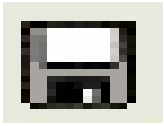

Figura 5.47: Visualización del icono Diskette.

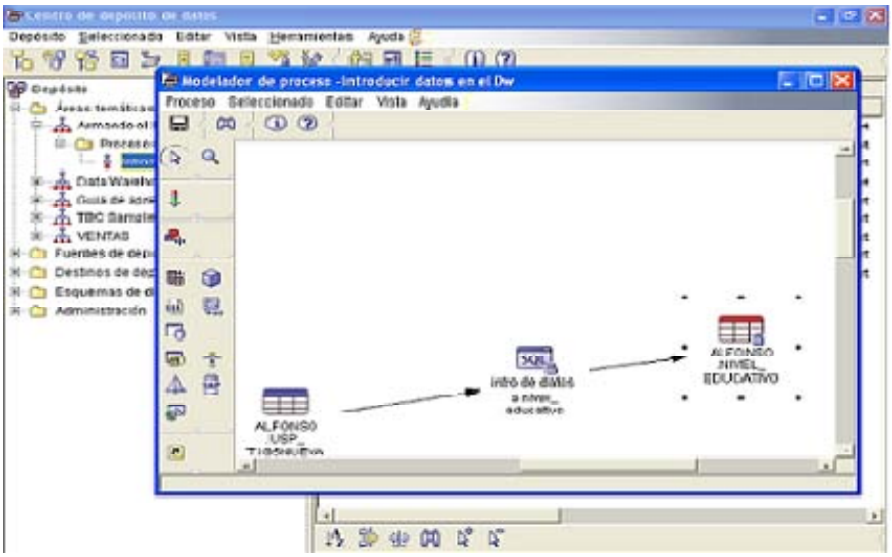

Figura 5.48: Visualizacion del Modelador de Proceso, que se encuentra bloqueado.

- Luego se debe selecionar la opción **Prueba**.

El centro de depósito de datos muestra una ventana de progreso una vez que finalizado.

- Pulsar con el botón derecho del ratón y escojer la opción muestreo de contenido en la tabla destino de depósito *NIVEL_EDUCATIVO* (ver fig. 5.49 de la pág.115).

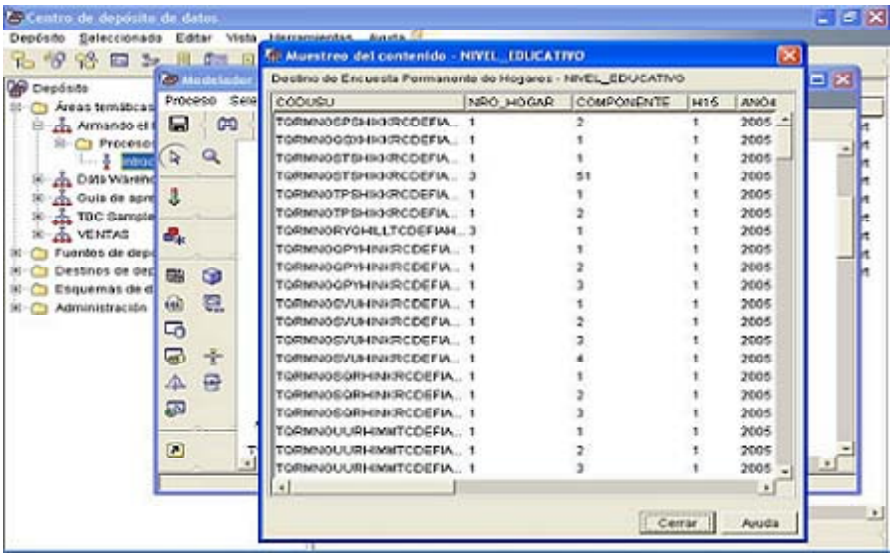

Figura 5.49: Visualización del contenido de la tabla destino de depósito *NIVEL_EDUCATIVO*.

5.4.5 Definición de Claves de Tablas de Destino de Depósito

En esta sección se definirán las *claves principales* y *foráneas* de tablas de destino para utilizarlas posteriormente en una unión. Previamente debe haberse definido las *tablas de mediciones y la tabla de hechos*.

En cada *tabla de destino*, se seleccionará una columna que se podrá utilizar para identificar de modo exclusivo las filas de la tabla. Esta será su *clave principal*.

Cualidades de las columnas que deben seleccionarse como clave principal:

- **Siempre debe tener un valor:** la columna de una clave principal no puede contener valores nulos.
- **Debe tener valores exclusivos:** cada valor de la columna debe ser distinto para cada fila de la tabla.
- **Los valores deben ser estables:** un valor nunca debe cambiar por otro valor.

La definición de una *clave principal* para una tabla es altamente recomendable porque la identificación exclusiva de cada fila agiliza el acceso a las mismas.

Las *claves foráneas* se utilizan para definir las relaciones entre tablas.

En un esquema en estrella, una clave foránea define la relación entre la tabla de hechos y las tablas de mediciones asociadas a la misma. La *clave principal* de la tabla de mediciones tiene una clave foránea correspondiente en la tabla de hechos.

La *clave foránea* requiere que todos los valores de una columna determinada de la *tabla de hechos* existan en la *tabla de mediciones.*

A continuación se definirán las *claves principales* y *foráneas.*

Definición de una Clave Principal

Procedimientos para obtener una clave principal:

- Seleccionar la carpeta *Destino de Deposito* del Centro de control del depósito de datos y escoger tabla *Individuos* . Se obtiene la ventana Propiedades.
- Presionar la pestaña *claves primaria de depósito.*
- En *columnas disponibles*, seleccionar los campos: *(CODUSU, NRO_HOGAR, COMPONENTE, H15, ANO4, TRIMESTRE, REGION, MAS_500, AGLOMERADO, PONDERA)* como clave principal.

- Oprimir > para trasladar los campos a *Columnas de claves principales.*
- Dejar el campo *nombre de restricción* vacío, de modo que *DB2 Universal Database* genere un nombre de restricción.

Una *clave principal* puede considerarse como una restricción, porque todos los valores de la columna seleccionada deben ser exclusivos (ver fig. 5.50 de la pág.117).

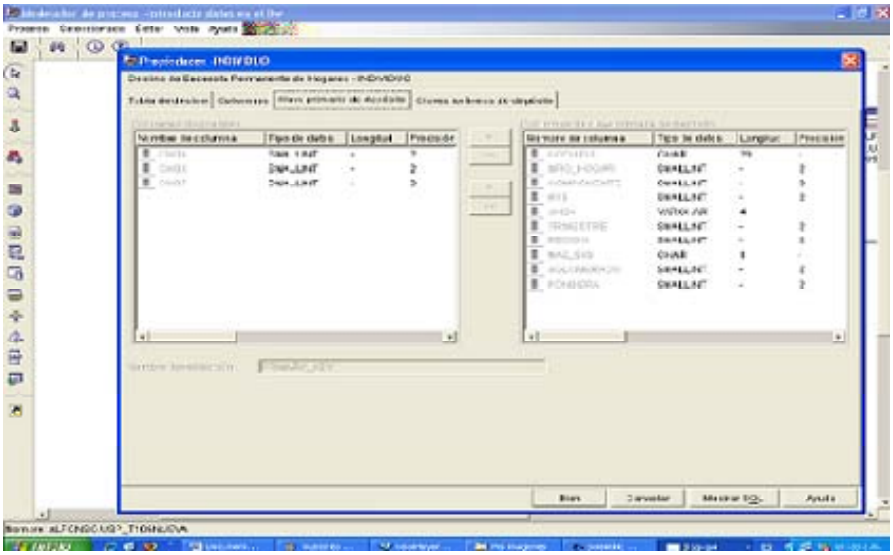

Figura 5.50: Obtencion de claves primarias de depósito.

- Presionar el botón Bien para guardar las definiciones.

Se deberá relizar los mismos pasos para definir claves principales para las otras tablas de destino.

Definición de Clave Foránea

Es necesario definir *claves foráneas* para las relaciones entre la tabla *Individuos* y las demás tablas de destino *(Asalariados, Independientes, Individuo,*

Nivel _ educativo, Ocupación _ principal, Pob_ con_ Plan_ Jefes_ y_ Jefas, Pob_ Desocupada, Pob_ Desocupada_ con_ empleo_ Anterior, Pob_ Ocupado).

Para definir las *claves foráneas*:

- Visualizar la tabla *Asalariados* en la lista de tablas de la base de datos *PDESTINO*. Luego pulsar con el botón derecho del ratón en la tabla y presionar Modificar.

  Apertura del cuaderno Modificar tabla:

- Pulzar pestaña *Claves foráneas.*
- Pulsar Añadir. Se abrirá la ventana Añadir clave foránea.
- **Esquema de tabla**: escribir el ID de usuario.
- **Nombre de tabla**: especificar *Individuos*, que es la tabla padre. El campo *Clave principal* muestra la clave principal para *Individuos.*
- **Columna disponible**: se deberá seleccionar *(CODUSU, NRO_ HOGAR, COMPONENTE, H15, ANO4, TRIMESTRE, REGION, MAS_ 500, AGLOMERADO, PONDERA)* como las columnas que se desean definir como *clave foránea.*
- Pulsar > para trasladar *(CODUSU, NRO_ HOGAR, COMPONENTE, H15, ANO4, TRIMESTRE, REGION, MAS_ 500, AGLOMERADO, PONDERA)* a la lista *Clave foránea.*
- Aceptar los valores por omisión para los campos en la supresión y en la actualización.
- Dejar el campo Nombre de restricción vacío, de modo que *DB2 Universal Database* genere un nombre de restricción.

Una clave foránea puede considerarse como una restricción, porque para cada valor de la columna de clave foránea de la tabla dependiente hay una fila de la tabla padre con un valor coincidente en la columna de clave principal del padre.

- Una vez finalizado, se deberá pulsar el botón Bien para guardar las definiciones.

- Realizar los mismos pasos para definir *claves foráneas* para las otras tablas de destino (ver fig. 5.51 de la pág.119).

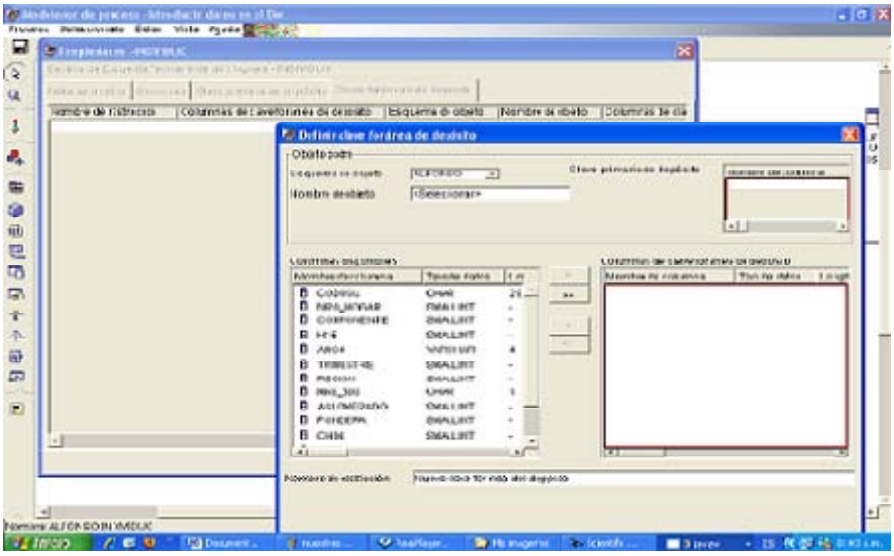

Figura 5.51: Definición de claves foráneas de depósitos.

Creación de un Esquema en Estrella Desde el Centro de Depósito de Datos

Se creará un *esquema en estrella* desde las tablas de depósito especificadas con anterioridad.

Se podrá utilizar este *esquema en estrella* para consultas en la base de datos de depósito. También se podrá exportar el esquema en estrella a *OLAP Integration Server* para crear una base de datos *OLAP*.

Definición de un Esquema en Estrella

En este apartado se definirá el *esquema en estrella* que debe contener las tablas de mediciones y de hechos ya definido en la sección *Instalación del Ambiente Datamart.*

Para definir un *Esquema en estrella* se debe realizar los siguentes pasos:

- Desde el *Centro de depósito de datos*, pulsar con el botón derecho del ratón en la carpeta Esquemas de depósito y luego en Definir.
- Se abrirá el cuaderno Definir esquema de depósito.
- **Nombre**: del esquema, *Esquema de la EPH.*
- **Administrador**: contacto para el esquema.
- **Descripción**: breve comentario del esquema: *esquema en estrella de Encuesta Permanente de Hogares.*
- Aceptar el resto de los valores.
- Seleccionar el recuadro Utilizar solo una base de datos.
- Desde la lista Base de datos de destino de depósito, seleccionar *PDESTINO* (ver fig. 5.52 de la pág.121).
- Pulsar el botón **Bien** para definir el esquema de depósito.

El esquema de depósito se añade al árbol debajo de la carpeta *Esquemas de depósito.*

Apertura del Esquema en Estrella

Para abrir el *esquema de depósito* se debe ralizar lo siguente:

- Hacer una doble pulsación en el esquema de depósito *Esquema de la EPH.*
- Pulsar el botón **Abrir**.

Adición de Tablas al Esquema en Estrella

Para adicionar las tablas de mediciones y las tablas de hechos *(Asalariados, Independientes, Individuo, Nivel _ educativo, Ocupación _ principal, Pob_con_Plan_Jefes_y_Jefas, Pob_Desocupada, Pob_Desocupada_ con_*

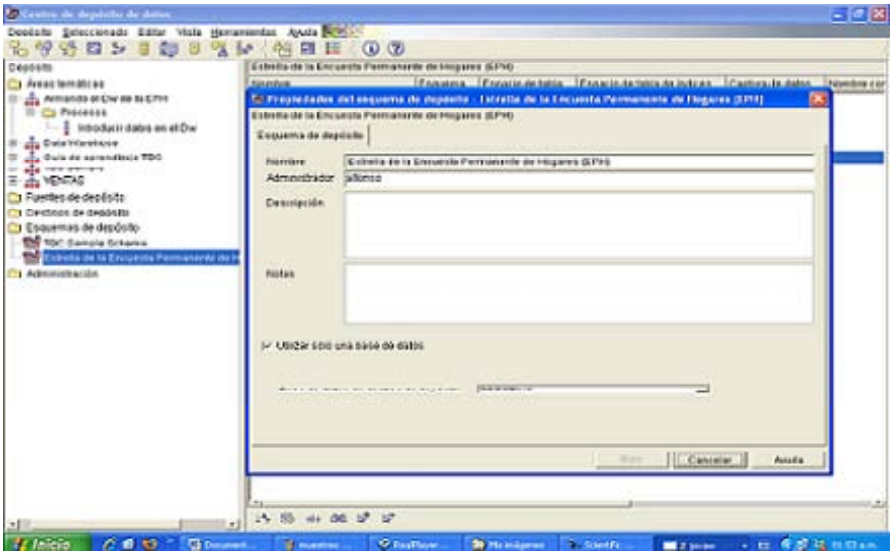

Figura 5.52: Visualización del Cuaderno de Definicíon del esquema de depósito.

empleo_Anterior, Pob_Ocupado) al esquema en estrella se debe desarrollar los siguientes procedimientos:

- Pulsar el ícono Añadir datos.
- Pulsar sobre la cuadrícula en el punto donde desea colocar las tablas. Se abrirá la ventana Añadir datos.
- Expandir el árbol Destinos de depósito hasta que se visualice una lista de tablas bajo la carpeta Tablas.
- Seleccionar la tabla *Asalariados.*
- Pulsar > para añadir la tabla *Asalariados* a la lista Tablas fuente y destino seleccionadas.
- Repitir los dos últimos pasos, para añadir el resto de las tablas.
- Pulsar el botón **Bien**.

  Las tablas que se han seleccionado anteriormente se visualizarán en la ventana *Modelo de esquema de depósito* (ver fig. 5.53 de la pág.122).

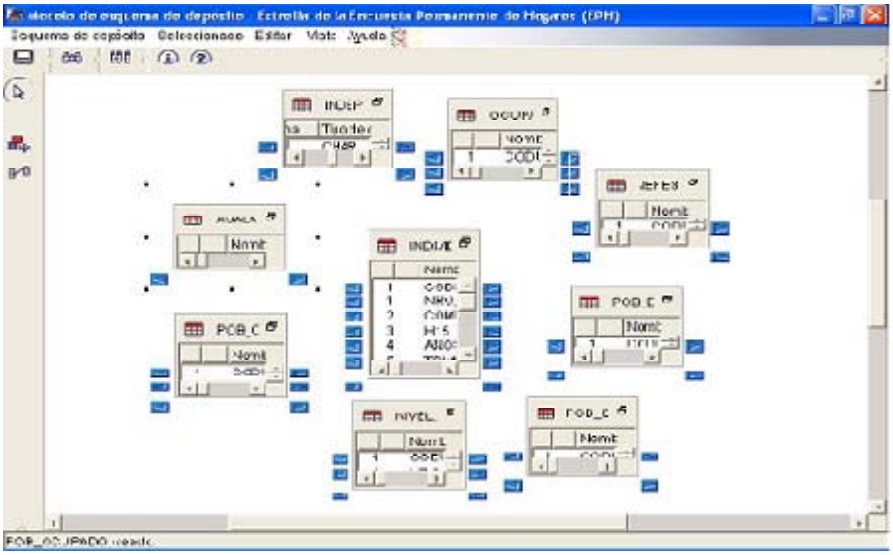

Figura 5.53: Adición de las tablas de mediciones y las de hechos al esquema de estrella.

Unión Automática de Tablas en el Esquema en Estrella

Para lograr la fusión de las tablas dentro del *Centro de Deposito de Datos*, se utilizarán las *claves principales y foráneas*.

Para realizar la unión automática de las tablas de mediciones y las tablas de hechos se debe tener en cuenta lo siguiente:

- Selección de tablas *(Asalariados, Independientes, Individuo, Nivel _ educativo, Ocupación _ principal, Pob_ con_ Plan_ Jefes_ y_ Jefas, Pob_ Desocupada, Pob_ Desocupada_ con_ empleo_ Anterior, Pob_ Ocupado)* manteniendo pulsada la tecla Ctrl y pulsando en cada tabla.
- Presionar el ícono Unión automática en la barra de herramientas.
- El Centro de depósito de datos traza líneas verdes entre las claves principales de las tablas de mediciones y las tablas foráneas de la tabla de hechos. Se pueden cambiar estas líneas hasta que guarde el trabajo (para eliminar las líneas, se debe realizar una selección múltiple pulsando con el botón derecho del ratón y luego Eliminar).

- Pulsar el ícono Guardar de la barra de herramientas para guardar el trabajo: las líneas verdes de unión automática se vuelven negras (ver fig. 5.54 de la pág.123 y 5.55 de la pág.124).

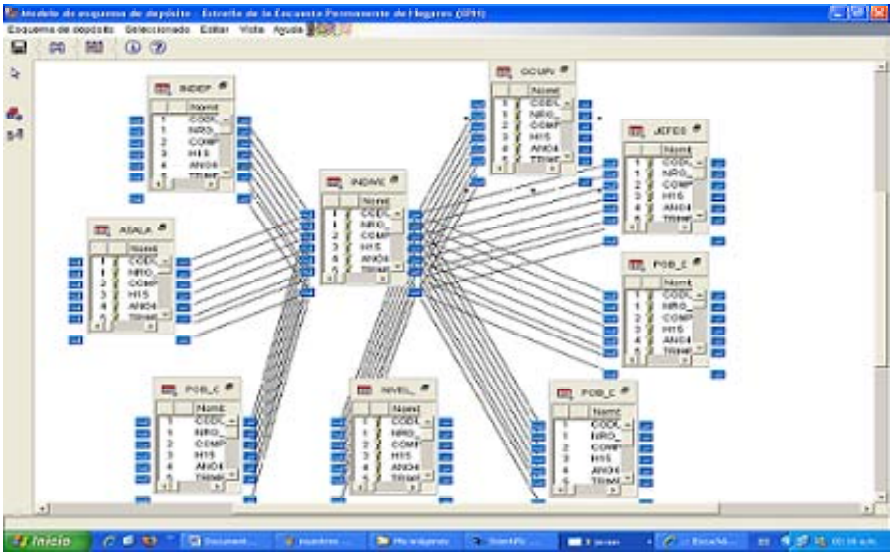

Figura 5.54: Visualización del Modelo de Estrella después de la unión automática.

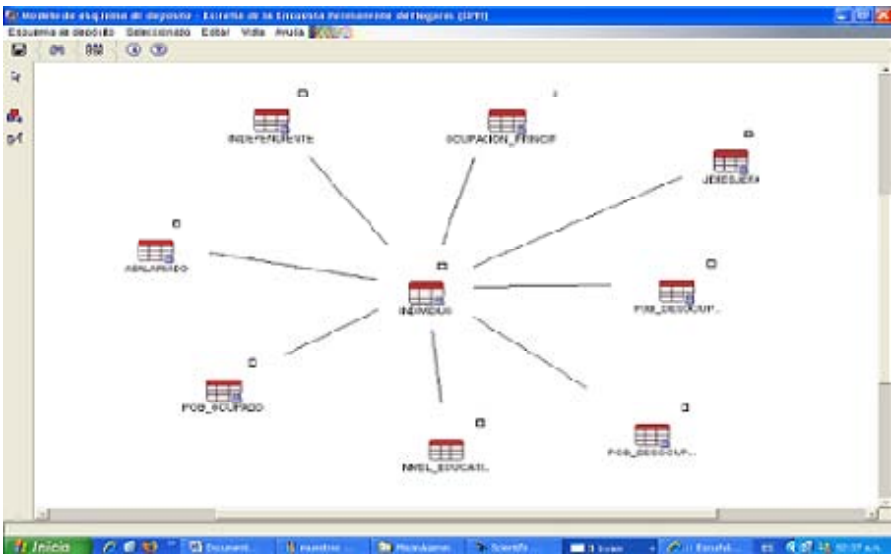

Figura 5.55: Visualización del Modelo de Estrella luego de utilizar la opción ocultar columnas.

Capítulo 6

Extracción de Conocimiento con IBM DB2 Intelligent Miner for Data

6.1 Conceptos de Minería de Datos

La *minería de datos* suele describirse cómo *"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. Los modelos de minería de datos se pueden aplicar a situaciones empresariales como las siguientes [9]:

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

El siguiente diagrama describe las relaciones entre cada paso del proceso (ver fig. 6.1 de la pág. 126).

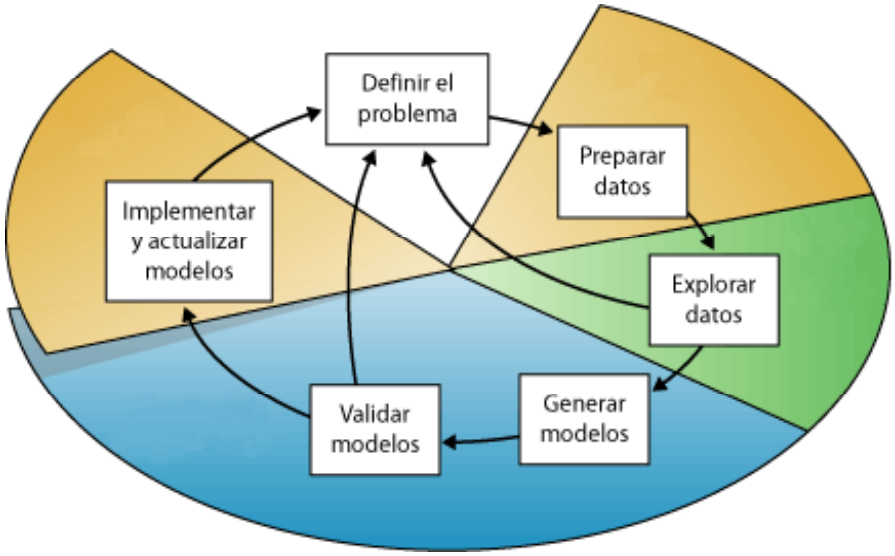

Figura 6.1: Proceso que se ilustra la generación de un modelo de minería de datos.

Aunque el proceso que se ilustra en el diagrama es circular, esto no significa que cada paso conduzca directamente al siguiente. La creación de un modelo de minería de datos es un proceso dinámico e iterativo. Una vez que se han explorado los datos, puede que se descubra que resultan insuficientes para crear los modelos de minería de datos adecuados y que, por tanto, se debe buscar más datos.

Se puede generar varios modelos y descubrir que no responden al problema planteado cuando se lo definió y que, por tanto, se debe volver a definir el problema. Es posible que se deba actualizar los modelos una vez implementados debido a que haya más datos disponibles. Por esto, es importante comprender que la creación de un modelo de minería de datos es un proceso, y que cada paso del proceso puede repetirse tantas veces como sea necesario para crear un modelo válido.

IBM DB2 Intelligent Miner for Data V8.1 ofrece un entorno integrado para crear y trabajar con modelos de minería de datos. El entorno incluye algoritmos y herramientas de minería de datos que facilitan la generación de

una solución completa para diversos proyectos. Para obtener más información acerca de cómo usar *IBM DB2 Intelligent Miner for Data V8.1* ver el Capítulo N^{o}4 *(Introducción a Intelligent Miner for Data).*

6.1.1 Definir el Problema

El primer paso del proceso de *minería de datos*, como se resalta en el siguiente diagrama, consiste en definir claramente el problema a resolver (ver fig. 6.2 de la pág. 127).

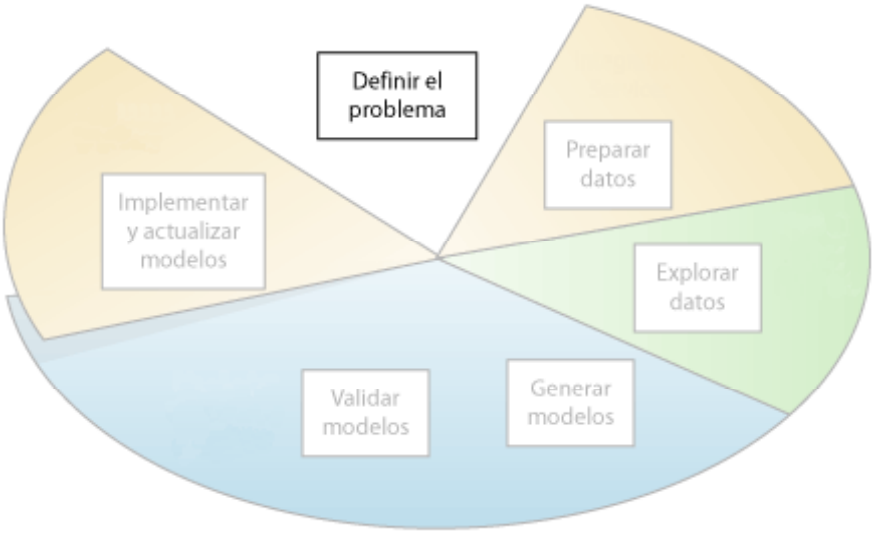

Figura 6.2: El primer paso del proceso, implica en definir claramente el problema.

Este paso incluye analizar los requisitos de la organización, definir el ámbito del problema, definir las métricas por las que se evaluará el modelo y definir el objetivo final del proyecto de *minería de datos.* Estas tareas se traducen en preguntas como las siguientes:

- *¿Qué se está buscando?.*
- *¿Qué atributo del conjunto de datos se desea intentar predecir?.*
- *¿Qué tipos de relaciones se intenta buscar?.*

- *¿Se desea realizar predicciones a partir del modelo de minería de datos o sólo buscar asociaciones y patrones interesantes?.*
- *¿Cómo se distribuyen los datos?.*
- *¿Cómo se relacionan las columnas?, o en caso de haber varias tablas, ¿cómo se relacionan las tablas?.*

Para responder a estas preguntas, es probable que se deba dirigir un estudio de disponibilidad de datos para investigar las necesidades de los usuarios de la organización con respecto a los datos disponibles. Si los datos no son compatibles con las necesidades de los usuarios, puede que se deba volver a definir el proyecto.

6.1.2 Preparar los Datos

El segundo paso del proceso de minería de datos, como se indica en el siguiente diagrama, consiste en consolidar y limpiar los datos identificados en el paso *Definir el Problema* (ver fig. 6.3 de la pág. 128).

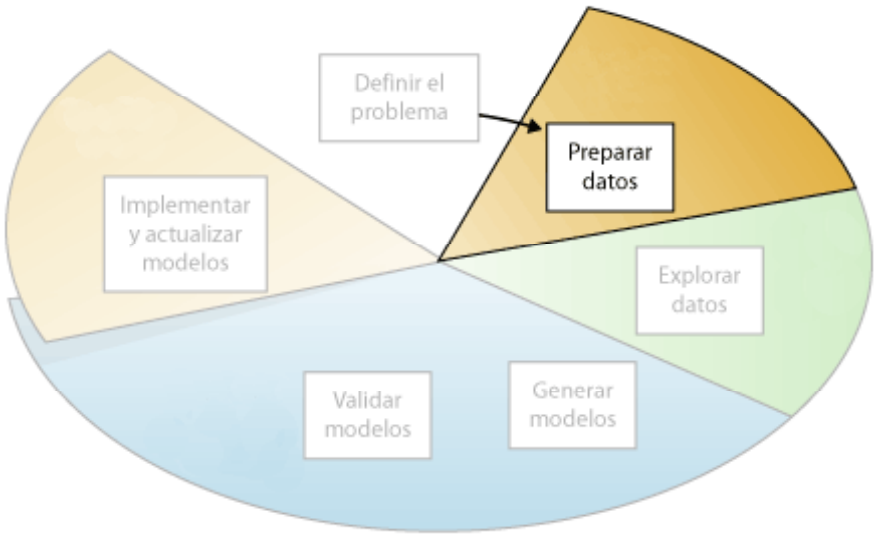

Figura 6.3: El segundo paso, consiste en la depuración y consolidación de los datos.

Los datos pueden estar dispersos en la organización y almacenados en distintos formatos. *IBM DB2 Intelligent Miner for Data* puede utilizar como datos de entrada archivos planos, donde estos también pueden contener incoherencias como datos faltantes *"missings"* , fuera de rango *"outliers"* o simplemente contener errores.

Por ejemplo: los datos pueden mostrar que un cliente adquirió un producto incluso antes de haber nacido o que el cliente compra regularmente en una tienda situada a 3.000 kilómetros de su casa. Antes de empezar a generar modelos, se debe solucionar estos problemas. Normalmente se trabaja con un conjunto de datos muy grande y no se puede comprobar cada transacción. Es por ello que este paso es de suma importancia ya que es aquí donde se tendrá que realizar las correspondientes y verificaciones para obtener resultados fehacientes.

Calidad en los Datos

El éxito de las actividades de Data Mining se relaciona directamente con la calidad de los datos.

Muchas veces resulta necesario pre-procesar los datos antes de derivarlos al modelo de análisis. El pre-procesamiento puede incluir transformaciones, reducciones o combinaciones de los datos.

La semántica de los datos debe ayudar para la selección de una conveniente representación y las bondades de la representación elegida gravitan directamente sobre la calidad del modelo y de los resultados posteriores.

Problemas con los Datos

En la fase de Preparación de Datos, pueden suceder una diversidad de casos:

- *Demasiados datos:*
  - Datos corruptos o con ruido.
  - Datos redundantes *(requieren factorización).*
  - Datos irrelevantes.
  - Excesiva cantidad de datos *(muestreo).*

- *Pocos datos:*
  - Atributos perdidos *(missings).*
  - Valores perdidos.
- *Poca cantidad de datos*
  - Datos fracturados.
  - Datos incompatibles.
  - Múltiples fuentes de datos.

6.1.3 Explorar los Datos

El tercer paso del proceso de minería de datos, como se resalta en el siguiente diagrama, consiste en explorar los datos preparados (ver fig. 6.4 de la pág. 130).

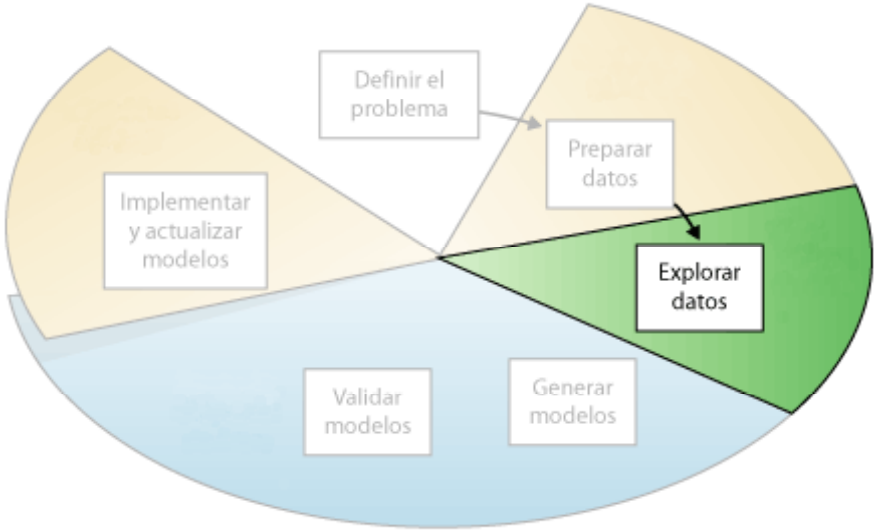

Figura 6.4: Se debe comprender los datos para selecionar un modelo adecuado.

Se debe comprender los datos para tomar las decisiones adecuadas al crear los modelos. Entre las técnicas de exploración se incluyen calcular los valores mínimos y máximos, calcular la media y las desviaciones estandar y examinar

la distribución de los datos. Una vez explorados los datos, se puede decidir si el conjunto de datos contiene datos con errores y, a continuación, crear una estrategia para solucionar los problemas.

6.1.4 Generar Modelos

El cuarto paso del proceso de minería de datos, como se resalta en el siguiente diagrama, consiste en generar los modelos de minería de datos (ver fig. 6.5 de la pág. 131).

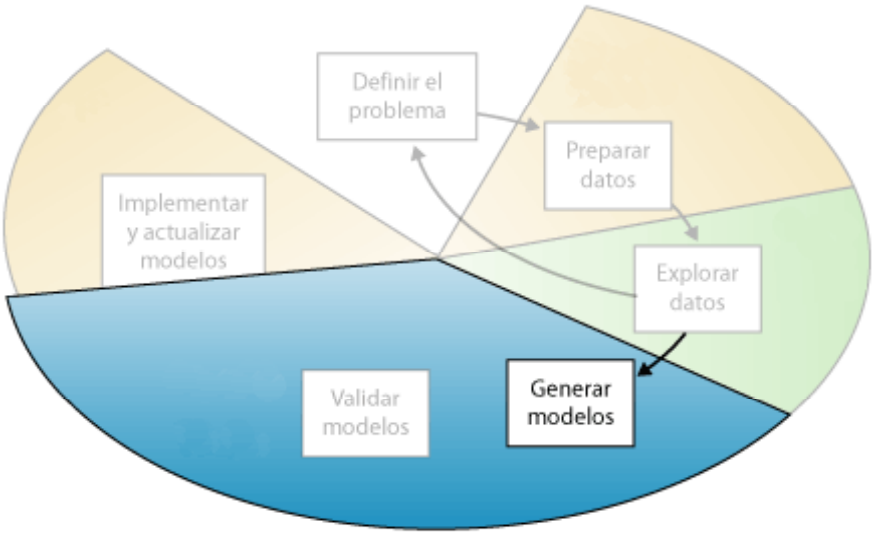

Figura 6.5: Un modelo, es una tabla de datos compuesta por filas y columnas.

Antes de generar un modelo, se deben separar aleatoriamente los datos preparados en conjuntos de datos de entrenamiento y comprobación independientes. El conjunto de datos de entrenamiento se utiliza para generar el modelo y el conjunto de datos de comprobación para comprobar la precisión del modelo mediante la creación de consultas de predicción.

Se utilizarán los conocimientos adquiridos en el paso *Explorar los Datos* para definir y crear un modelo de minería de datos. Normalmente, los modelos contienen:

- *Columnas de Entrada.*

- *Columna de Identificación.*
- *Columna dePredicción.*

Es decir que los datos para data mining se organizan en forma de una tabla plana compuesta por Filas y Columnas, donde:

- Las Filas: Son las *unidades de análisis.* Por ejemplo: una cuenta bancaria, un ticket de un supermercado, etc.
- Las Columnas: Los *atributos* de cada unidad de análisis. Por ejemplo: la frecuencia de uso de la tarjeta de crédito, sexo, edad, etc.

Una vez definida la estructura del modelo de minería de datos, se la procesa rellenando la estructura vacía con los patrones que describen el modelo. Esto se conoce como entrenar el modelo.

Los patrones se encuentran al pasar los datos originales por un algoritmo matemático. *IBM DB2 Intelligent Miner for Data V8.1* contiene un algoritmo diferente para cada tipo de modelo que se puede generar. Se puede utilizar parámetros para ajustar cada algoritmo.

El modelo de minería de datos se define mediante:

- *Objeto de estructura de minería de datos.*
- *Objeto de modelo de minería de datos.*
- *Algoritmo de minería de datos.*

Características de las Tablas de Datos Para Data Mining

Como se hacía referencia anteriormente, un modelo de *Minería de Datos* se organiza como una tabla plana, con filas y columnas. En donde en ella se tiene las siguientes particularidades:

- Cada fila debe corresponder a una instancia relevante al caso de estudio.
- Todos los datos deben estar en una sola tabla o "vista" de la Base de Datos.

- Las columnas sin variabilidad deben ser ignoradas.
- Los atributos con valores únicos para cada caso deben ser ignoradas *(nro. de cuenta, DNI, etc.)*. Muchas veces este tipo de información contiene datos sensibles.

Datos sensibles: Datos personales que revelan origen racial y étnico, opiniones políticas, convicciones religiosas, filosóficas o morales, afiliación sindical e información referente a la salud o a la vida sexual. *Art. Nº 7 de la Ley Nº25326 Protección de los Datos Personales.*

Resumiendo, para tener una mejor comprensión del problema se debe *factorizar* (*reducir dimensionalidad*) logrando así un modelo terminado.

6.1.5 Explorar y Validar los Modelos

El quinto paso del proceso de *Minería de Datos*, como se resalta en el siguiente diagrama, consiste en explorar los modelos que se han generado y comprobar su eficacia (ver fig. 6.6 de la pág. 133).

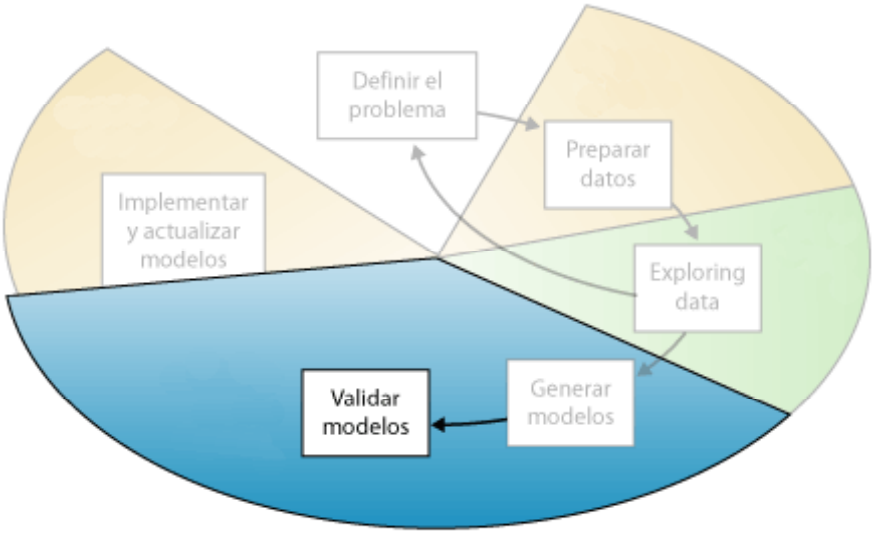

Figura 6.6: La validación implica la seleción del modelo que se adapte mejor.

No se debe implementar un modelo en un entorno de producción sin comprobar primero si el modelo funciona correctamente. Además, puede que se hayan creado varios modelos y se deba decidir cuál funciona mejor. Si ninguno de los modelos que se han creado en el paso Generar Modelos funciona correctamente, puede que se deba volver a un paso anterior del proceso y volver a definir el problema o volver a investigar los datos del conjunto de datos original.

6.1.6 Implementar y Actualizar los Modelos

El último paso del proceso de minería de datos, como se resalta en el siguiente diagrama, consiste en implementar los modelos que funcionan mejor en un entorno de producción (ver fig. 6.7 de la pág. 134).

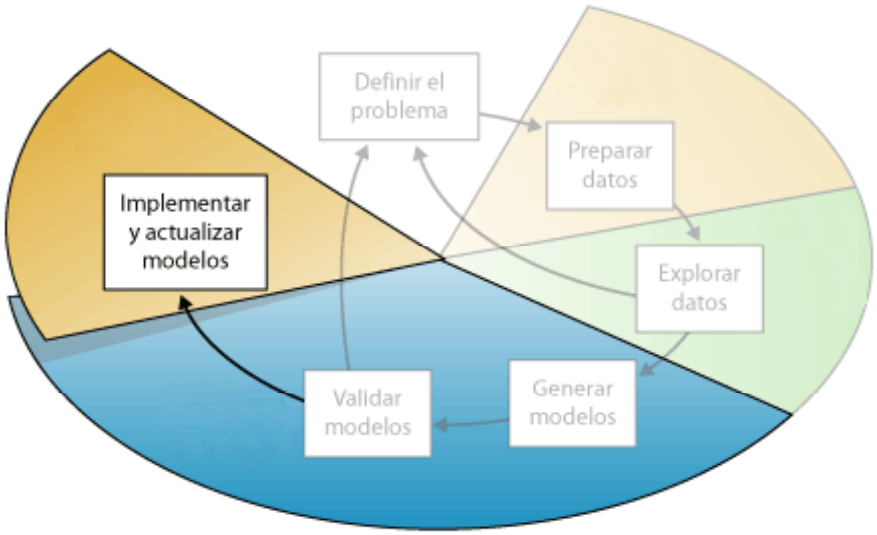

Figura 6.7: La implementación es el ultimo paso de el proceso.

Una vez que los modelos de minería de datos se encuentran en el entorno de producción, se pueden llevar acabo diferentes tareas, dependiendo de las necesidades. Éstas son algunas de las tareas que se pueden realizar:

- *Utilizar los modelos para crear predicciones que se puedan utilizar para tomar decisiones empresariales.* Por ejemplo: la predicción de demanda

, optimización de campañas - tracking de campañas y predicción de respuesta / no respuesta.

- *Incrustar la funcionalidad de minería de datos directamente en una aplicación.*
- *Aplicaciones del modelo de minería de datos a negocios electrónicos.* Por ejemplo: para mejorar la estructura del *Website* (mejora en tiempos de acceso, análisis de tráfico y uso de recursos de e-business), se pueden aplicar las siguientes técnicas:
  - Propensión a la fuga - modelos de predicción de abandono del sitio.
  - Propensión a la compra venta cruzada (afinidad) - canasta de consumo.
  - Reglas de asociación de páginas visitadas.
  - Segmentación de visitantes, panelistas.
  - Scoring de riesgo.
  - Análisis cross/up sell - caracterización de perfiles de clientes para definir acciones de up selling y cross selling.
  - Detección de fraude.
  - Modelo de valor del cliente - identificación de clientes potenciales.
- *Crear un informe que permita a los usuarios realizar consultas directamente en un modelo de minería de datos existente.*

La actualización del modelo forma parte de la estrategia de implementación. A medida que la organización recibe más datos, debe volver a procesar los modelos para mejorar así su eficacia.

6.2 Proceso de Minería Aplicado a la EPH

Como se había mencionado anteriormente, el *Proceso de Minería*, está compuesto por los siguientes pasos [15]:

- *Definir el problema.*
- *Preparar los datos.*

- *Explorar los datos.*
- *Generar modelos.*
- *Explorar y validar los modelos.*
- *Implementar y actualizar los modelos.*

6.2.1 Definición de los Problemas

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.*

Fundamentación: los problemas laborales persisten a pesar de que la economía crece.

En la medida que se mantenga la economía en crecimiento se supone que habrá generación de empleos, pero el problema es que puede llegar a hacerlo con un ritmo muy cansino para las necesidades laborales de la población.

Cuando el crecimiento del empleo es insuficiente, la falta de empleo no necesariamente se manifiesta a través del alto desempleo, sino en la caída de las tasas de participación laboral y en el mantenimiento de muchos empleos de baja calidad. En otras palabras, la baja tasa de participación (especialmente entre las mujeres y los jóvenes) es la otra cara de la falta de oportunidades laborales.

Estos datos deberían encender una luz de precaución, aún cuando se confíe en que el crecimiento económico durará, porque sugieren que las restricciones para salir a buscar y conseguir un empleo están resurgiendo, en particular, en el interior del país.

Esto lleva a un estudio más certero acerca de la idiosincracia de los individuos del interior del País, particularmente en la provincia de Corrientes, con cuyos datos muestrales se trabajará aplicándoles numerosas técnicas de *Minería de Datos (clústering , árboles de decisión , etc.),* para descubrir patrones de información ocultos en las bases usuarias de la *Encuesta Permanente de Hogares (EPH)* [16].

Hipótesis: la mayor fuente de empleo en la provincia de Corrientes la brinda el sector Público.

Objetivos Generales: caracterizar y describir el empleo público de la Provincia de Corrientes a través de la utilización de técnicas de *Minería de Datos.*

Objetivos Específicos:

- Describir la composición del empleo en Corrientes.
- Conocer los perfiles socio demográficos 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.

6.2.2 Preparación de los Datos

Es una etapa compleja y que requiere el mayor tiempo.

El éxito del trabajo dependerá de los datos recopilados, de una buena selección y preparación.

Inicialmente se dispone de 12 *bases de datos* o *bases usuarias* (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.

Todos los objetos de estudios realizados en este apartado se elaboraron con estos datos. Los mismos son suministrados, previa registración por el portal Web del *INDEC (Instituto Nacional de Estadística y Censos) http://www .indec. mecon. ar/*

(ver fig. 6.8 de la pág. 138). Allí se encontrarán las *bases usuarias* como se pueden ver en la fig. 6.9 de la pág. 139.

También por este medio se puede descargar *documentos de consulta para el uso de la base usuaria*, estos son:

- Diseño de registros y estructura para las bases preliminares.
- Estimación de los errores de muestreo en la EPH continua.

- Tablas de errores de muestreo trimestrales.
- Clasificador de Actividades para Encuestas Sociodemográficas (CAES - MERCOSUR).
- Clasificador Nacional de Ocupaciones.
- Código de países.
- Código de provincias.

Ver fig. 6.10 de la pág. 139.

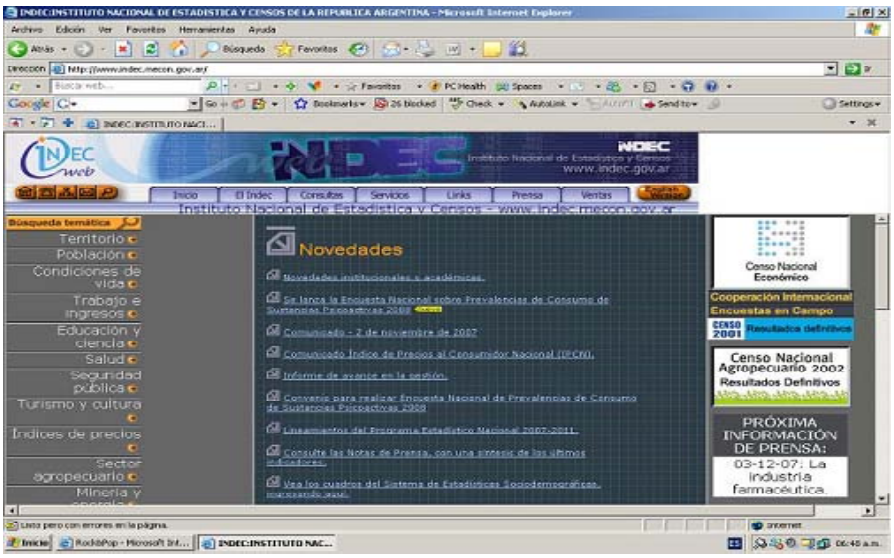

Figura 6.8: Visualización del site del INDEC, http://www.indec.mecon.ar/.

Para obtener más información sobre la etapa de Preparación de los Datos, se deberá referir al Capítulo Nº 5 "*Data Warehouse*".

6.2.3 Exploración de 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

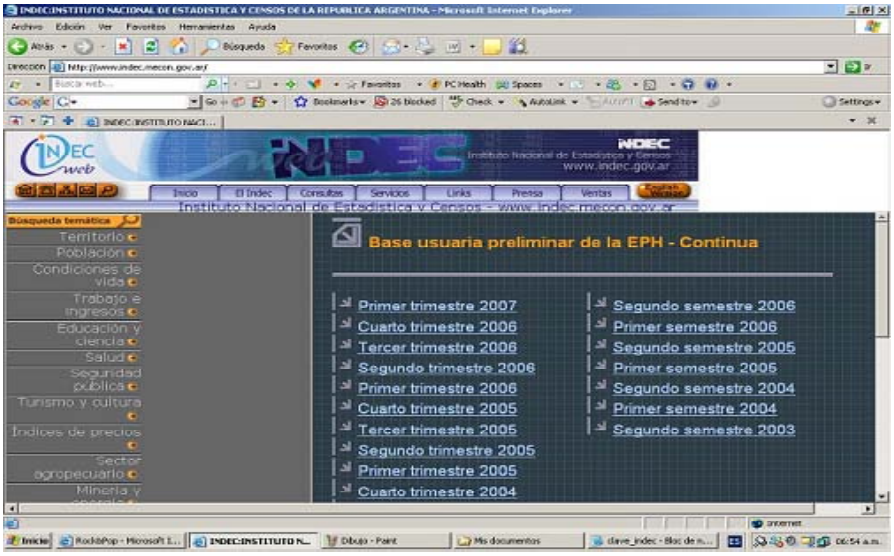

Figura 6.9: Visualización de las bases usuarias de la EPH (Encuesta Permanente de Hogares).

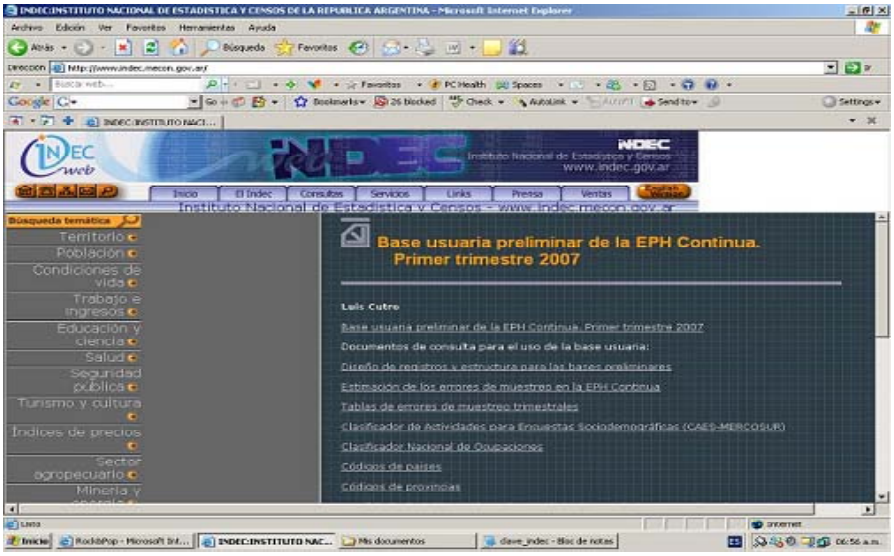

Figura 6.10: Visualización de los documentos de consulta para el uso de la base usuaria.

de los datos logrados, sería conveniente volver a redefinir el problema a tratar. Para luego continuar con el ciclo de vida del Proyecto.

Esta etapa de exploración se podría dividir en varias fases, dependiendo de los tipos de análisis y de herramientas a utilizar. En este apartado se utilizará *IBM DB2 UDB V8.1,* con el que se podrá realizar un análisis de composición de variables para cada uno de los objetivos fijados en la etapa de definición del problema. Por ejemplo: *conocer los perfiles socio demográficos de los Planes Jefes y Jefas.*

Se tendrá 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, (ver fig. 6.11 de la pág. 140).

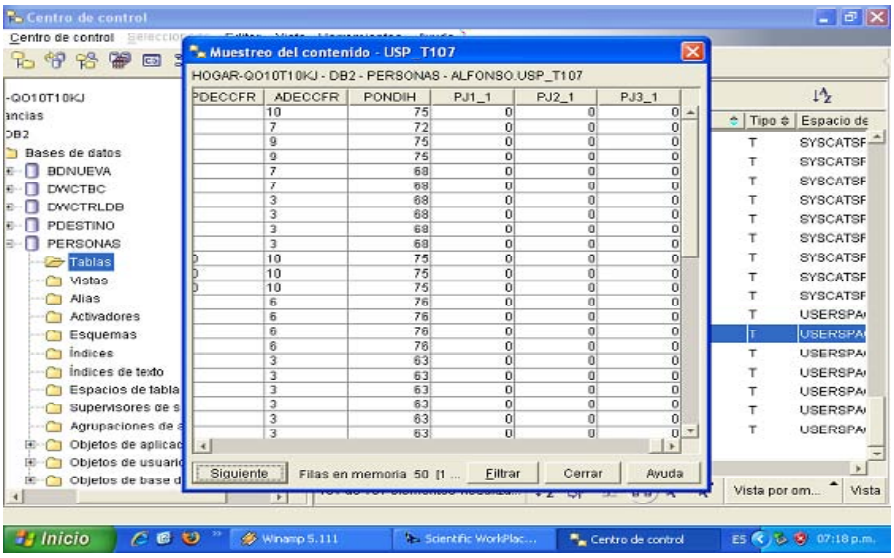

Figura 6.11: Muestreo del contenido de la variable *PJ1_1 (Existencia del plan Jefes Jefas).*

Para realizar un análisis más exhaustivo a la misma el *IBM DB2 UDB V8.1* permite aplicar técnicas de filtrado sin la utilización del códigos *SQL* (ver fig. 6.12 de la pág. 141).

Se puede obtener así un filtrado más preciso y también el número de registros exactos que cumplen con esas condiciones (ver fig. 6.13 de la pág. 141).

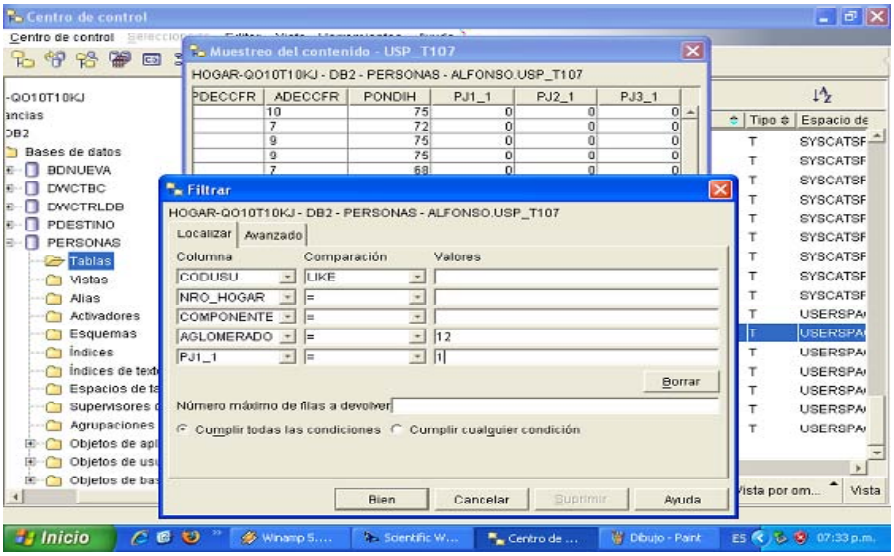

Figura 6.12: Filtrado por el Aglomerado Corrientes y por la existencia del Plan Jefa Jefe.

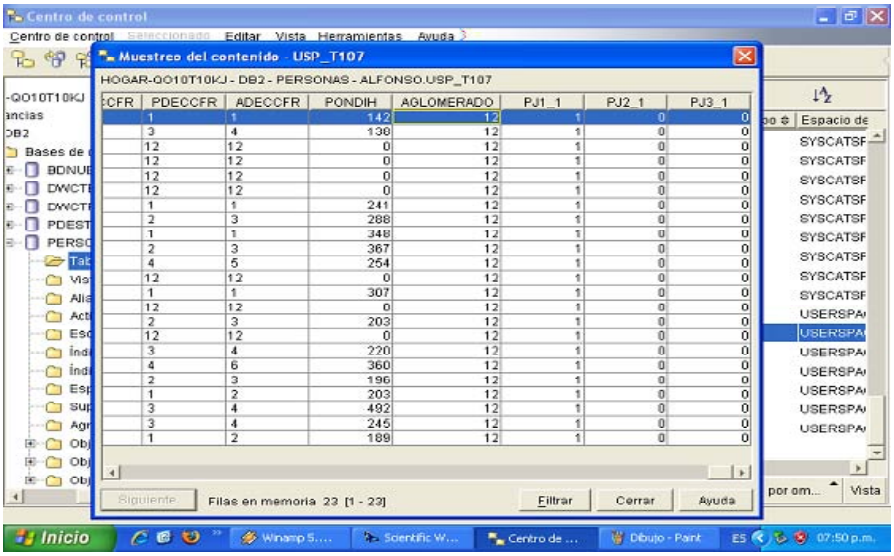

Figura 6.13: Visualización tanto del contenido como así tambien del número de los registros.

- *Indagar los perfiles Educativos de los Planes Jefes y Jefas.*

La dimensión educación está compuesta por las siguientes variables:

CH09: ¿Sabe leer y escribir?.

CH10: ¿Asiste o asistió a algún establecimiento educativo? (colegio, escuela, universidad).

CH11: Ese establecimiento es (público, privado).

CH12: ¿Cuál es el nivel más alto que cursa o cursó?.

CH13: ¿Finalizó ese nivel?.

CH14: ¿Cuál fue el último año que aprobó?.

NIVEL EDUCATIVO: Nivel Educativo.

(ver fig. 6.14 de la pág. 142).

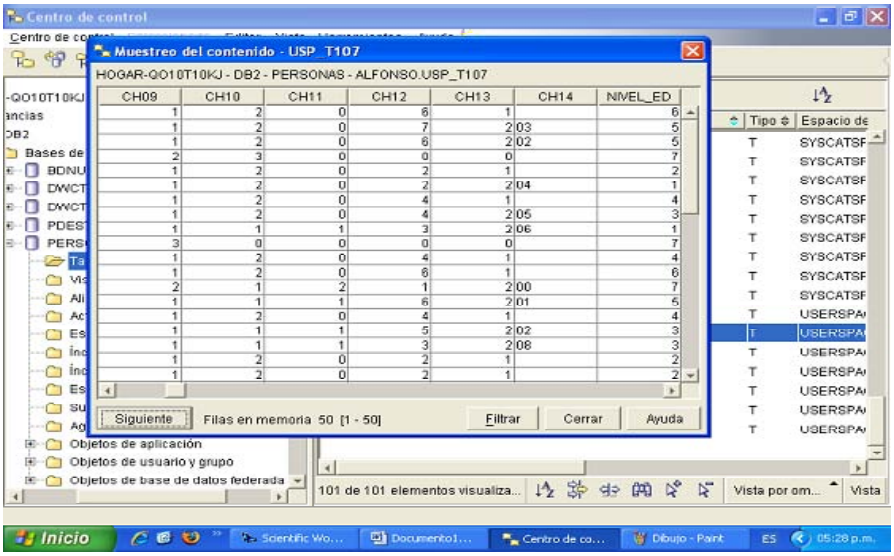

Figura 6.14: Muestreo de los valores que asumen las variables.

Para el resto de los objetivos específicos se tendrá que realizar lo antes visto, para continuar así con el ciclo de vida del Proyecto de Minería.

Considerando estos datos, simplemente se realiza un análisis exploratorios con *IBM DB2 UDB V8.1* en busca de información que pueda resultar interesante. Así mismo, se trata de comprender sobre el total de los datos, cuáles pueden ser los más importantes y determinar qué datos se pueden utilizar.

Esta fase es muy importante ya que determina que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original. Se debe determinar si los datos con los que se cuenta son suficientes para hallar conocimiento, es decir si son realmente válidos.

Algunas veces no resulta obvio que esos datos no puedan proveer las respuestas que se está buscando, por ello la importancia de prestar total atención a este punto.

6.2.4 Generación de los Modelos

En esta sección se plasmarán todos los objetivos específicos, para su posterior extracción de información.

Describir la Composición del Empleo en la Ciudad de Corrientes

Para esta problemática puntual se utilizará un software de índole netamente estadístico: *Infostat* , este permite realizar análisis de variables con múltiples funcionalidades adicionales. Esta herramienta permitirá obtener resultados, en los cuales se podrá visualizar cuáles son los perfiles de esta población.

Visualizando la fig. 6.15 de la pág. 144, se puede comprobar el elevado número de empleos que depende del Gobierno de la Provincia.

De la fig. 6.15 de la pág. 144 se pueden extraer los siguientes datos:

- Administración pública, defensa y seguridad social obligatoria: 16 %.
- Enseñanza: 13 %.
- Servicios de esparcimiento y servicios culturales y deportivos: 4 %.
- Construcción: 8 %.
- Servicios de hogares privados que contratan servicio doméstico: 13 %.

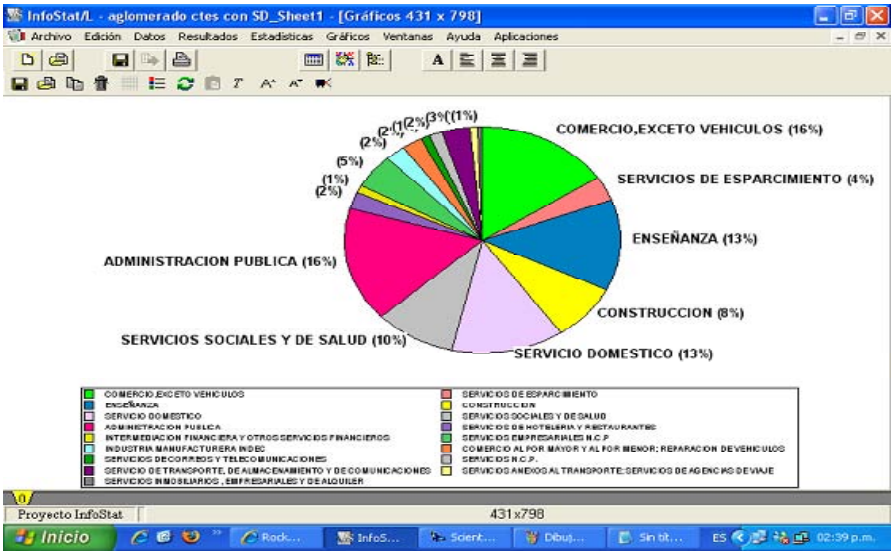

Figura 6.15: Visualización del grafico de frecuencias, de la composición del empleo de Corrientes.

- Servicios sociales y de salud: 10 %.
- Comercio al por mayor, en comisión y al por menor, excepto vehículos automotores y motocicletas: 16 %.
- Servicios de hotelería y restaurantes: 2 %.
- Intermediación financiera y otros servicios financieros: 1 %.
- Servicios empresariales N.C.P.: 5 %.
- Industria manufacturera INDEC: 2 %.
- Comercio al por mayor y al por menor, reparación vehículos automotores, motocicletas, efectos personales y enseres domésticos: 2 %.
- Servicios de correos y telecomunicaciones: 1 %.
- Servicios N.C.P.: 2 %.
- Servicios de transporte, de almacenamiento y de comunicaciones: 3 %.
- Servicios anexo al transporte; servicios de agencias de viaje: 1 %.

- Servicios inmobiliarios, empresariales y de alquiler: 1%.

Este gráfico permite sacar conclusiones, no solamente observando las frecuencias de los correspondientes rubros.

Conocer los Perfiles Socio Demográficos de los Planes Jefes y Jefas

Luego de obtener una visión general de las actividades económicas de la población en el punto anterior, se puede seguir con la investigación.

Es indispensable saber que hasta el momento no se han utilizado herramientas de extracción de conocimiento en bases de datos, *KDD (Knowledge Discovery in Databases).*

Lo que se realizará aquí es una descripción de perfiles de los individuos, en este caso los que posean planes asistenciales. Todo esto aplicando la técnica de *Clustering Demográfico* con el *IBM DB2 Intelligent Miner for Data V8.1.*

A partir de esta etapa se comienza a trabajar con *Intelligent Miner for Data* e *Intelligent Miner Visualizer,* el primero para el análisis en sí y el segundo para visualizar los resultados.

Para comprender la creación y utilización de los diferentes objetos de formulación es conveniente profundizar en primer lugar, con los conceptos claves que se explican en los Capítulos N^{o}4 *"Introducción a Intelligent Miner for Data".*

Básicamente, los pasos a llevar a cabo son:

- Creación de los Objetos de Datos (datos de entrada).
- Transformación de los datos aplicando funciones (Discretización, Correspondencia de valores, Correspondencia de nombres).
- Creación de la Base de Minería **PERSONAS**.
- Creación de Objetos Adicionales.

Creación de los Objetos de Datos (Datos de Entrada) Una vez ingresado al servidor *Intelligent Miner*, es necesario configurar correctamente la conexión al mismo (ver fig. 6.16 de la pág. 146).

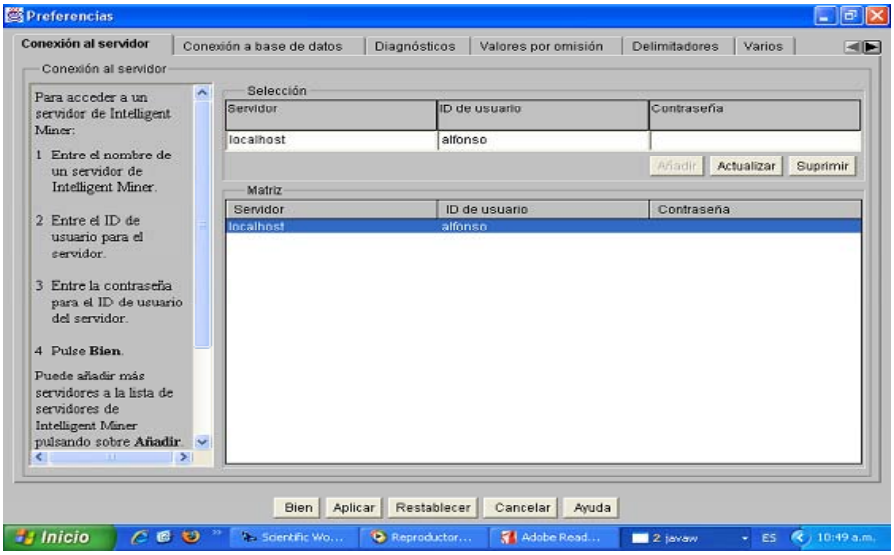

Figura 6.16: Para acceder al Intelligent Miner, deberá ingresar (Servidor, ID de Usuario y Contraseña).

Ya configuradas las opciones de conexión, queda indicarle al servidor cuáles serán los datos de entrada. Para ello se tendrá que presionar la opción crear Datos, inmediatamente aparecerá el asistente que guiará con las opciones correspondientes (ver fig. 6.17 de la pág. 147).

El asistente orientará a lo largo de los siguientes pasos:

- Selección del tipo de datos para la definición de los datos de entrada o de los datos de salida.
- Selección de los nombres de las tablas de base de datos, vistas o archivos planos.
- Especificación de parámetros para los datos de entrada o salida.
- Especificación del nombre de los datos de entrada o salida.

Una vez seleccionado el formato y el nombre de la entrada de datos (ver fig. 6.18 de la pág. 147).

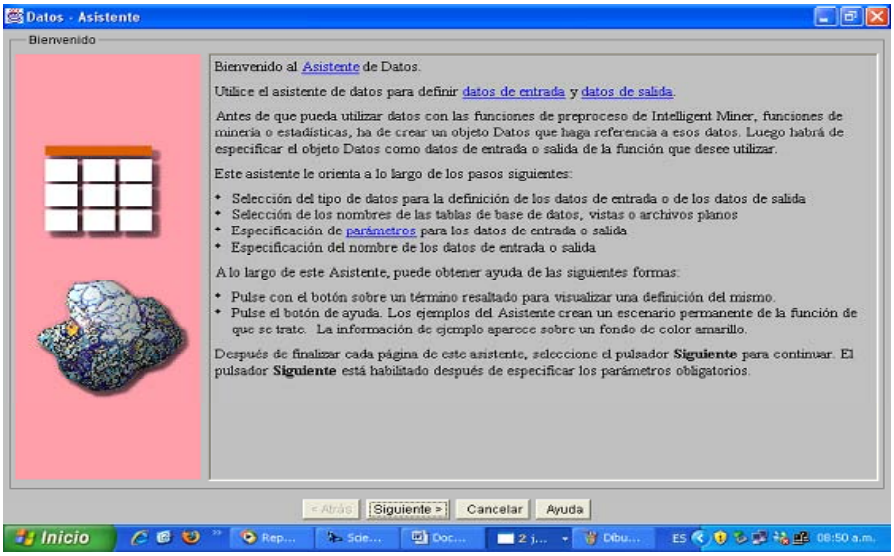

Figura 6.17: Iniciación del asistente de datos, este nos orientará a lo largo de todo este paso.

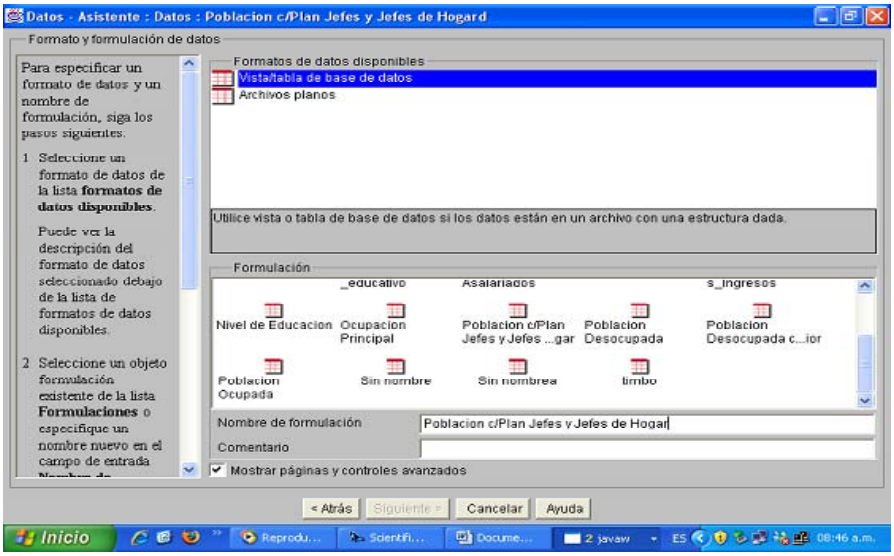

Figura 6.18: En la definición de los datos, escogemos el formato de vista/tabla de base de datos.

El siguiente paso es de seleccionar el servidor de base de datos, con su correspondiente Tabla asociada, en este caso será la *USP_ T107 (Base usuaria Persona del primer trimestres del 2007)* (ver fig. 6.19 de la pág. 148).

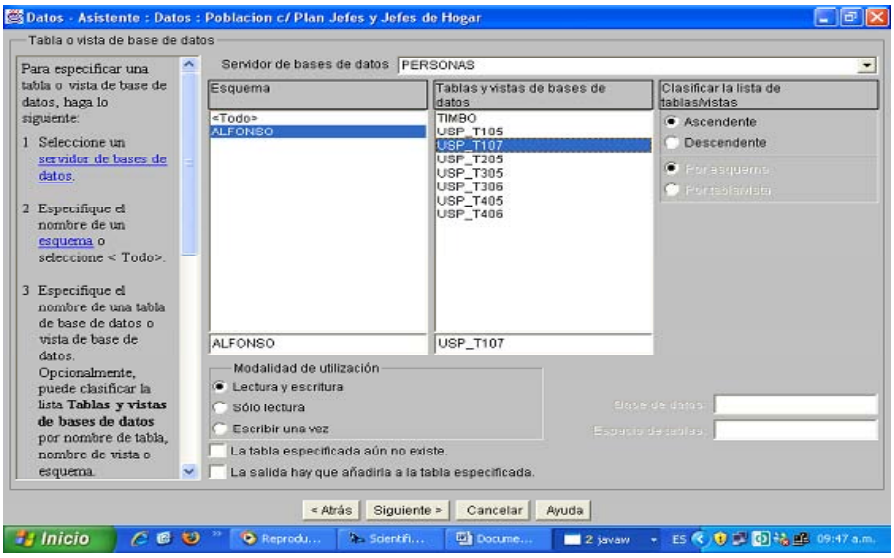

Figura 6.19: Selección del servidor, esquema, tablas/vistas de base de datos.

Como se podrá observar en la fig. 6.20 de la pág. 149, dentro de los parámetros de campo se encuentran:

- Nombre de campo en la *Base de Datos en DB2*.
- Tipo de datos del campo en la *Base de Datos en DB2*.
- Tipo de dato del campo en *Intelligent Miner for Data* (permite modificar).
- Correspondencia de nombres (permite aplicar una determinada correspondencia para un campo específico).

En el gráfico anterior se pueden visualizar los diferentes parámetros de campos (Nombre de campo, Tipo de datos y Correspondencia de nombres).

Para finalizar, queda la etiqueta de *"Campos calculados"* (ver fig. 6.21 de la pág. 150). Estos son campos adicionales que contienen información resultante, ya sea mediante una:

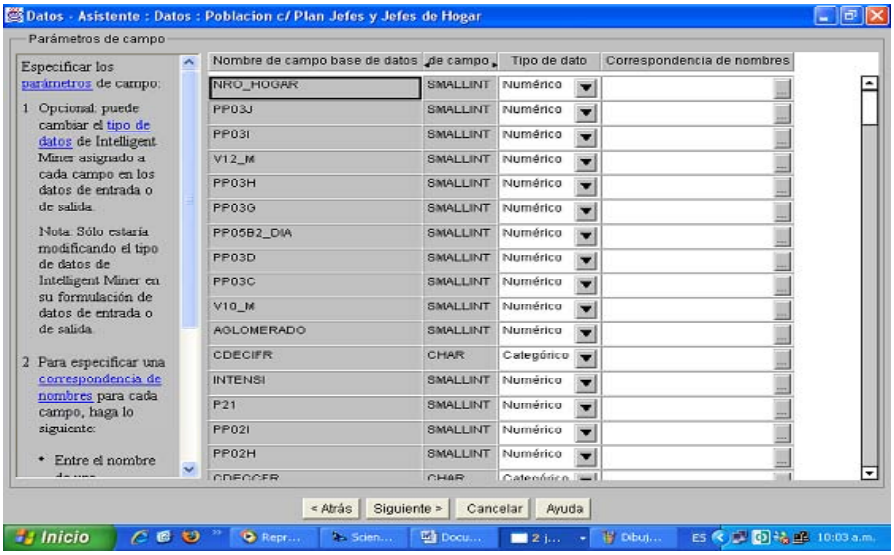

Figura 6.20: Selección o modificación de los parámetro de los campos.

- *Discretización* .
- *Correspondencia de Valores* .
- *Correspondencia de Nombres* .
- *Función* .

Los mismos son calculados por el *Intelligent Miner for Data* durante una ejecución de minería y serán profundizados más adelante.

Transformación de los Datos Aplicando Funciones En este apartado se crearán diferentes funciones con el fin de relacionarlos con los Datos de entrada correspondientes.

Correspondencia de Nombres Estos objetos convierten los valores una vez finalizada la ejecución de *Minería de Datos,* con lo cual es el Visualizador de Resultados el que muestra los valores convertidos.

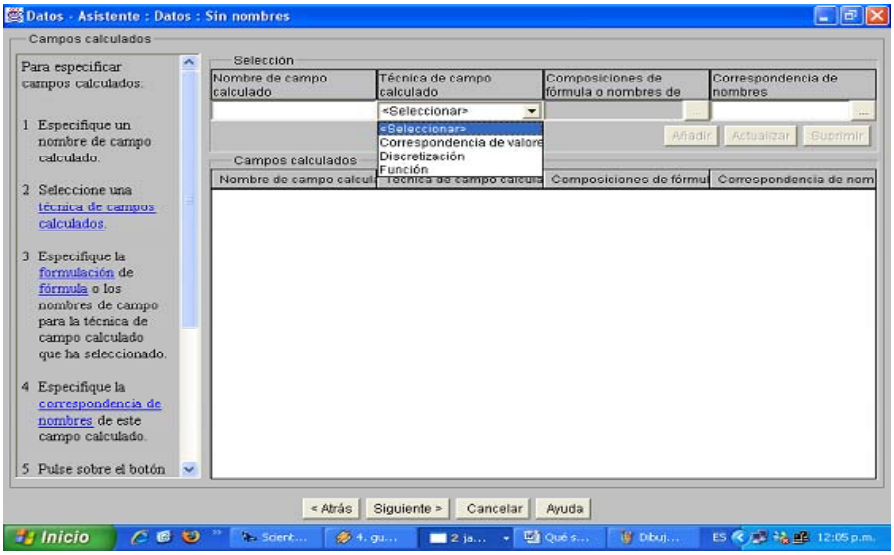

Figura 6.21: Selección de una tecnica de campo calculado (Discretización, Función, Etc.).

Se utilizan para dar nombres más descriptivos a valores de campos, son especialmente útiles cuando se trata de campos que contienen códigos numéricos.

Por ejemplo: El campo *ESTADO (Condición de Actividad)* donde:

- 0 = Entrevista individual no realizada.
- 1 = Patrón.
- 2 = Cuenta propia.
- 3 = Obrero/Empleado.
- 4 = Trabajador familiar sin remuneración.
- 9 = Ns. /Nr.

Las correspondencia de nombres que se crean son:

- *Analfabetismo:* para el campo elemento de 1, el valor es "Si sabe leer y escribir", para el campo elemento de 2, es "No" y para el campo elemento de 3, es "Menor de 2 años".

- *Asist. Educativa:* para el campo elemento de 1, el valor es "Si, asiste a algún establecimiento educativo (colegio, escuela, universidad)", para el campo elemento de 2, es "No asiste , pero asistió" y para el campo elemento de 3, es "Nunca asistió".

- *Categoría de Inactividad:* para el campo elemento de 1, el valor es "Jubilado/Pensionado", para el campo elemento de 2, el valor es "Rentista", para el campo elemento de 3, el valor es "Estudiante", para el campo elemento de 4, el valor es "Ama de casa", para el campo elemento de 5, el valor es "Menor de 6 años", para el campo elemento de 6, el valor es "Discapacitado" y el 7, para el valor "Otros".

- *Categoría Ocupacional:* para el campo elemento de 1, el valor es "Patrón", para el campo elemento de 2, el valor es "Cuenta propia", para el campo elemento de 3, el valor es "Obrero o empleado", para el campo elemento de 4, el valor es "Trabajador familiar sin remuneración" y para el campo elemento de 9, el valor "Ns./Nr."

- *Cobertura Médica:* para el campo elemento de 1, el valor es "Obra social (incluye PAMI)", para el campo elemento de 2, el valor es "Mutual /Prepaga/Servicio de emergencia", para el campo elemento de 3, el valor es "Planes y seguros públicos", para el campo elemento de 4, el valor es "No paga ni le descuentan" , para el campo elemento de 9, el valor "Ns./Nr.", para el campo elemento de 12, el valor es "Obra social y Mutual /Prepaga/Servicio de emergencia", para el campo elemento de 13, el valor es "Obra social y Planes y Seguros Públicos", para el campo elemento de 23, el valor es "Mutual /Prepaga/Servicio de emergencia/Planes y Seguros Públicos", para 123, el valor es "Obra social, Mutual /Prepaga/Servicio de emergencia y Planes y Seguros Públicos".

- *Condición de Actividad:* para el campo elemento de 0, el valor es "Entrevista individual no realizada", para el campo elemento de 1, el valor es "Ocupado", para el campo elemento de 2, el valor es "Desocupado", para el campo elemento de 3, el valor es "Inactivo" y para el campo elemento de 4, el valor "Trabajador familiar sin remuneración" y para el 9 , el valor "Ns./Nr.".

- *Estado Civil:* para el campo elemento de 1, el valor es "Unido", para el campo elemento de 2, es "Casado" y para el campo elemento de 3, es "Separado/a o divorciado/a", para el campo elemento de 4, es "Viudo/a" y el campo elemento de 5 , es "Soltero/a".

- *Nivel Educativo:* para el campo elemento de 1, el valor es "Primaria Incompleta (incluye educación especial)", para el campo elemento de 2, el valor es "Primaria Completa", para el campo elemento de 3, el valor es "Secundaria Incompleta", para el campo elemento de 4, el valor es "Secundaria Completa", para el campo elemento de 5, el valor "Superior Universitaria Incompleta", para el campo elemento de 6, el valor es "Superior Universitaria Completa", para el campo elemento de 7, el valor es "Sin Instrucción" y para el campo elemento de 9 , el valor es "Ns./Nr.".

- *Sexo:* para el campo elemento de 1, el valor es "Varón" y para el 2, es "Mujer".

- *Región:* para el campo elemento de 01, el valor es "GBA", para el campo elemento de 40, es "Noroeste", para el campo elemento de 41, es "Nordeste", para campo elemento de 43, es "Pampeana" y para campo elemento 44, es "Patagonia".

A las variables de entrada se las relaciona con el campo que corresponda en la pestaña Parámetros de campo, columna Correspondencia de Nombres, en el Objeto de Datos (ver fig. 6.22 de la pág. 153).

Creación de la Base de Minería Al finalizar los anteriores pasos, se deben crear los *Objetos de Minería* , que no son más que funciones analíticas aplicadas a los datos.

Estos objetos generan *Objetos de Resultados* , que se pueden ver y analizar con las herramientas de visualización incorporadas a *Intelligent Miner Visualizer*. Los resultados se analizan en páginas posteriores, aquí simplemente se describen los *Objetos de Minería* que se crean.

Considerando el análisis de carácter exploratorio que se desea realizar, se utiliza la *Función de Clustering Demográfico*.

Los objetos que se generan son:

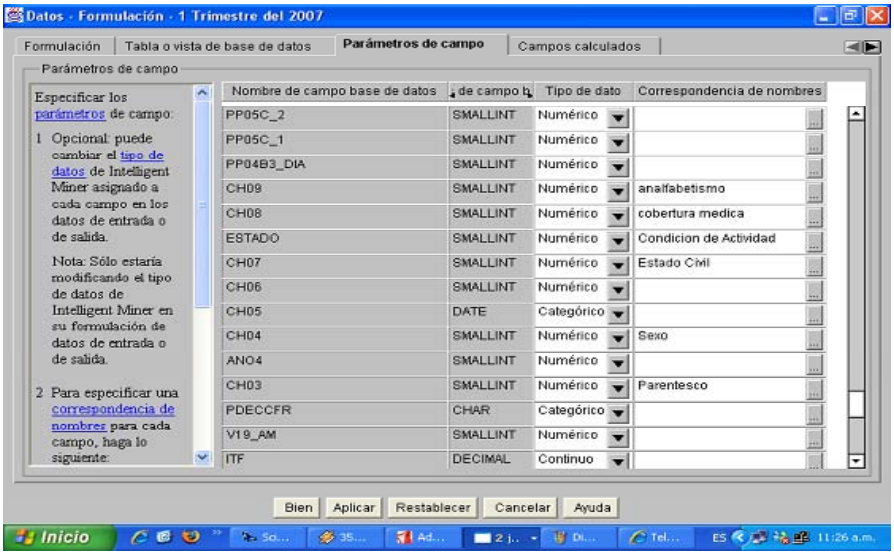

Figura 6.22: Selección de correspondencia de nombres, en la pestaña parámetros de campo.

- *PJ1_1=1 and Aglomerado =12.*

Clustering de los Perfiles de los Planes Jefes y Jefas de la Prov. de Ctes.

- *personas EPH.*

Contiene información de la tabla de personas de la tabla USB_T105 con datos, de la Base de Datos personas.

- *Datos de la EPH , con Ctes.*

Contiene información de las variables a trabajar, como así como también de Ctes.

- *Estudio de la Var CAES con respecto de la Población del NEA.*

Contiene información de la población del NEA.

Los mismos se pueden visualizar, cuando se presiona Base de minería, Abrir base de minería como se puede ver en la siguente fig. 6.23 de la pág. 154.

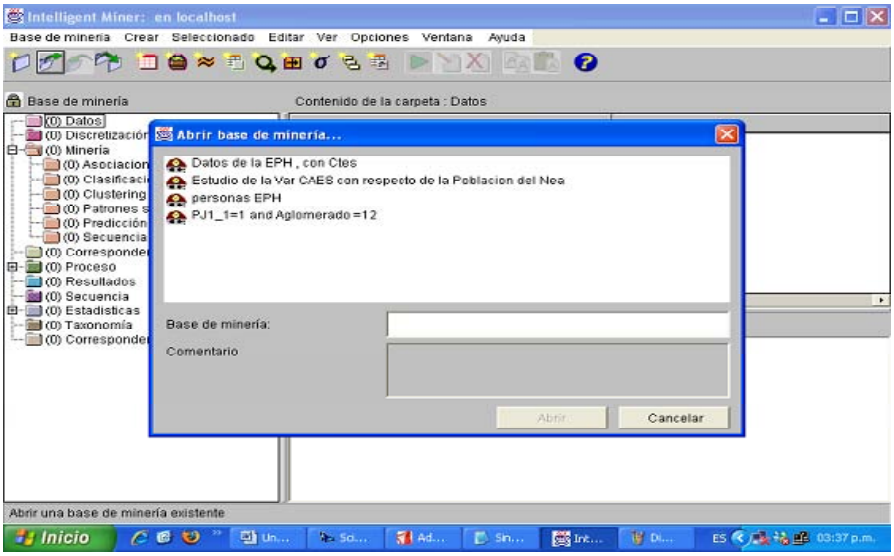

Figura 6.23: Visualización de las distintas bases de minería creadas en el Intelligent Miner.

Una vez seleccionada la base de minería a utilizar, se debe crear la función de minería propiamente dicha. Para ello se deben realizar los siguientes pasos:

- Selección de una *Función de Minería*.
- Selección de los *Datos de Entrada*.
- Especificación de *Parámetros*.
- Especificación de los *Campos de Salida*.
- Especificación del nombre de *Datos de Salida*.
- Especificación del nombre de *Resultado*.

Selección de una Función de Minería Para seleccionar una *Función de Minería*, se debe escoger una de ellas en la listas de *Función de Minería* disponibles.

Las funciones de minería disponibles, como se puede ver en la fig. 6.24 de la pág. 155, son las siguientes:

- *Asociación.*
- *Clasificación - Árbol.*
- *Clasificación - Neuronal.*
- *Clustering - Demográfico.*
- *Clustering - Neuronal.*
- *Patrones secuenciales.*
- *Predicción - Función base radial.*

La que se utilizará con más frecuencia en este apartado es la de *Clustering - Demográfico.*

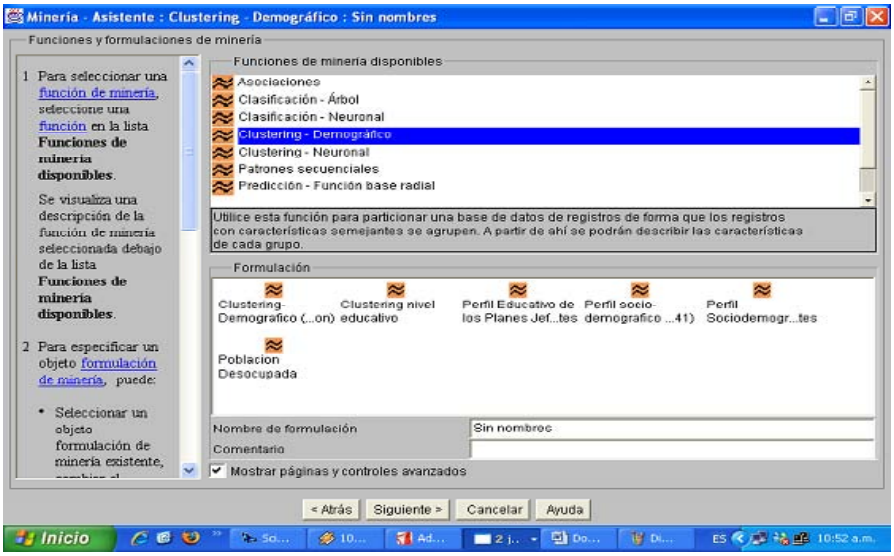

Figura 6.24: Selección de la función de minería, *Clustering - Demográfico.*

Selección de los Datos de Entrada Como se puede ver en la siguiente fig. 6.25 de la pág. 156 el *Intelligent Miner for Data*, nos permite seleccionar los Datos de entrada, ya sea mediante los Datos de entrada disponibles o caso contrario se pueden crear Datos de entrada, presionando el botón Crear datos.

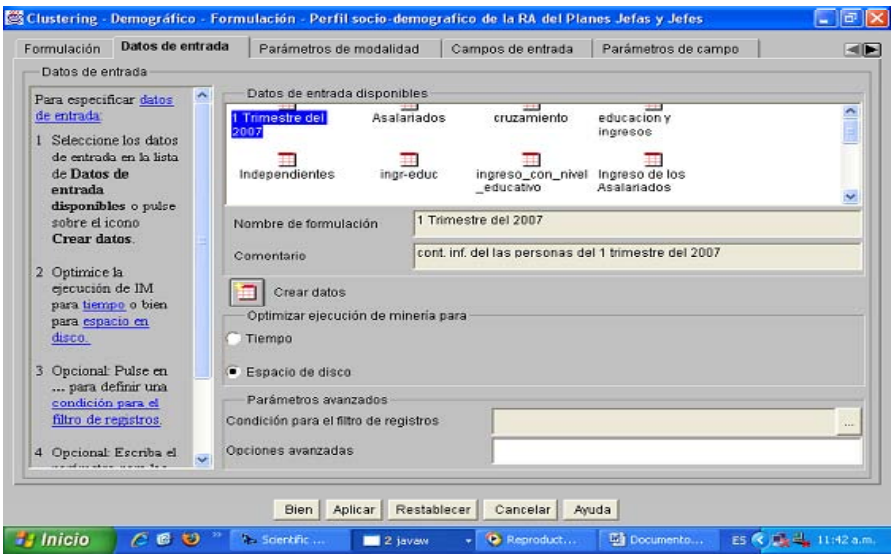

Figura 6.25: Selección de los Datos de entrada, *1 Trimestre del 2007.*

Una vez seleccionados estos, se deberá realizar las Especificaciones de los *Parámetros.*

Especificación de Parámetros En la sección de Parámetros de modalidad (ver fig. 6.26 de la pág. 157), se pueden realizar numerosas modificaciones como ser las Pasadas máximas. Estas maximizan el número de veces que la función se aplica sobre los datos de entrada.

En este caso de estudio, en las secciones Especificación de *parámetros* y Especificación de los *campos de salida,* no se realizan modificaciones.

Campos de Entrada Los campos de entrada son campos de datos que una función de minería utiliza para su posterior procesamiento (ver la fig. 6.27 de la pág. 157).

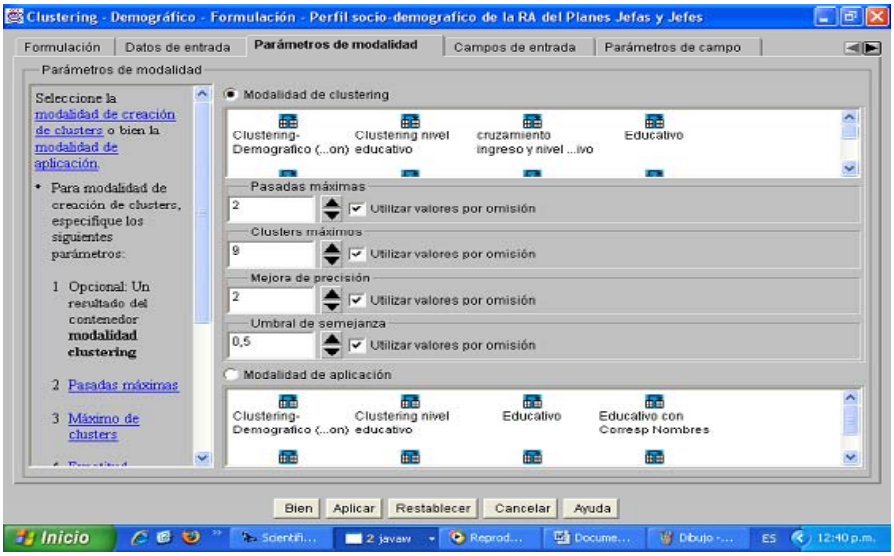

Figura 6.26: Especificación de los parámetros de modalidad.

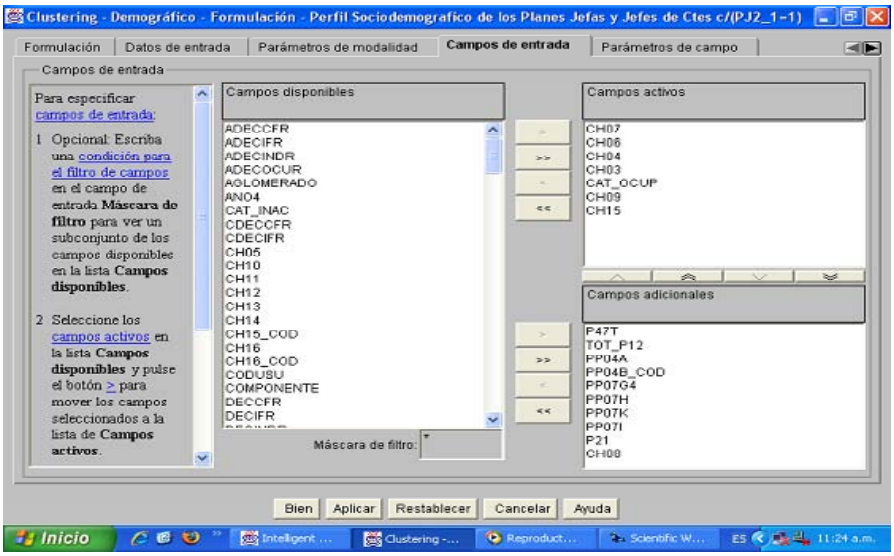

Figura 6.27: Selección de los campos de entrada (Campos activos y Campos adicionales).

Objetos de Resultados En este apartado se expondrán todos los resultados que proporcione el *Intelligent Miner Visualizer*. Esta herramienta permitirá visualizar, analizar y hasta explorar los resultados obtenidos en cada ejecución.

Al ejecutar la función de minería, como se puede ver en la fig. 6.28 de la pág. 158, el *Intelligent Miner* proveerá de la siguiente información: hora de inicio, tiempo transcurrido, información de estado adicional y criterio de *condorcet*. Este último permitirá encontrar la optimización en la seleccion de los Cluster.

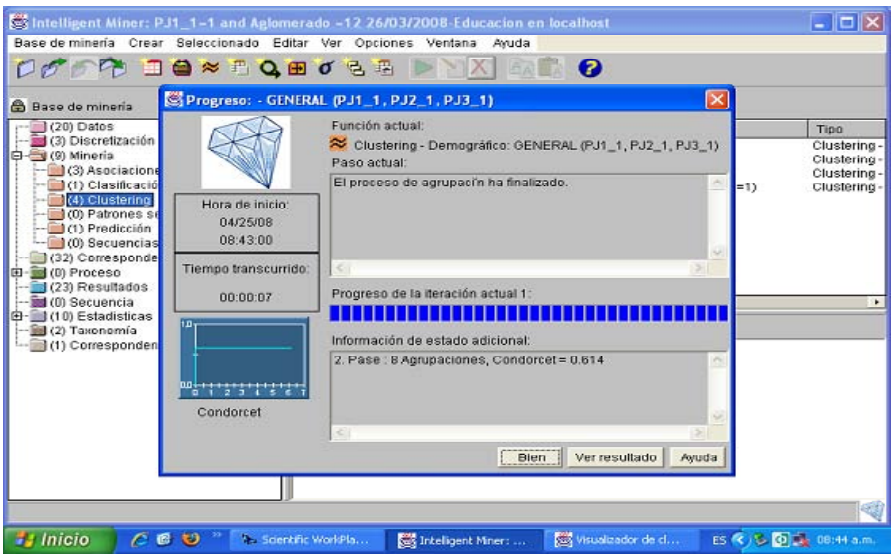

Figura 6.28: El criterio de condorcet es de 0.614 (donde aceptable es 0,65).

Como se puede observar en fig. 6.28 de la pág. 158, el criterio de *Condorcet* toma el valor 0.614. Esta información es suficiente para determinar que la similitud de los registros dentro de cada cluster es excelente dado que un valor mínimo usual que se considera aceptable es 0,65.

Esto no implica que no se puede obtener mejores resultados seleccionando las variables de entrada.

Al visualizar los objetos de resultados (ver fig. 6.29 de la pág. 159) se nota la existencia de 8 clusters identificados por la ejecución de minería. En cada clúster, los diagramas y gráficos de barras representan los campos activos y

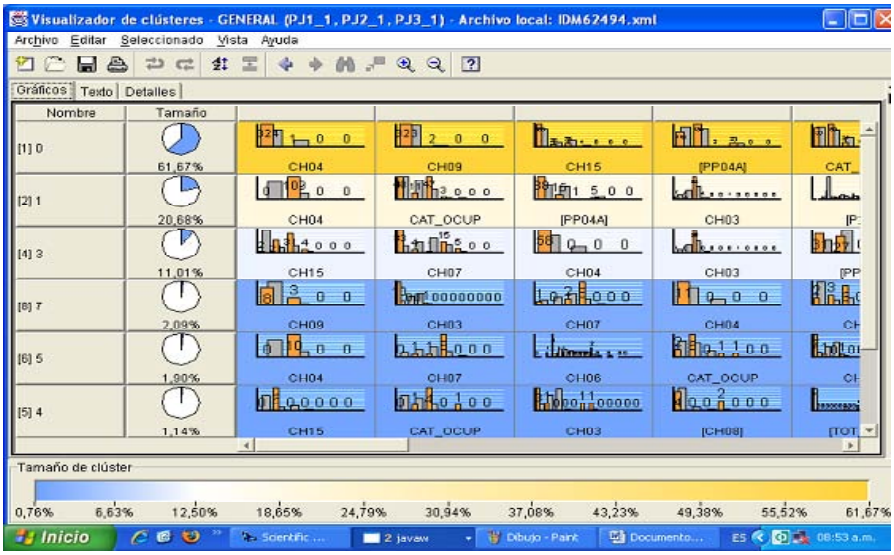

Figura 6.29: Intelligent Miner nos provee los resultados mediante *Visualizador de clústeres.*

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.)*.

La primera columna contiene el nombre y el ID del cluster, la siguiente representa el tamaño de cluster en porcentaje con respecto a la muestra; por ejemplo: el cluster superior representa un 61,67% de los datos, el siguiente un 20,68%, el siguiente un 11,01% y así sucesivamente.

En este caso prácticamente un 93,36% de la población está representada sólo por estos tres primeros clústeres, dividiéndose el 6,64% restante entre los demás.

Al contemplar la figura 6.30 de la pág. 160 se obtienen las siguientes conclusiones.

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 cónjugue (ver fig. 6.31 de la pág.

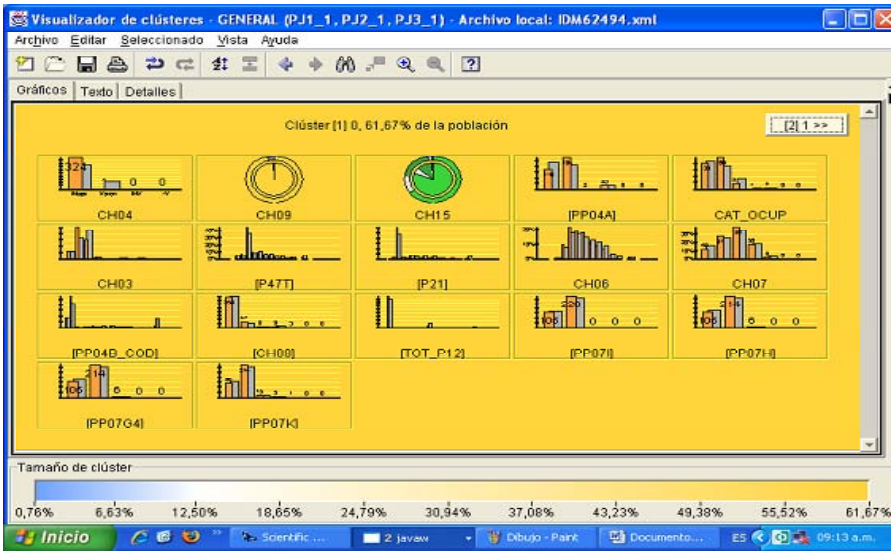

Figura 6.30: Visualización general del Clúster Nº1 de 61,67% de la población total.

161).

Con respecto a lo laboral, estás personas trabajan en hogares privados como servicio doméstico (ver fig. 6.32 de la pág. 161), donde no paga ni le descuentan mensualmente una cobertura médica como se puede ver fig. 6.33 de la pág. 162, tampoco tiene contrato de trabajo ni obra social y mucho menos descuento jubilatorio (ver fig. 6.34 de la pág. 162), respectivamente (ver fig. 6.35 de la pág. 163).

El ingreso total individual predominantemente de estas personas esta en promedio entre los 100 a 200 pesos (ver fig. 6.36 de la pág. 163), y es de 0 pesos el ingreso proveniente de otras actividades (ver fig. 6.72 de la pág. 183).

En la segunda agrupación, del 20,68% de la población total, se puede observar que el sexo predominantemente es el masculino (ver fig. 6.38 de la pág. 164).

Sin diferenciarse con el primer clúster, en este en su mayoría siguen siendo de esta localidad o sea Corrientes como se puede apreciar en la fig. 6.39 de la pág. 165, con un estado civil de viudo/a y con una edad sobresaliente de 46 años (ver fig. 6.40 de la pág. 166) , respectivamente (ver fig. 6.40 de la pág.

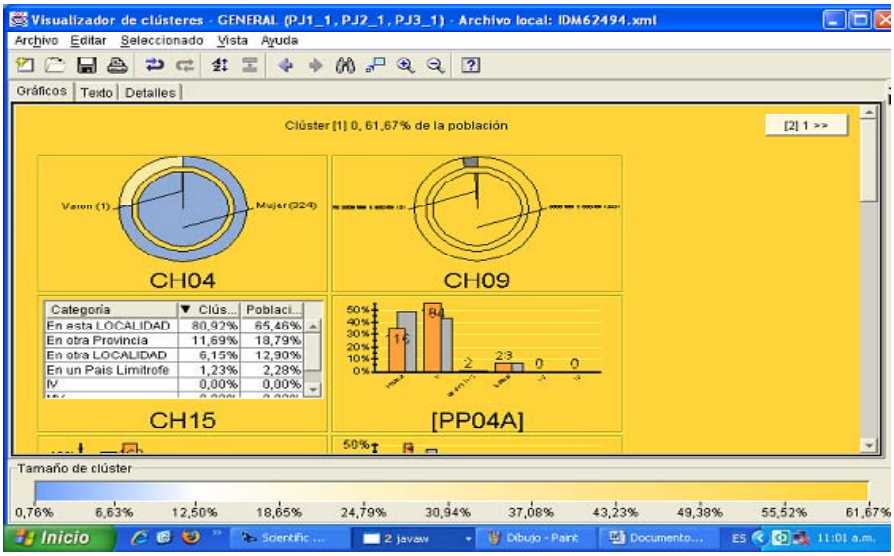

Figura 6.31: Visualización de las variables CH04 (sexo), CH15 (¿Donde nació?).

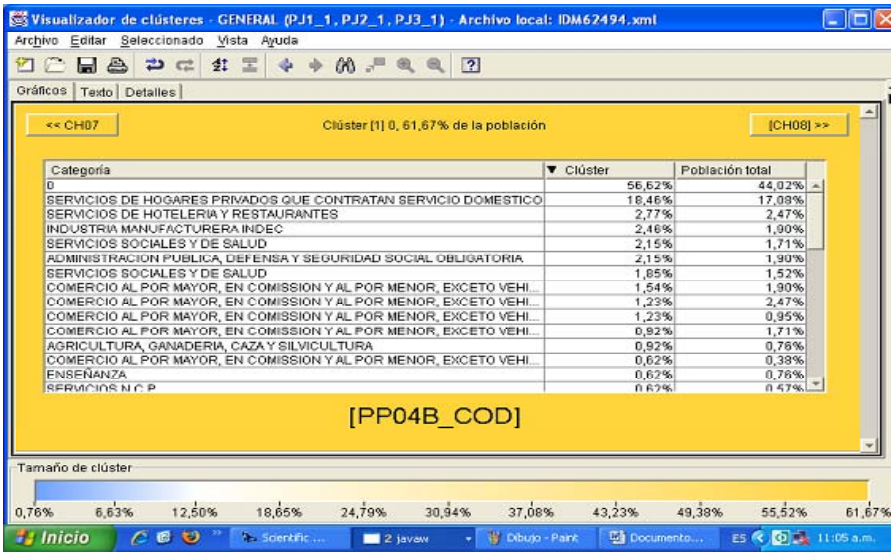

Figura 6.32: Visualización, del contenido de la variable PP04B_COD (Clasificación de Actividades Económicas para Encuestas Socioeconómicas *CAES*).

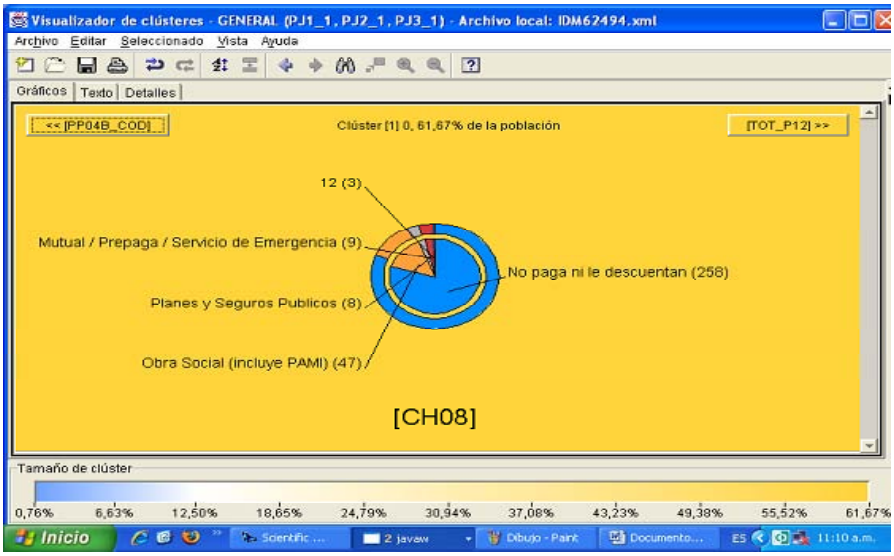

Figura 6.33: Muestreo del contenido de la variable CH08 (¿Tiene algún tipo de cobertura médica por la que paga o le descuentan?).

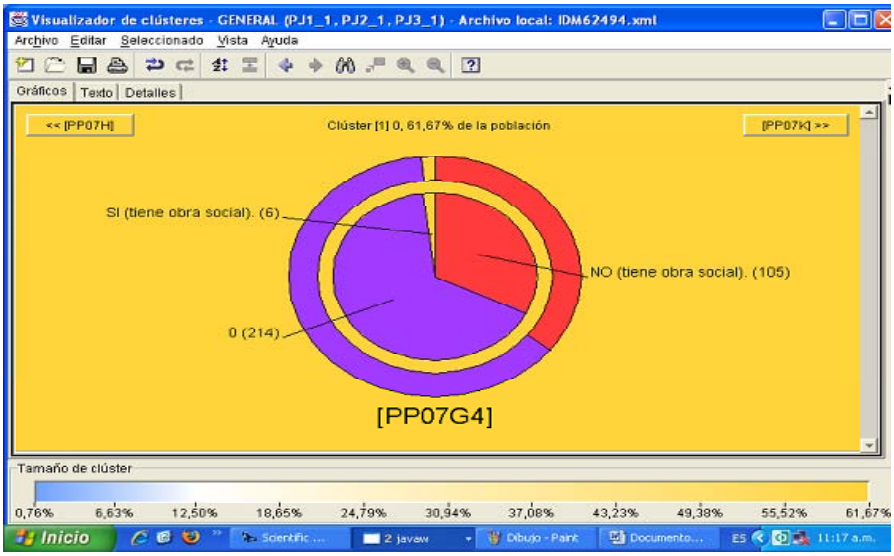

Figura 6.34: En el resultado de la variable PP07G4 (obra social) se puede observar que en su gran mayoría estas personas no la poseen.

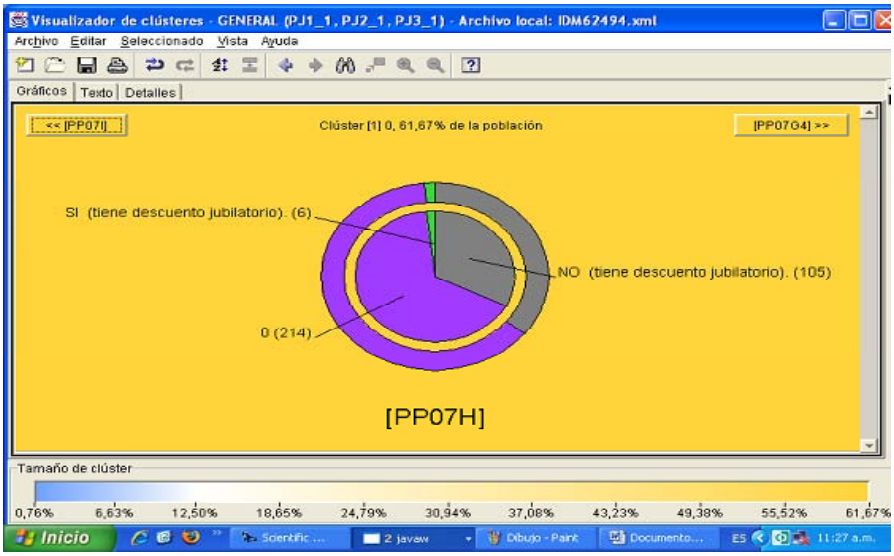

Figura 6.35: Visualización del resultado de la variable PP07H (¿Por ese trabajo tiene descuento jubilatorio?).

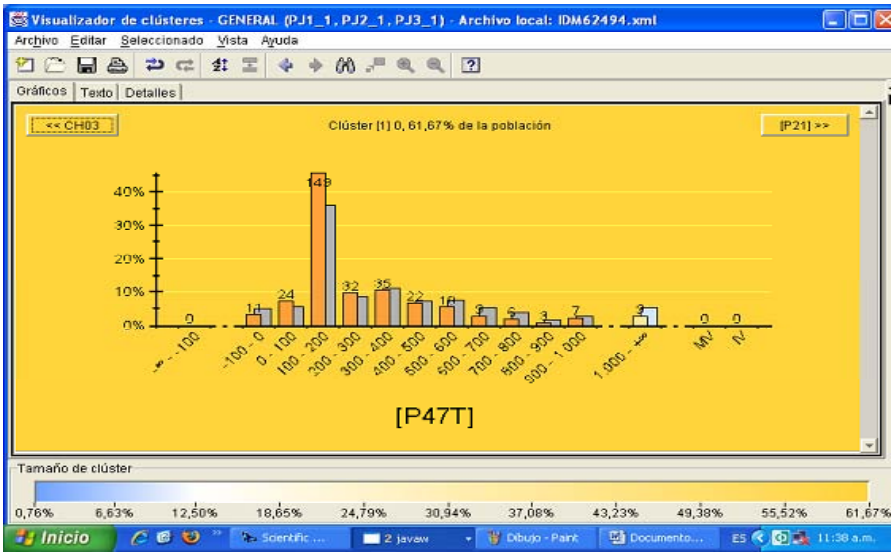

Figura 6.36: El monto del ingreso total individual de estas personas esta entre los 100 a 200 pesos.

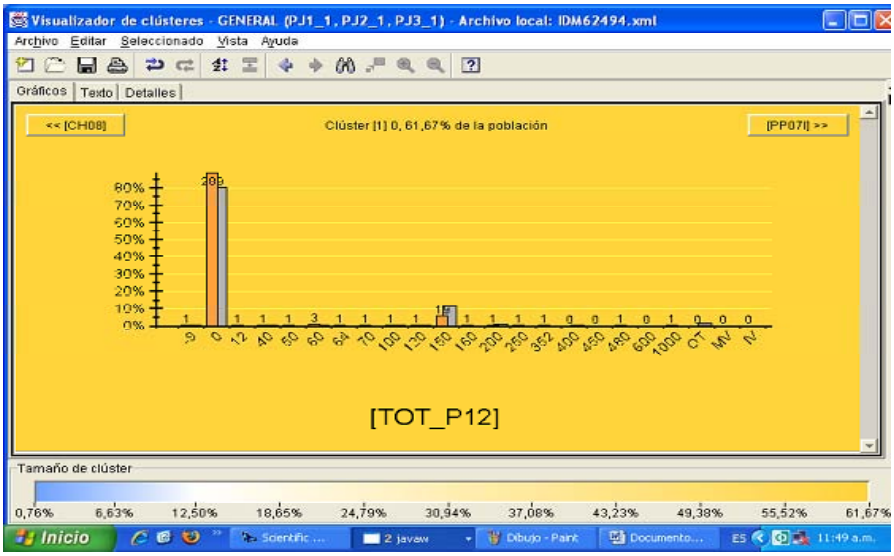

Figura 6.37: Visualización, de la variable TOT_P12 (ing. de otras ocupaciones).

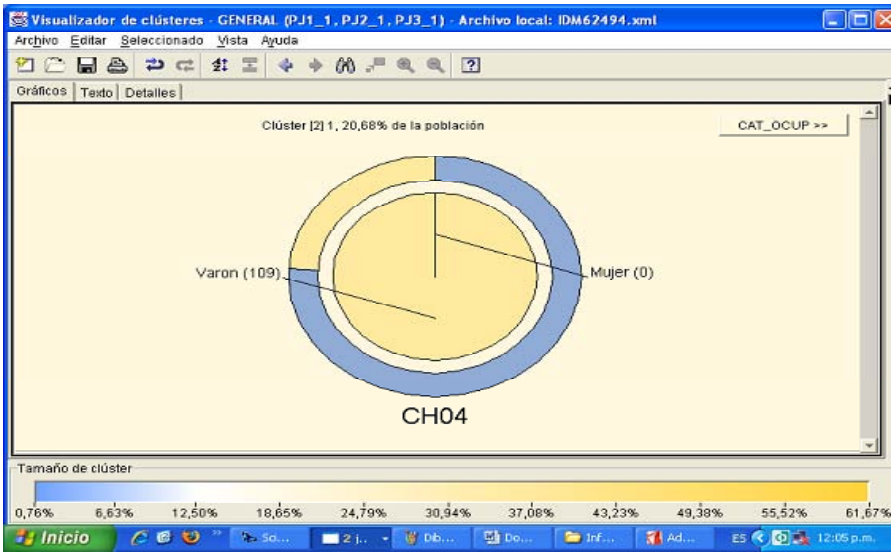

Figura 6.38: En el segundo clúster, del 20,68% de la población total se puede apreciar al sexo masculino como el predominante.

166).

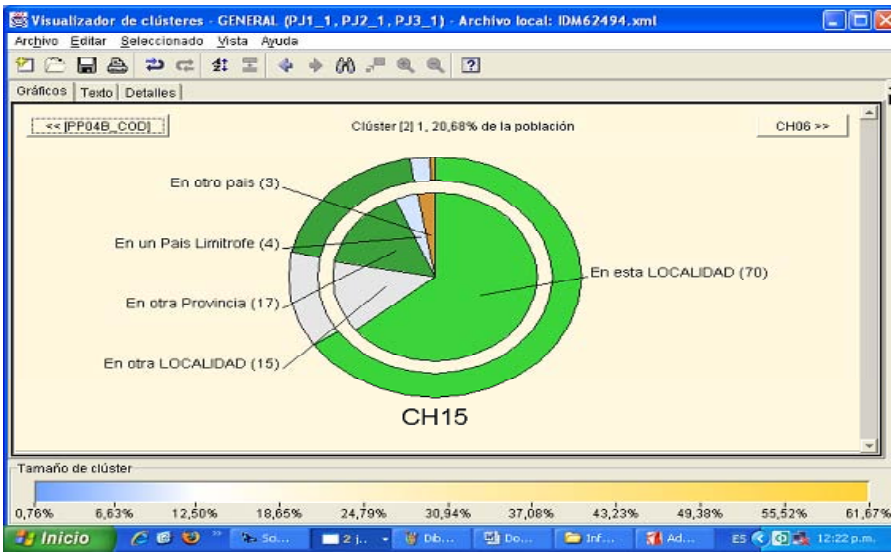

Figura 6.39: La opción "en esta localidad" de la variable CH15 (¿Dónde nació?) sigue siendo la predominante.

En la categoría ocupacional el rubro predominante se lo puede visualizar en la fig. 6.41 de la pág. 166 como el rubro de *" obrero o emplea".*

La actividad económica que resulta ser predominante es la *construcción* (ver fig. 6.42 de la pág. 167).

En relación a lo laboral se puede decir que estas personas no poseen cobertura medica, obra social, ni tampoco aportes jubilatorios e incluso no realizan aportes por sí mismos, todo esto se pude comprobar en las siguientes figuras: (ver fig. 6.43 de la pág. 167), (ver fig. 6.44 de la pág. 168), (ver fig. 6.45 de la pág. 168), respectivamente en la fig. 6.46 de la pág. 169.

El tipo de contrato, con la opción *no le dan ni le entregan nada* cuando el empleado recibe sus haberes es la opción más frecuente, como puede verse en la fig. 6.47 de la pág. 169.

El ingreso de la ocupación principal (P21) contiene al rango 0 al 50 pesos como el predominante en dicha variable (ver fig. 6.48 de la pág. 170).

El ingreso total individual (P47T) contienen la misma distribución de los

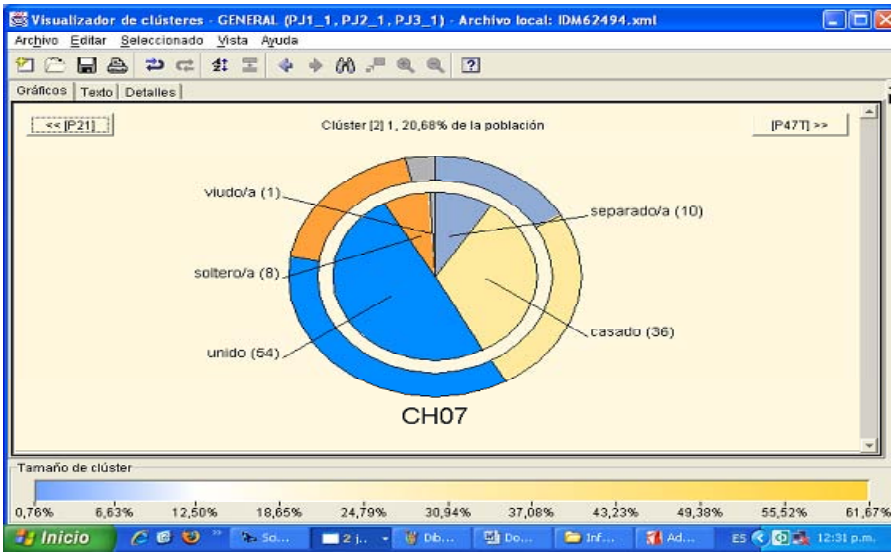

Figura 6.40: Visualización de la variable CH07(estado civil).

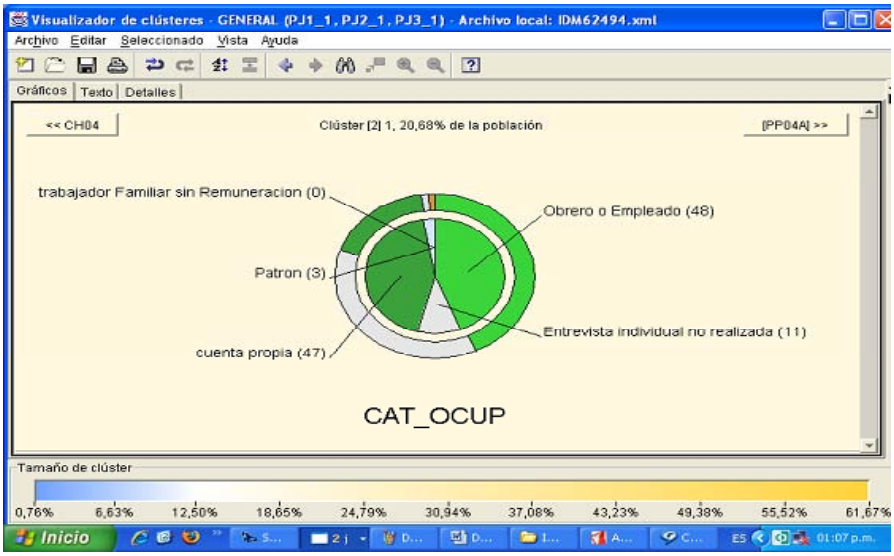

Figura 6.41: Visualización de las variables CAT_OCUP(categoría ocupacional).

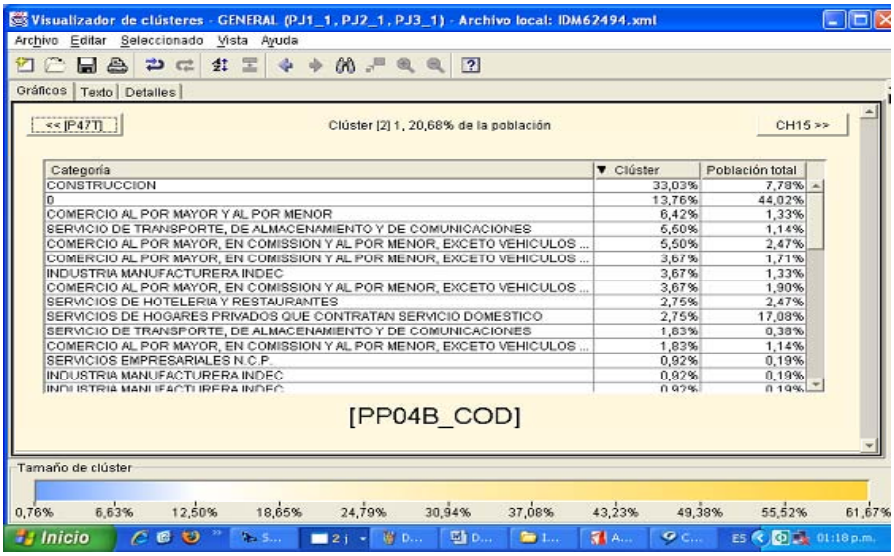

Figura 6.42: La variable PP04B_COD (rubro de las actividades económicas para el MERCOSUR).

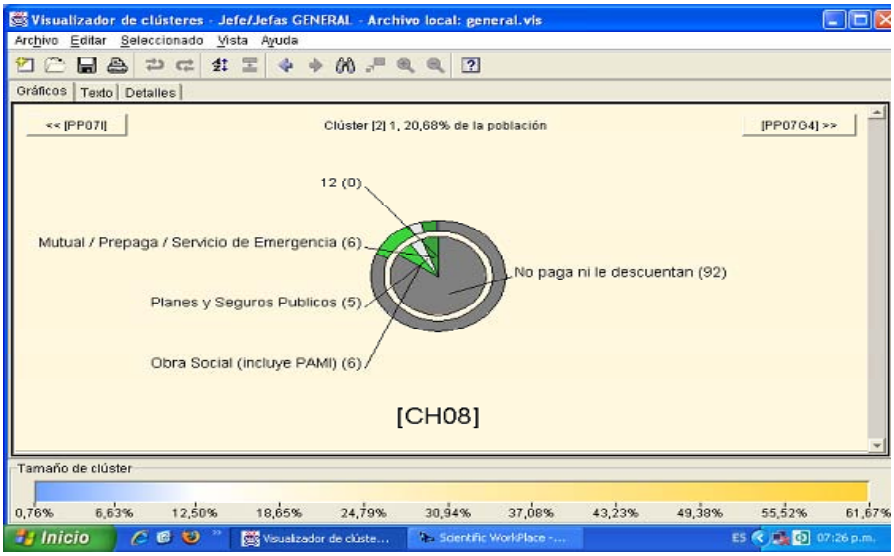

Figura 6.43: Visualización de la variables CH08 (cobertura medica).

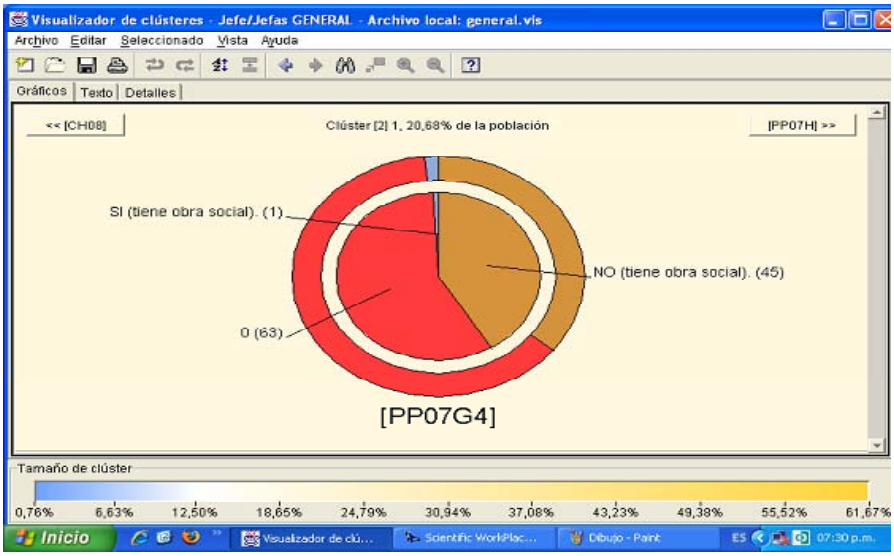

Figura 6.44: Visualización del diagrama circular de la variable PP07G4 (obra social).

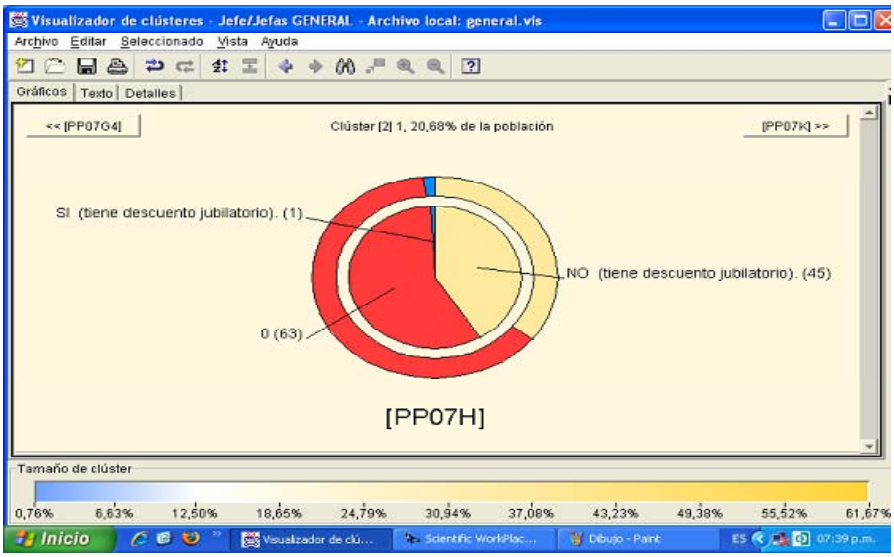

Figura 6.45: La opción “No tienen descuento jubilatorio” es la predominante en la variable PP07H (¿Por ese trabajo tiene descuento jubilatorio?).

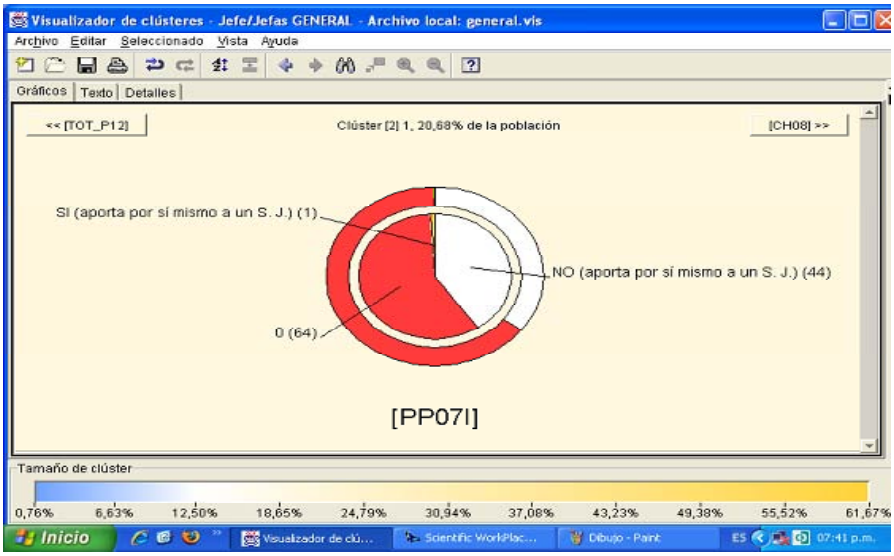

Figura 6.46: Resultado en formato de diagrama circular de la variable PP07I (¿Aporta por sí mismo a algún sistema jubilatorio?).

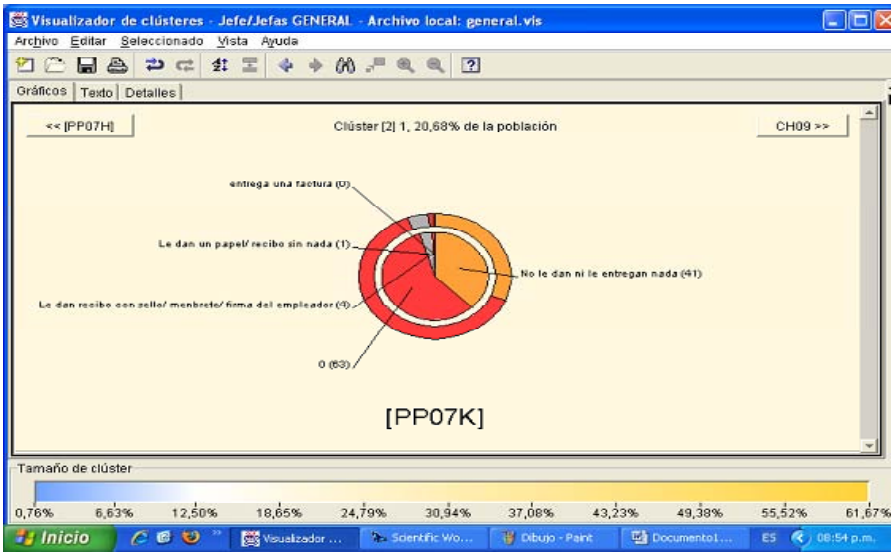

Figura 6.47: Muestreo del diagrama circular de la variable PP07K.

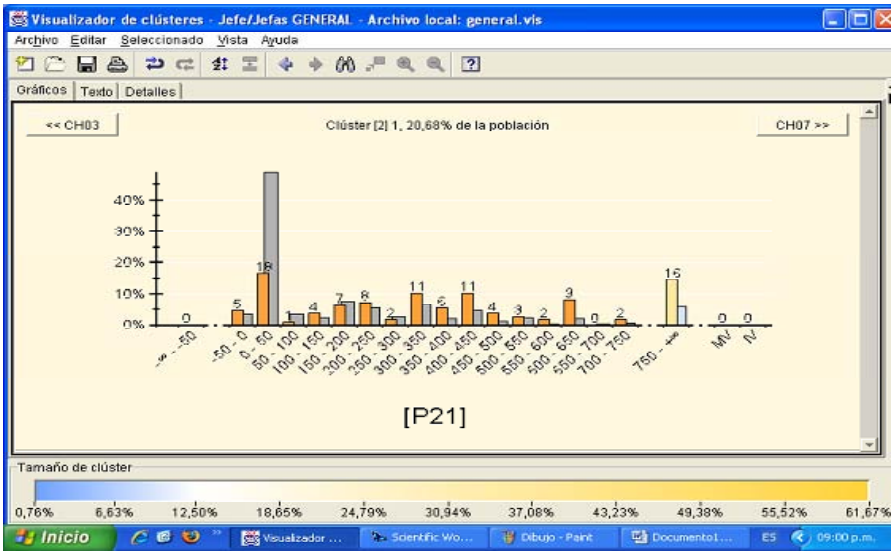

Figura 6.48: Visualización de la variable P21 (Monto del ingreso de la ocupación principal).

ingresos como puede verse en la fig. 6.49 de la pág. 171.

El monto de ingreso de otras ocupaciones (TOT_P12), resulta ser predominantemente 0 pesos (ver fig. 6.50 de la pág. 171).

El tercer cluster es de 11,01 % de la población total, 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 (ver fig. 6.51 de la pág. 172), (ver fig. 6.52 de la pág. 172), respectivamente (ver fig. 6.53 de la pág. 173).

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"* (ver fig. 6.65 de la pág. 180), (ver fig. 6.55 de la pág. 174).

También en este grupo el empleo en negro sigue siendo el predominante (ver fig. 6.56 de la pág. 174).

Lo mismo sucede con la variable PP07G4 (obra social) ya que estas personas no la poseen (ver fig. 6.57 de la pág. 175).

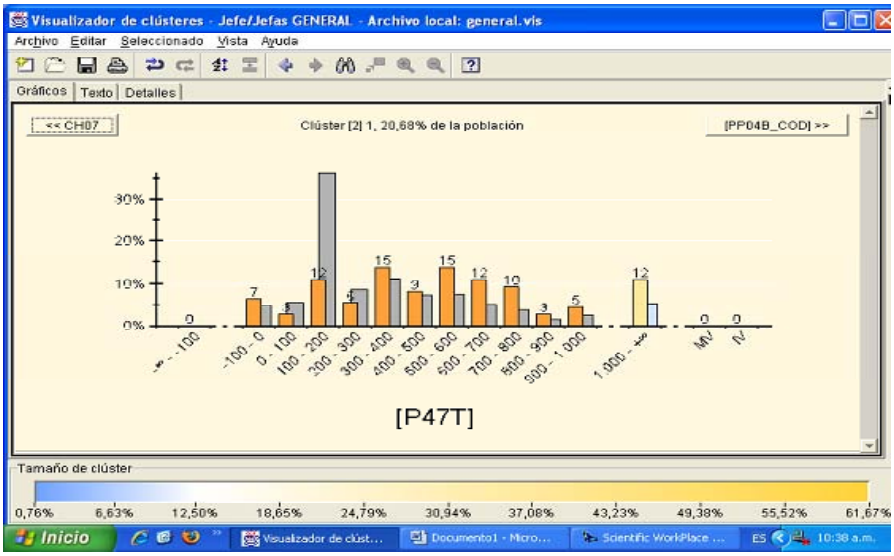

Figura 6.49: Visualización de la variable P47T (Monto del ingreso total individual).

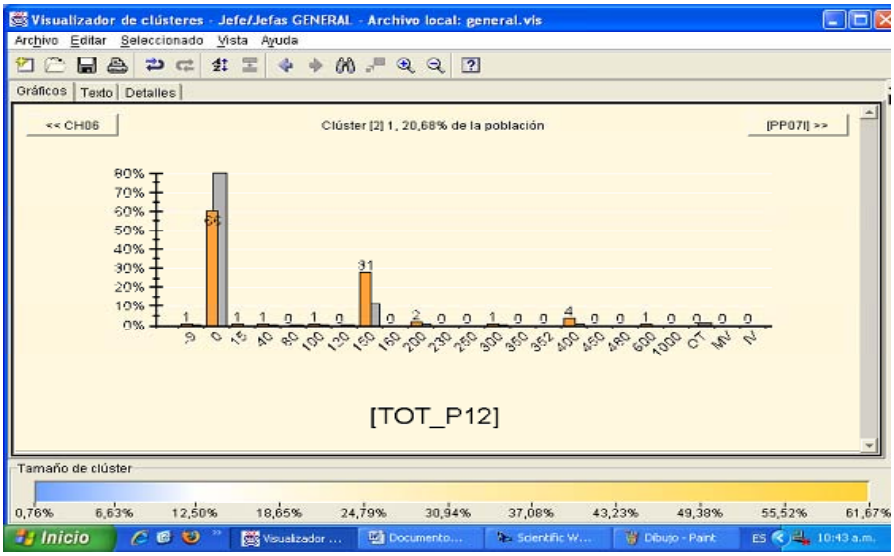

Figura 6.50: El contenido de la variable TOT_P12, demuestra el predominio 0 pesos.

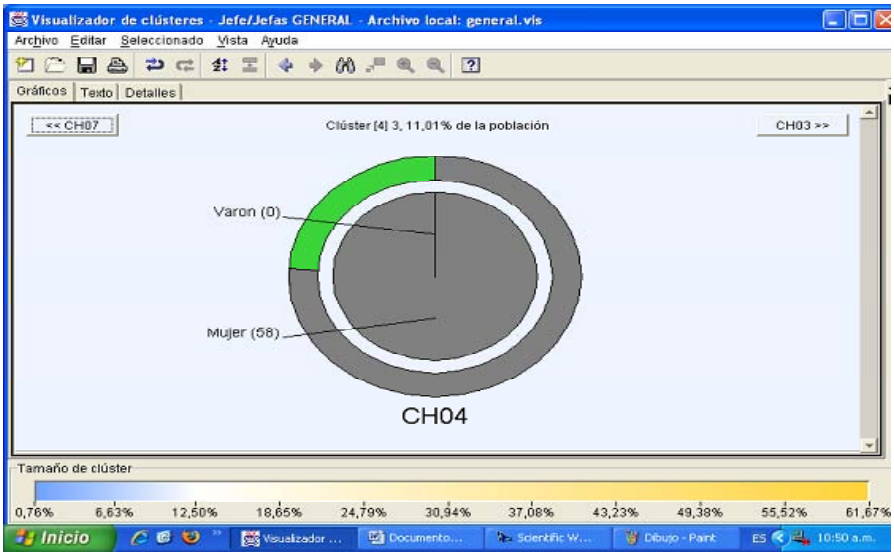

Figura 6.51: El sexo femenino es el predominante en el Clúster Nº3 (11,01 de la población total).

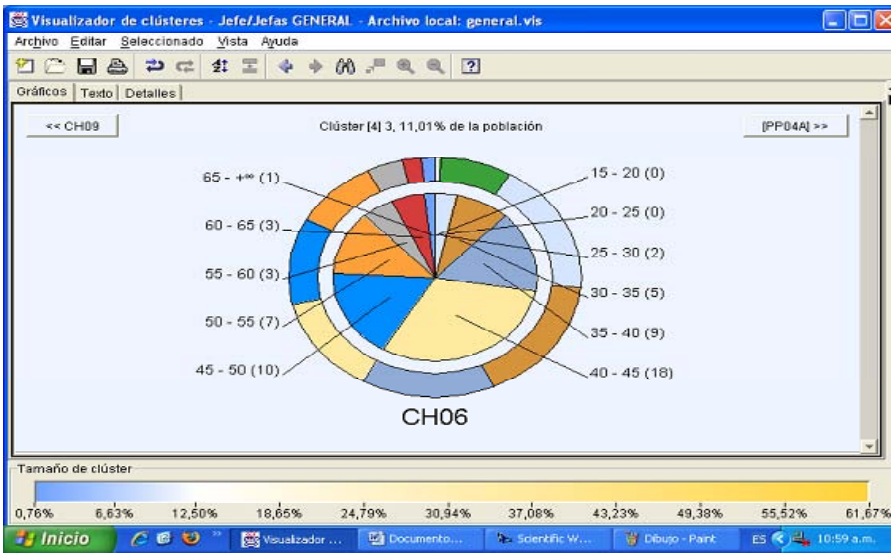

Figura 6.52: En este diagrama circular se puede observar que el rango de edad con mayor frecuencia es el [40-45].

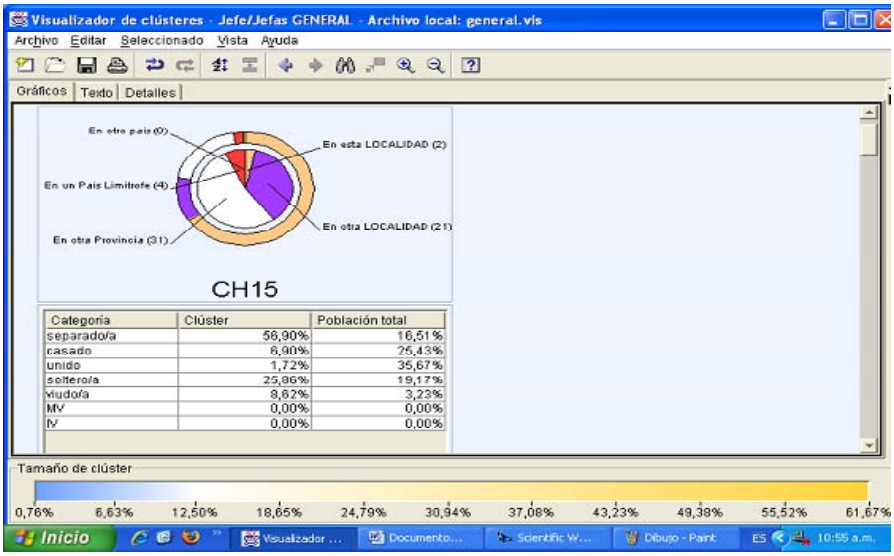

Figura 6.53: Visualización de las siguientes variables: CH15 (¿Dondé nació?) y CH07 (Estado Civil).

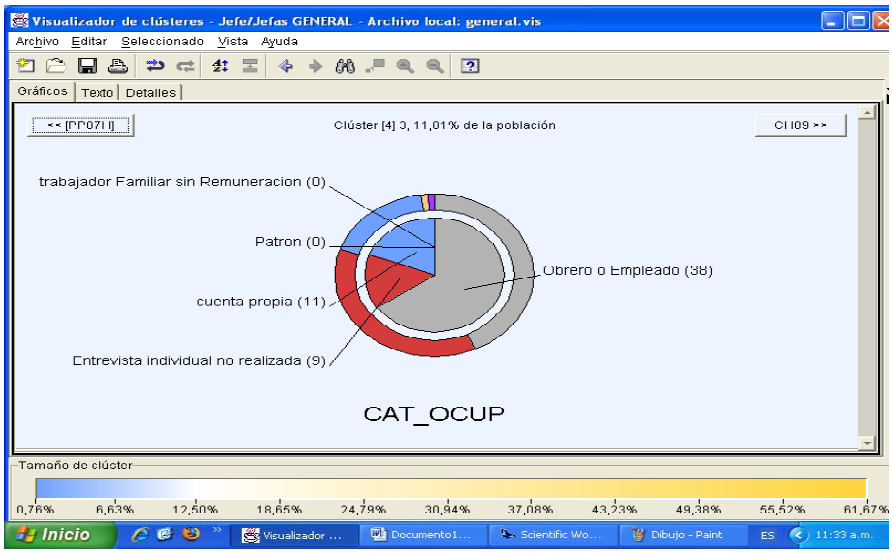

Figura 6.54: Diagrama circular de la variable CAT_OCUP (categoría ocupacional).

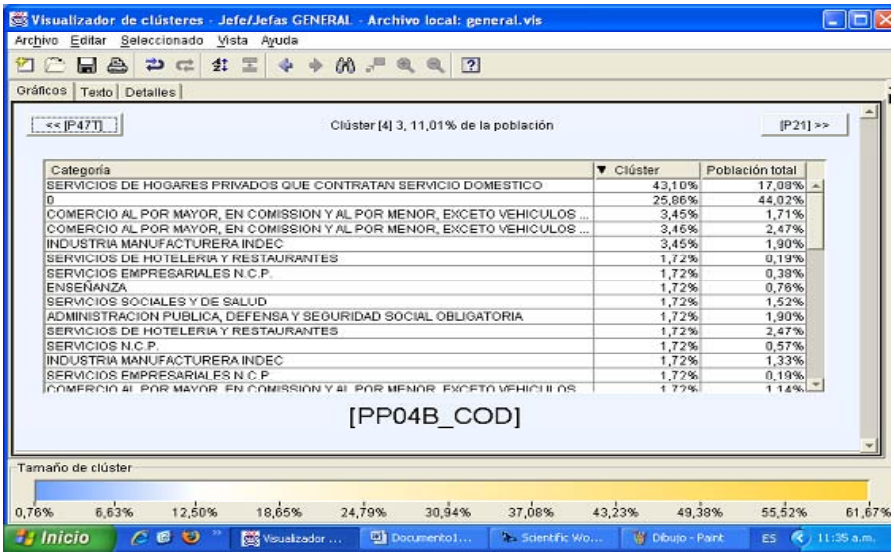

Figura 6.55: Visualización del resultado en formato tabla de la variable PP04B_ COD (rubro de actividades económicas).

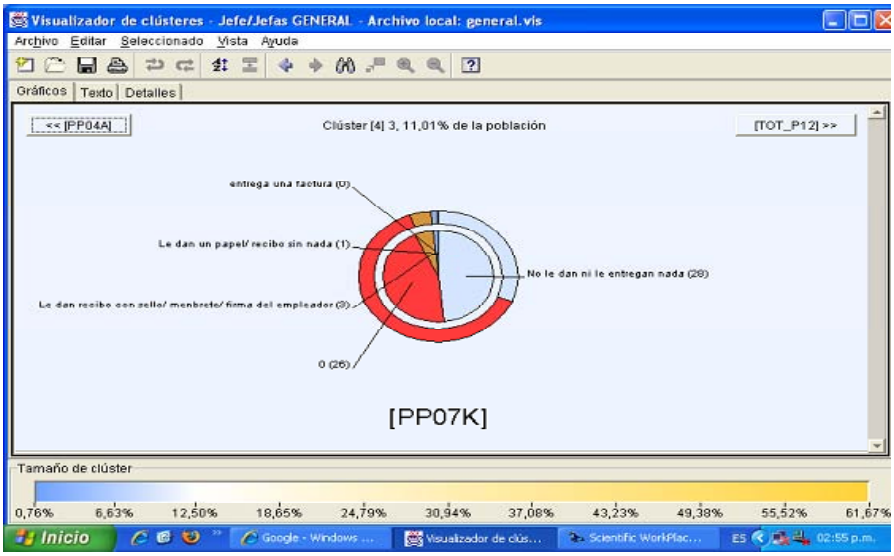

Figura 6.56: El tipo de contrato en negro, es el de mayor presencia en la variable PP07K.

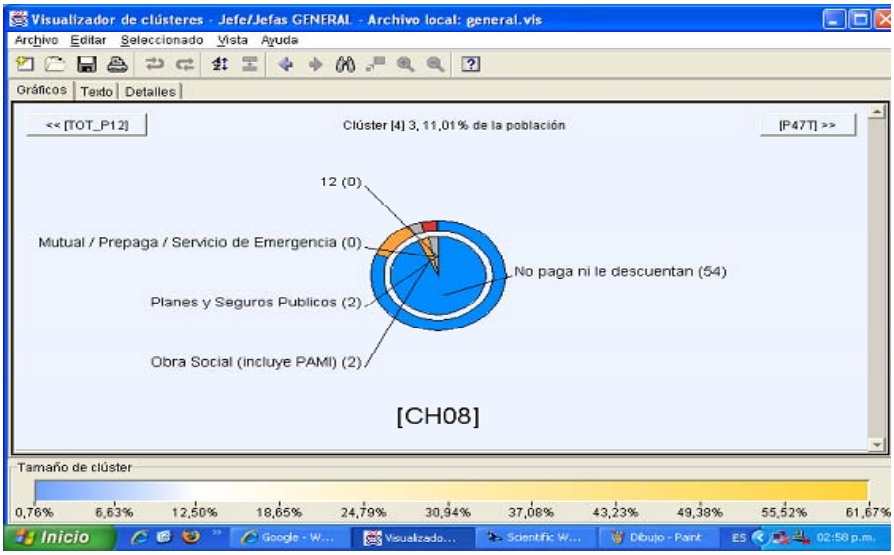

Figura 6.57: Visualizacíon de la variable CH08 (obra social).

Tampoco cuentan con un descuento jubilatorio con puede verse en la fig. 6.58 de la pág. 176.

Y mucho menos un aporte a algún sistema jubilatorio (ver fig. 6.59 de la pág. 176).

El nivel de analfabetismo también se lo tiene en cuenta para agrupar los distintos perfiles de los individuos. Éste está contemplado en la variable CH09 (sabe leer y escribir), como se puede ver en la fig. 6.60 de la pág. 177.

El cuarto cluster también contiene un 2,09 % población total.

En esta agrupación el sexo predominante es el femenino, con unos 25 a 30 años de edad y con un estado civil de soltero/a (ver fig. 6.61 de la pág. 177), (ver fig. 6.62 de la pág. 178) y (ver fig. 6.63 de la pág. 178).

Además, éstas ha nacido en otra provincia como se puede ver en la fig. 6.64 de la pág. 179.

En cuanto a lo laboral, esta persona tiene una categoría de ocupación de obrero o empleado como puede visualizarse en fig. 6.65 de la pág. 180 perteneciendo al rubro de actividades económicas de “Servicios de Asociaciones”

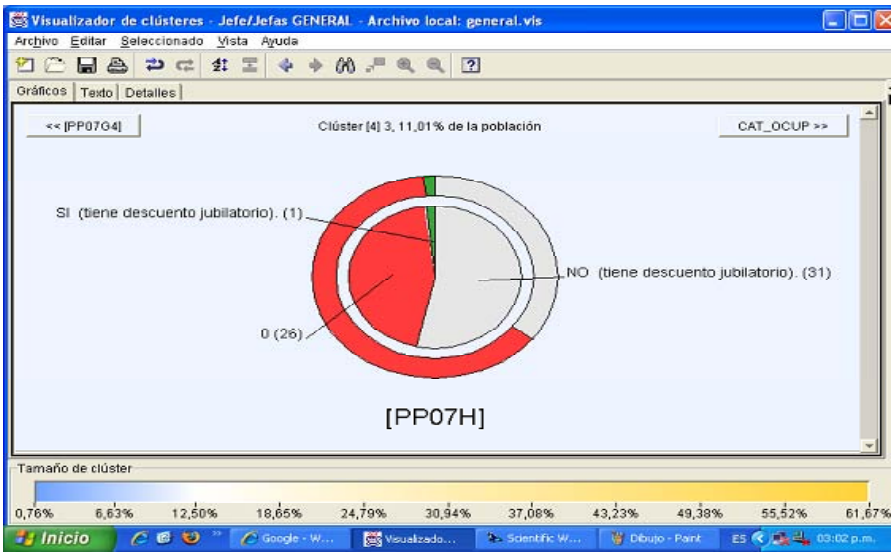

Figura 6.58: Muestreó de la variable PP07H (si tiene descuento jubilatorio).

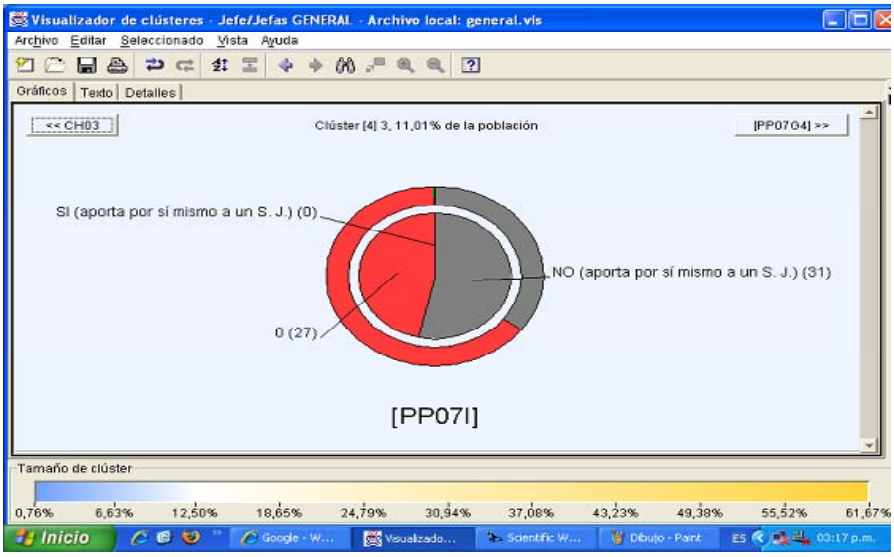

Figura 6.59: El aporte individual a algún sistema jubilatorio es nulo.

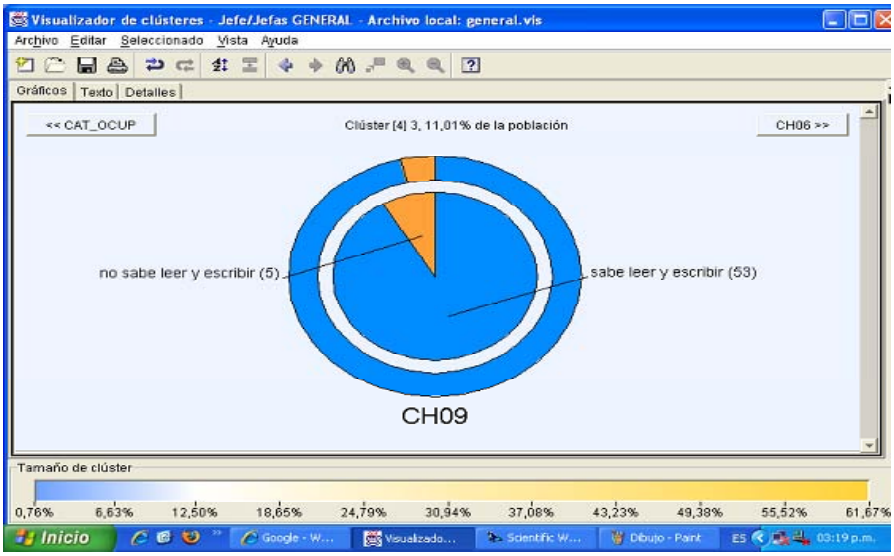

Figura 6.60: Visualización de la variable CH09 (sabe leer y escribir).

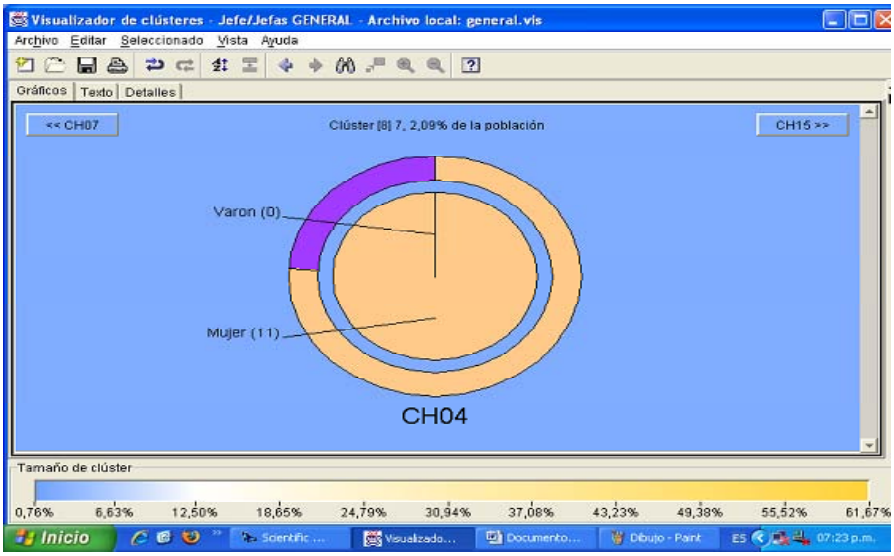

Figura 6.61: El sexo femenino es el predominante en la cuarta agrupación (2,09 % de la población total).

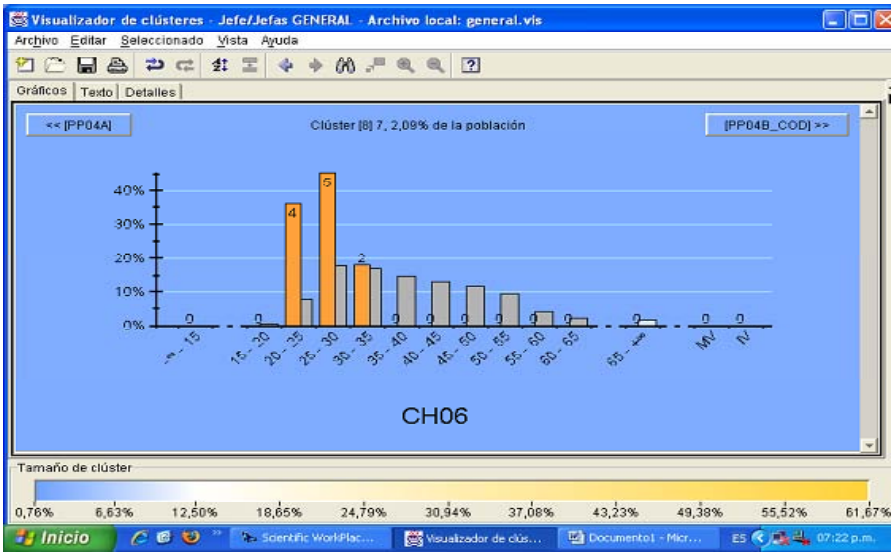

Figura 6.62: Visualización de la variable CH06 (años de edad).

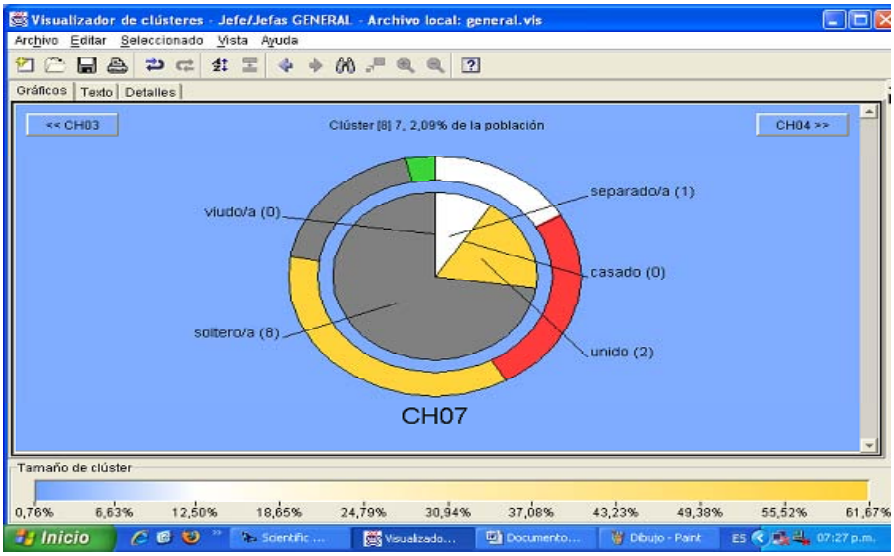

Figura 6.63: La opción *"soltero / a"* es la más frecuente en la variable CH07 (estado civil).

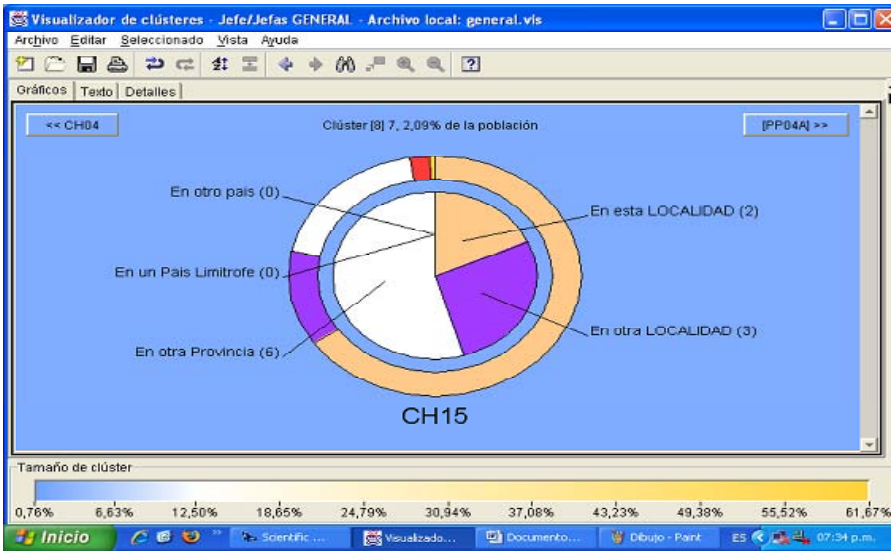

Figura 6.64: Resultado en formato de diagrama circular de la variable CH15 (¿Donde nació?).

(ver fig. 6.64 de la pág. 179).

El tipo de contrato de este trabajo, como se puede observar no varía con respecto a los anteriores clústeres, ya que sigue siendo predominante el trabajo en negro (ver fig. 6.67 de la pág. 181).

En cuanto a las condiciones laborales se puede decir que no tiene obra social (ver fig. 6.68 de la pág. 181).Tampoco descuento jubilatorio como se puede apreciar en la fig. 6.69 de la pág. 182, no obstante esta persona aporta por sí mismo a un sistema jubilatorio (ver fig. 6.70 de la pág. 182).

El ingreso de la ocupación principal no supera los pesos 350 como se puede observar en la fig. 6.71 de la pág. 183.

El monto de ingresos de otras ocupaciones, que incluye ocupación secundaria, ocupación previa a la semana de referencia, duedas / retroactivos por ocupaciones anteriores al mes de referencia, etc., se puede ver en la fig. 6.72 de la pág. 183; teniendo luego al ingreso total individual que es la sumatoria de ingresos laborales y no laborales con un monto que varia de los 100 a los 200 pesos (ver fig. 6.73 de la pág. 184).

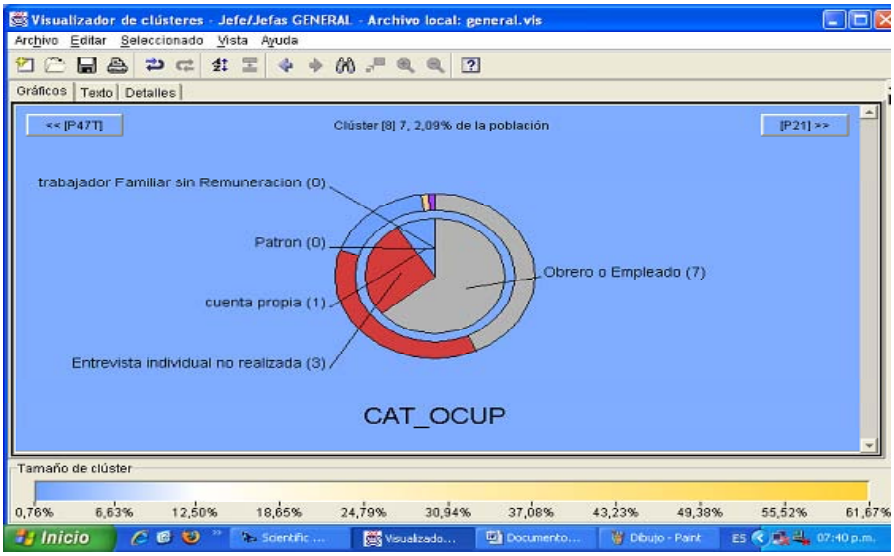

Figura 6.65: Visualización de la variable CAT_OCUP (Categoría Ocupacional).

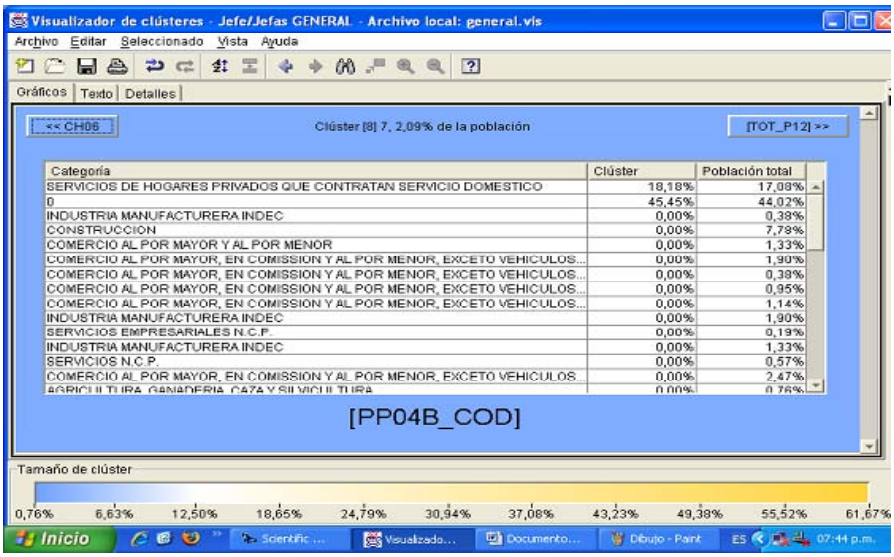

Figura 6.66: La categoría "servicios de hogares privados que contratan servicio domestico" es la opción con más frecuencia en la variable PP04B_COD.

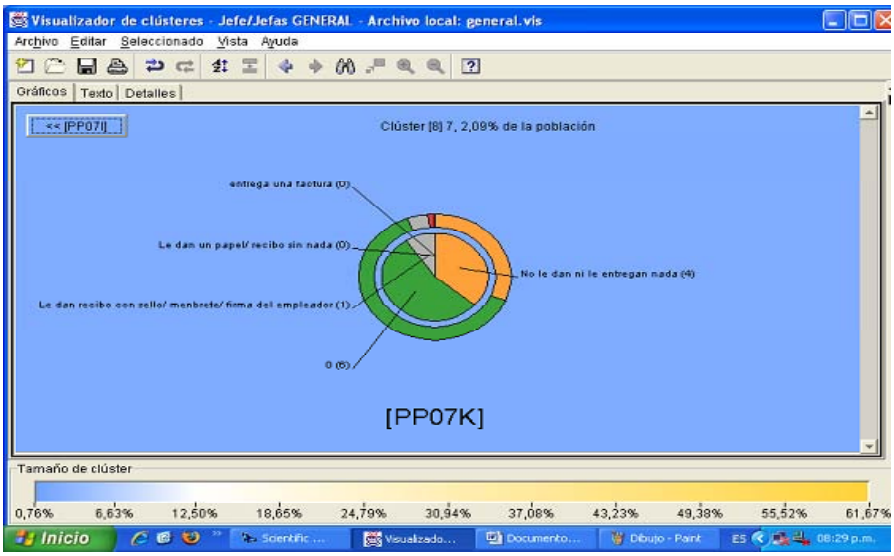

Figura 6.67: El siguiente resultado denota una vez más los índices de trabajo en negro.

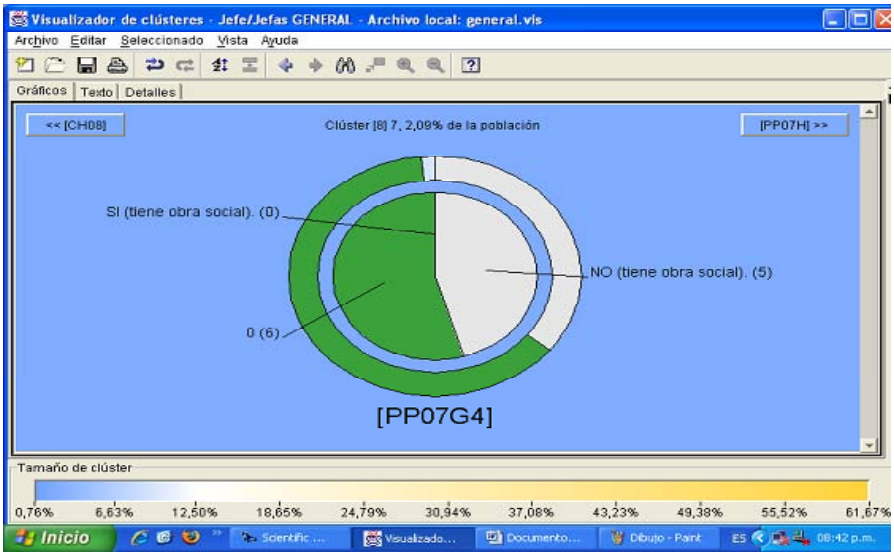

Figura 6.68: También es el cuarto clúster la opción “no tiene obra social” es la que posee mayor frecuencia.

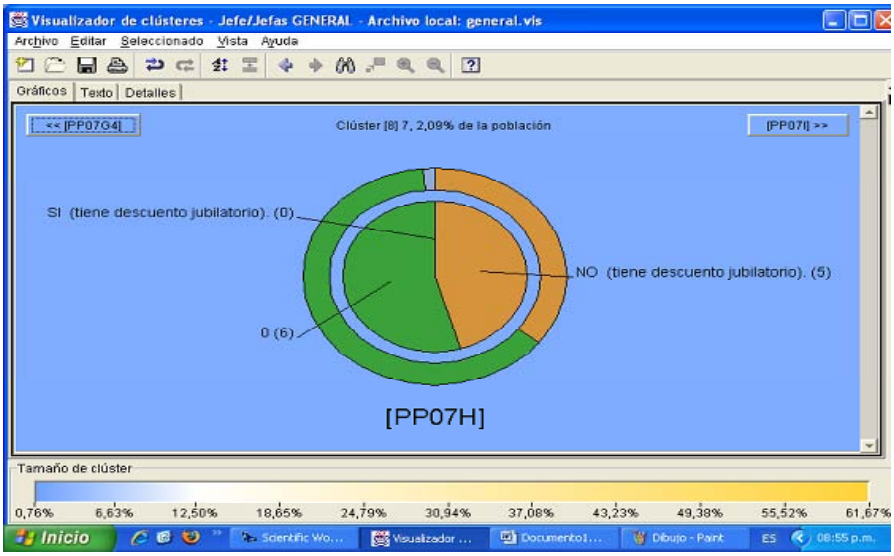

Figura 6.69: Podemos observar que no posee descuento jubilatorio.

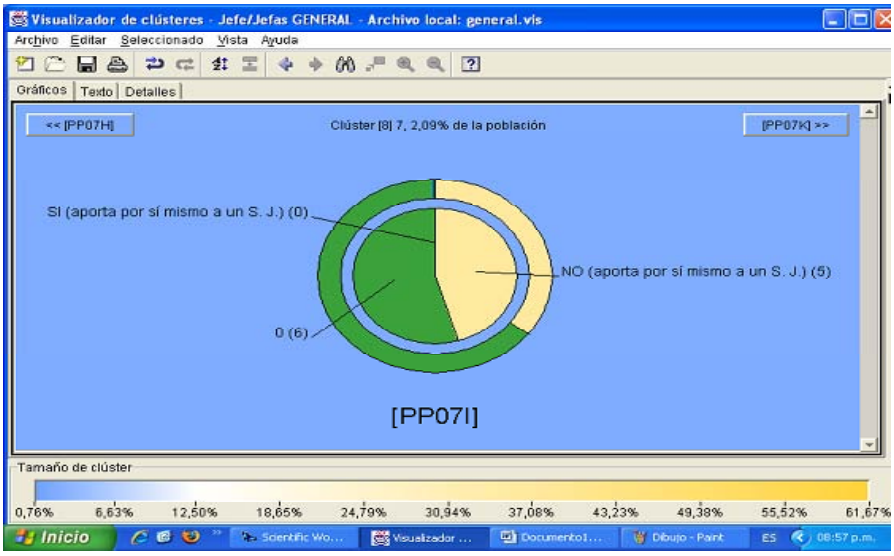

Figura 6.70: Visualizacíon de la variable PP07I (aporta por sí mismo a un S.J.).

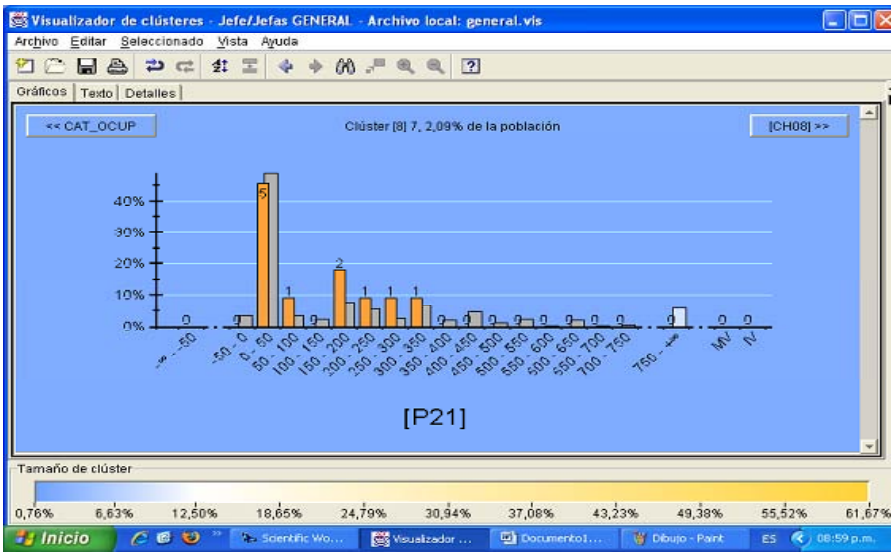

Figura 6.71: El ingreso de la ocupación principal de esta agrupación esta entre los 100 a 200 pesos.

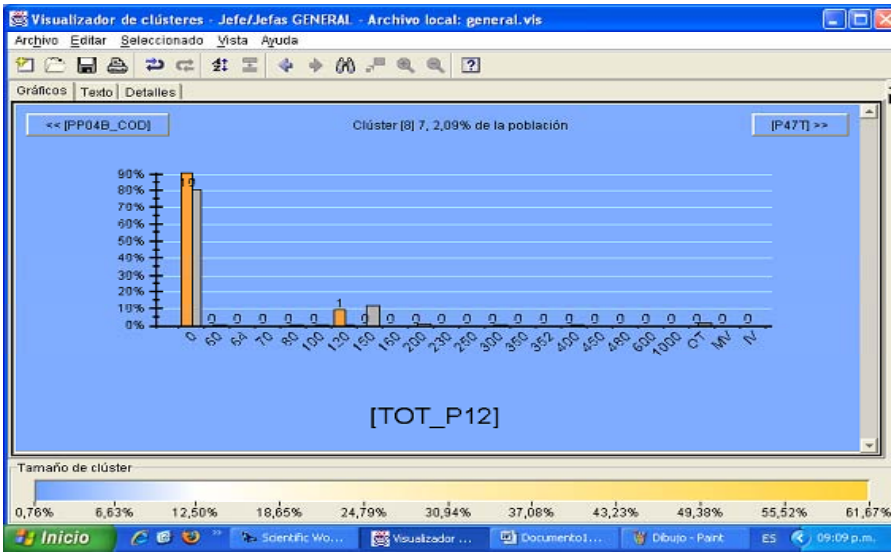

Figura 6.72: El ingreso de otras ocupaciones no supera los 120 pesos en esta agrupación.

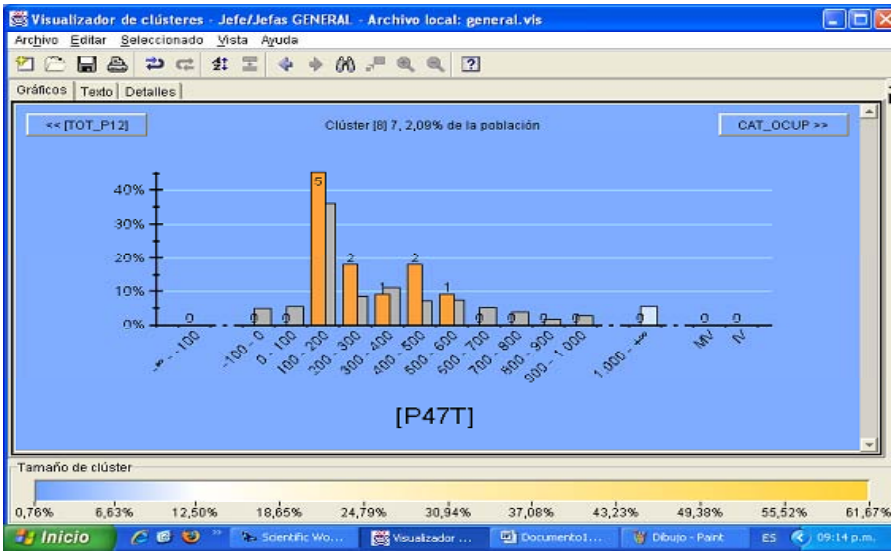

Figura 6.73: Visualización de la variable p47t (monto del ingreso total individual).

En el quinto cluster se puede observar que el perfil predominante es del sexo varón de unos 20 años de edad, con un estado civil de soltero, con un elevado grado de analfabetismo y ha nacido en Corrientes (ver fig. 6.74 de la pág. 185), (ver fig. 6.75 de la pág. 185), respectivamente (ver fig. 6.76 de la pág. 186).

También se observa posee una actividad económica que resulta ser predominantemente *"servicios de esparcimiento y servicios culturales y deportivos"* (ver fig. 6.77 de la pág. 186).

En el aspecto laboral se tiene que agregar que estas personas no cuentan con obra social, ni descuentos jubilatorios (ver fig. 6.78 de la pág. 187).

El trabajo en negro también está presente en esta agrupación como se puede ver en la fig. 6.79 de la pág. 187.

En cuanto a los ingresos, se puede observar que el monto de ingreso de la ocupación principal es de 0 pesos (ver fig. 6.80 de la pág. 188).

El monto de ingreso de otras ocupaciones es de 150 pesos y por ende el monto de ingreso total individual es un valor entre 100 y 200 pesos como se

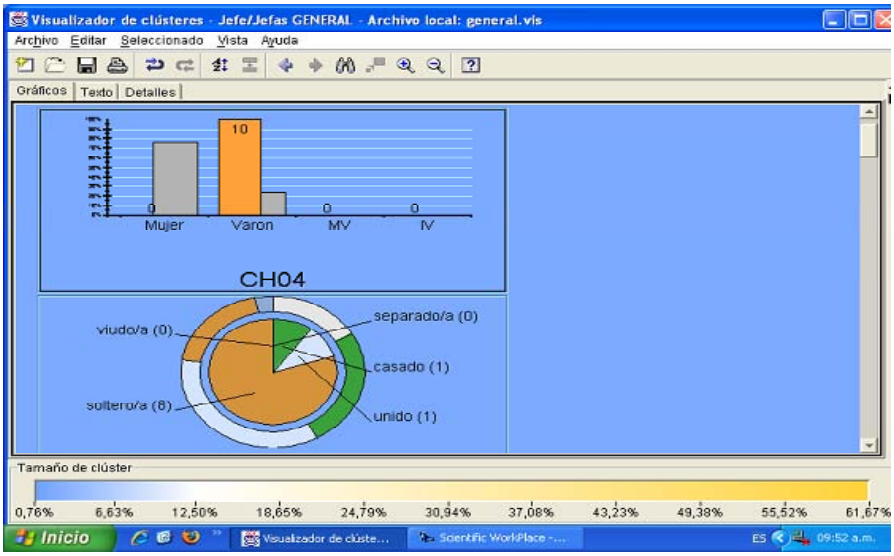

Figura 6.74: Las variables CH04 (sexo) y CH07 (estado civil).

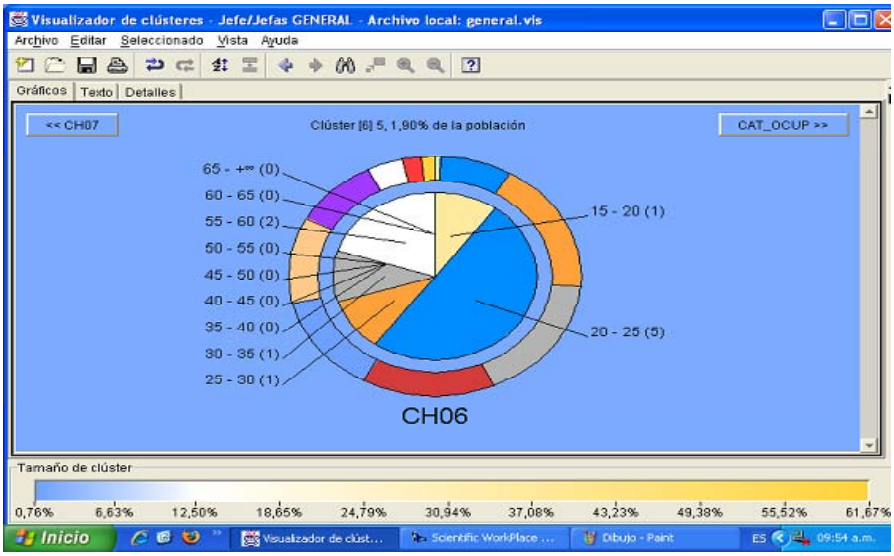

Figura 6.75: En la variable CH06 (años) el rango de edad con mayor representación es el de [20-25].

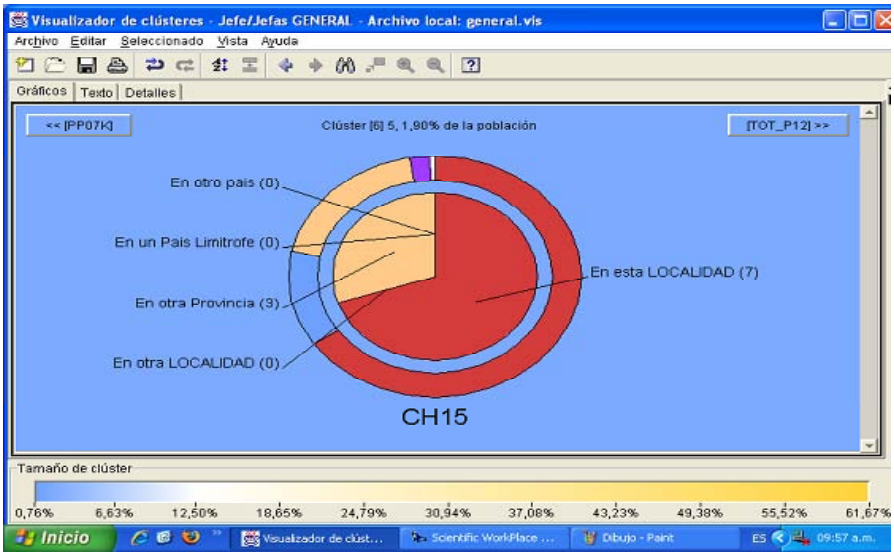

Figura 6.76: Estos individuos han nacido en su mayoría en esta localidad, es decir en Corrientes (Capital).

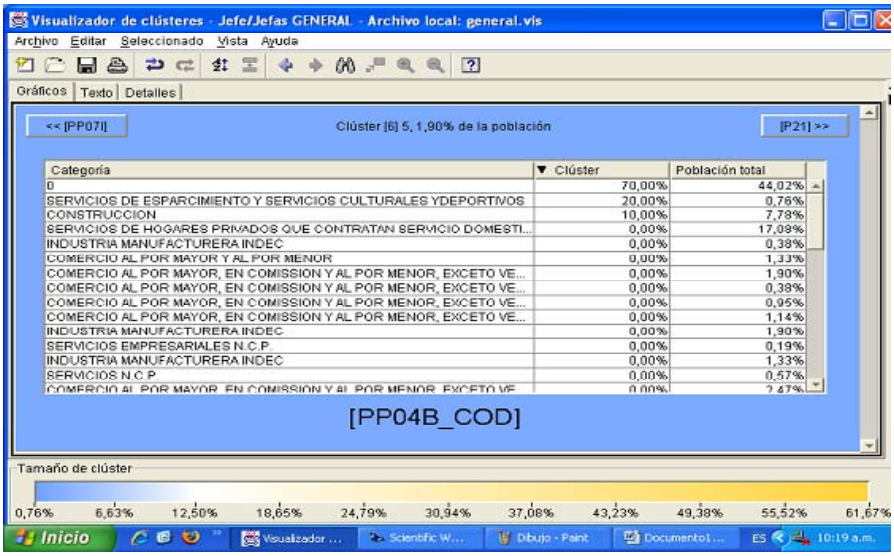

Figura 6.77: La construcción tambíen en esta agrupacion es la predominante.

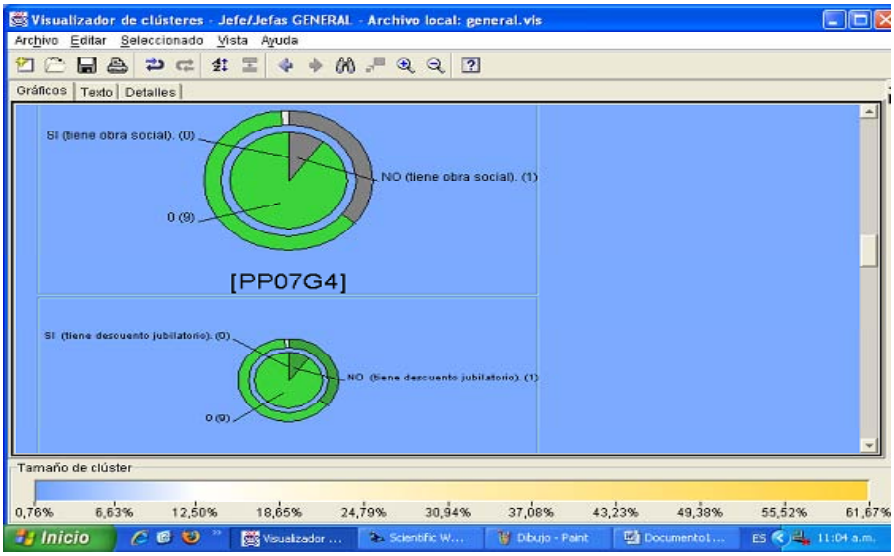

Figura 6.78: Visualización, de las variables PP07G4 (O. S.), PP07H (Desc. Jubilatorio).

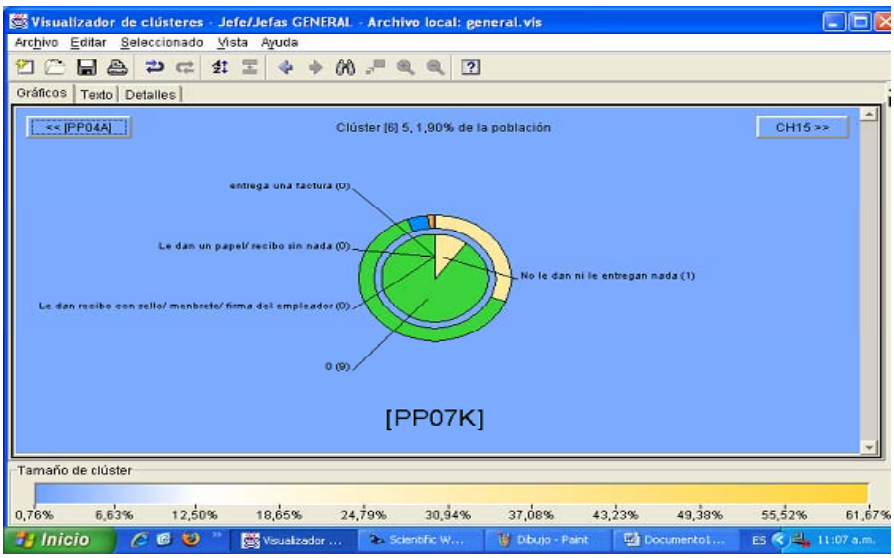

Figura 6.79: Visualización, de la variable PP07K (tipo de contrato laboral).

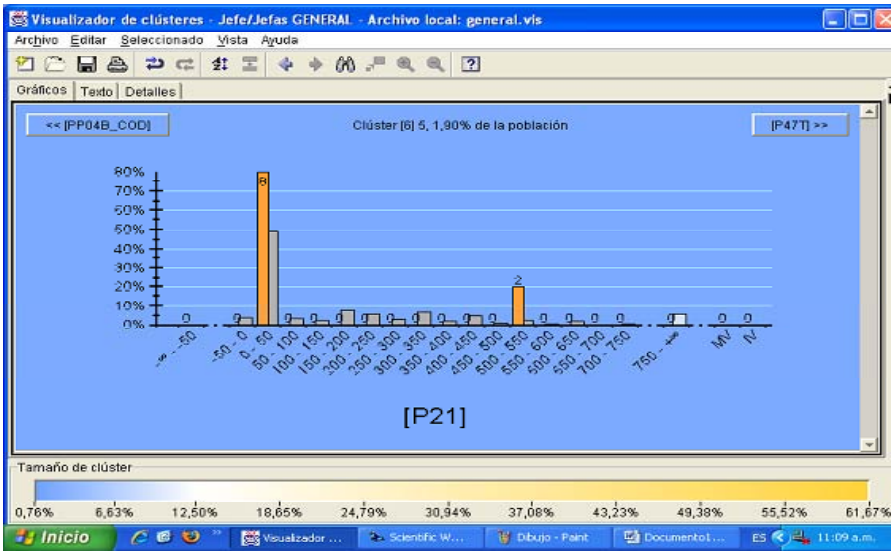

Figura 6.80: Visualización del monto del ingreso de la ocupación pincipal.

puede visualizar en las siguientes fig. 6.81 de la pág. 189 y la fig. 6.82 de la pág. 189.

En el sexto cluster (1,14% de la poblacíon total) se puede visualizar un predominio del sexo femenino con una edad de 49 años, un estado civil de separado / al y además la misma no ha nacido en otra provincia (ver fig. 6.83 de la pág. 190), (ver fig. 6.84 de la pág. 190), respectivamente (ver fig. 6.85 de la pág. 191).

La categoría ocupacional predominantemente es la de *"cuenta propia"* como se puede observar en la fig. 6.86 de la pág. 191, dedicándose al rubro de la construcción (ver fig. 6.87 de la pág. 192).

En el séptimo cluster (0,76 % de la población total) se podrá observar que el sexo predominante es el masculino, de estado civil separado y el mismo ha nacido en otra localidad de Corrientes (Capital), todo esto se puede visualizar en las siguientes graficas: (ver fig. 6.88 de la pág. 192), (ver fig. 6.89 de la pág. 193), respectivamente (ver fig. 6.90 de la pág. 193).

Con respecto al perfil de estas personas se puede decir que poseen un nivel de analfabetismo elevado como se puede comprobar en la (ver fig. 6.91 de la pág. 194).

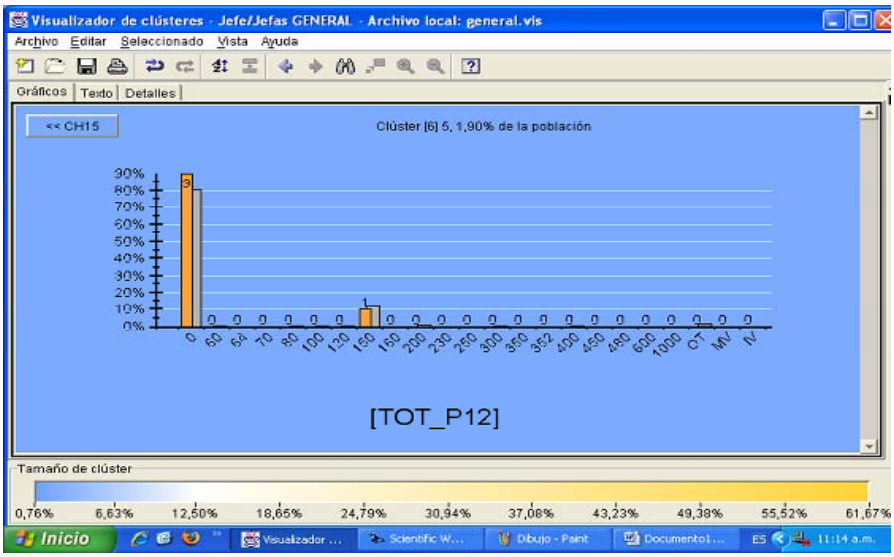

Figura 6.81: Visualización del monto del ingreso de otras ocupaciónes.

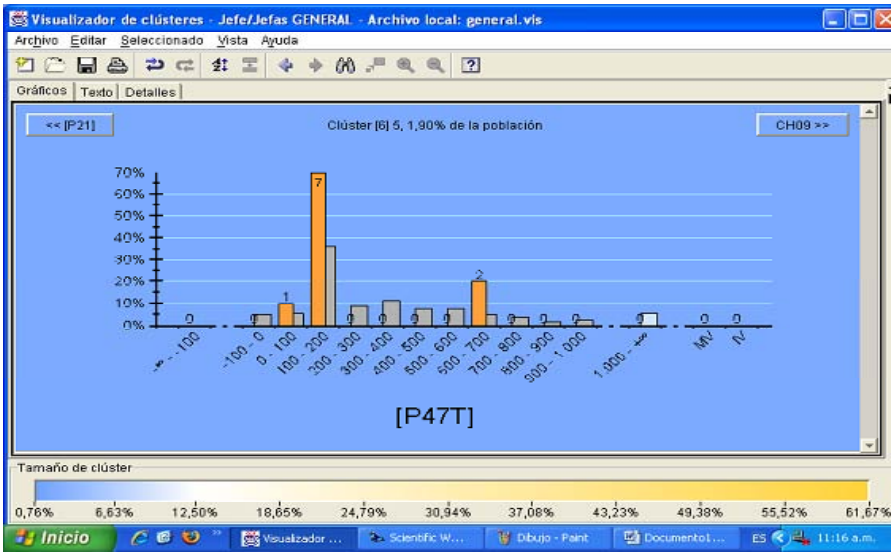

Figura 6.82: Visualización del monto del ingreso de total individual.

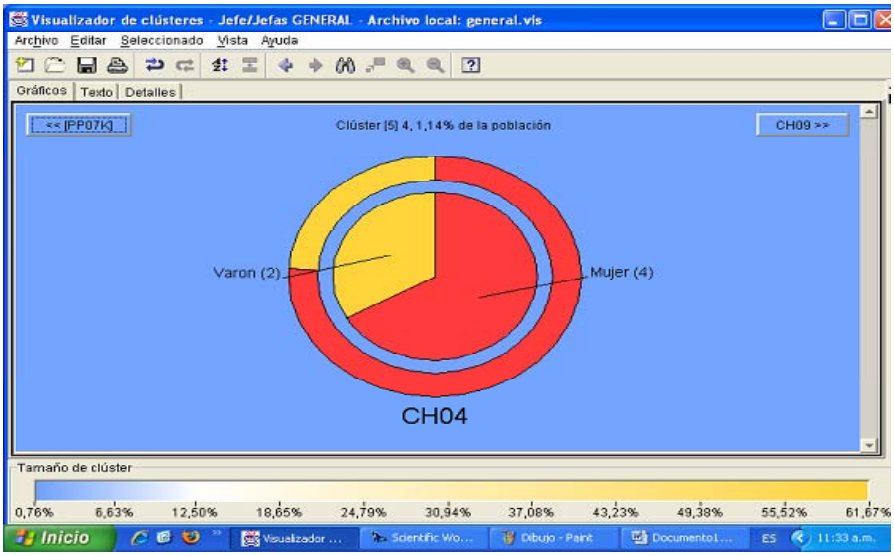

Figura 6.83: Visualización de la variable CH04 (sexo).

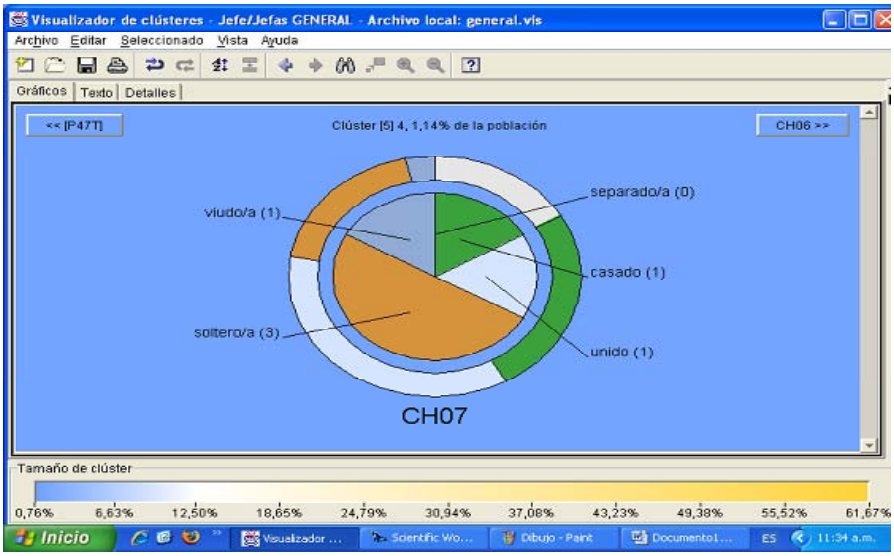

Figura 6.84: Visualización de la variable CH07(estado civil).

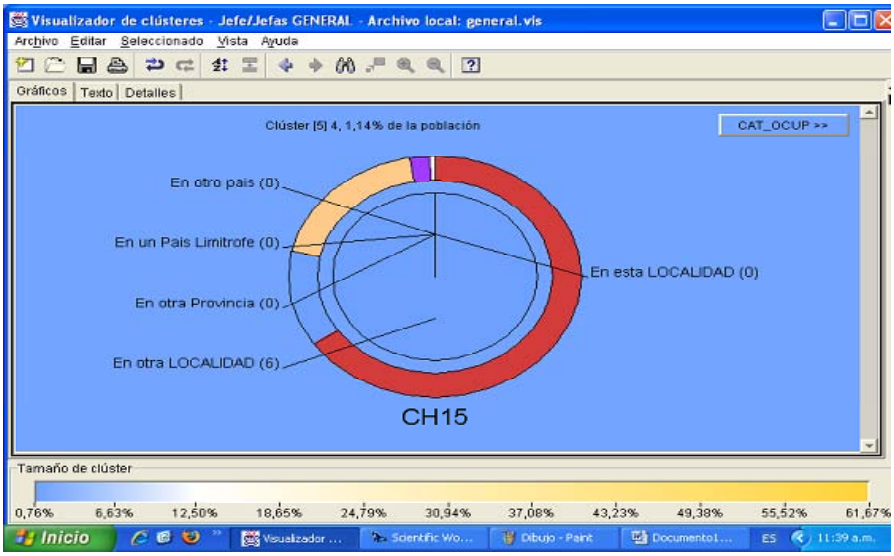

Figura 6.85: La opción "En otra localidad" es la predominante en la variable CH15 (¿Dónde nació?).

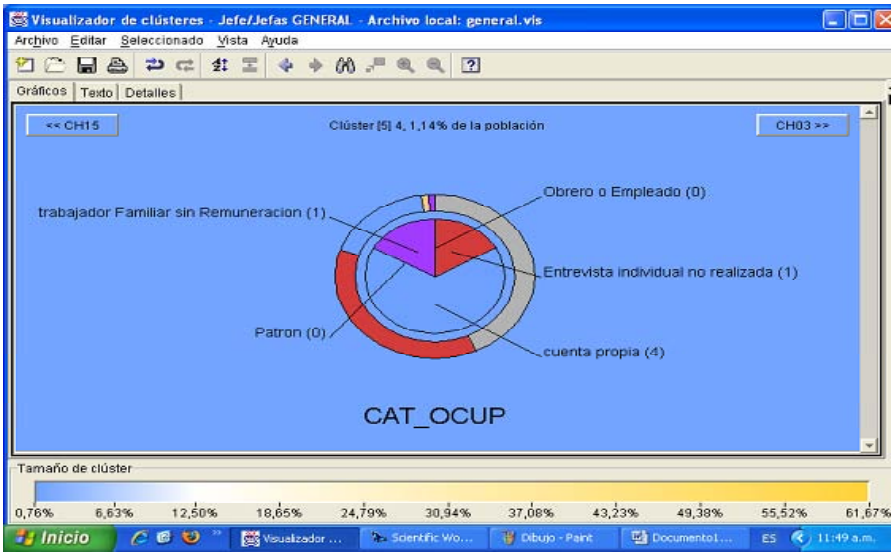

Figura 6.86: En la variable CAT_OCUP (Categoría Ocupacional) se puede observar a la opción con mayor representación "Cuenta propia".

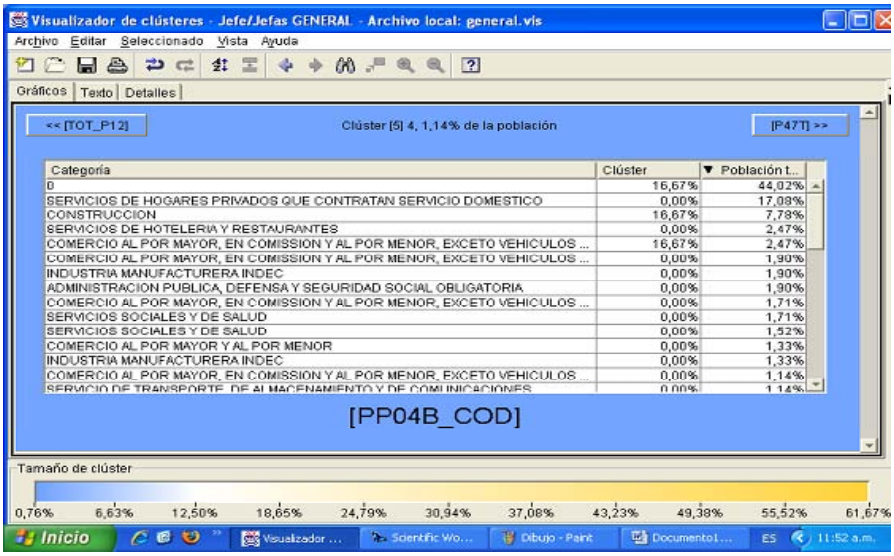

Figura 6.87: Visualización, de las variables PP04B_COD.

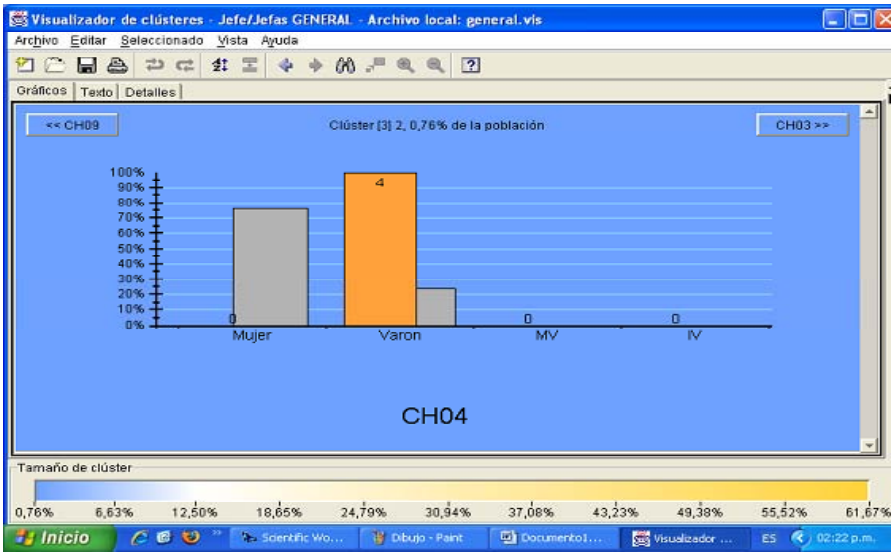

Figura 6.88: Visualización de la variable CH04 (sexo).

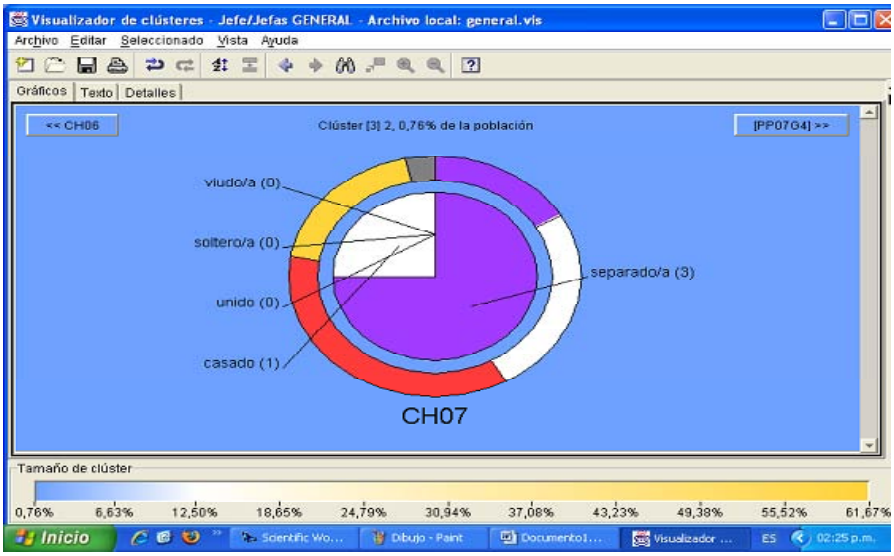

Figura 6.89: Visualización del diagrama circular de la variable CH07 (estado civil).

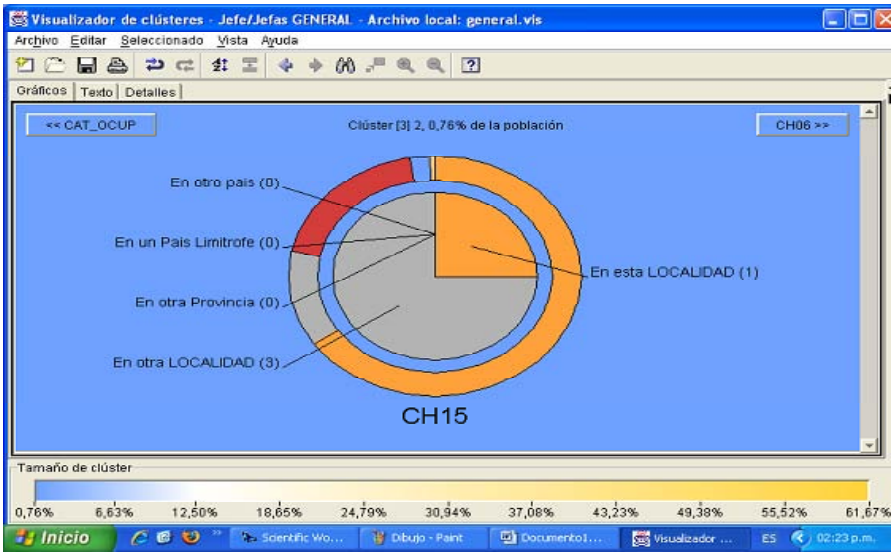

Figura 6.90: La opcíon "En otra localidad" es la que posee más frecuencia en la variable CH15(¿Dónde nacio?).

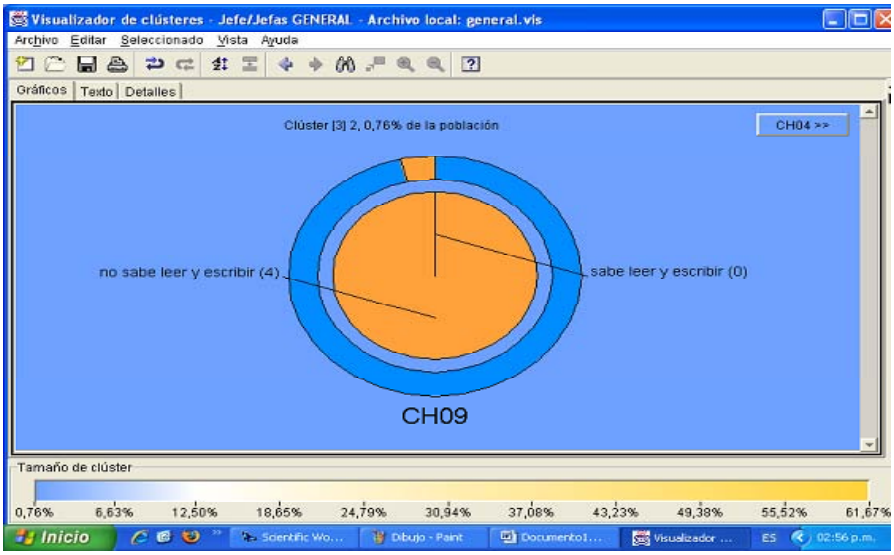

Figura 6.91: En la siguiente figura se puede comprobar el nivel de analfabetismo que poseen las personas de está agrupación.

La variable CAT_OCUP (categoría ocupacional) no existe un predominio de alguna categoría como se puede observar en la fig. 6.92 de la pág. 195, dedicándose estos al rubro de la *"construcción"* y a el *"servicios empresariales"* de (ver fig. 6.93 de la pág. 195).

En esas tareas o labores estás personas no poseen cobertura médica como se puede contemplar en la siguiente fig. 6.97 de la pág. 198. También cabe destacar que no tienen obra social y mucho menos descuento jubilatorio (ver fig. 6.93 de la pág. 195), (ver fig. 6.95 de la pág. 196).

En el octavo y último cluster se puede contemplar que también posee un 0,76 % de la población total como lo demuestra la siguiente fig.6.96 de la pág. 197.

Indagar los Perfiles Educativos de los Planes Jefes y Jefas

En este punto se estudiarán las principales variables relacionadas a la educación de las personas que poseen planes jefes/jefas.

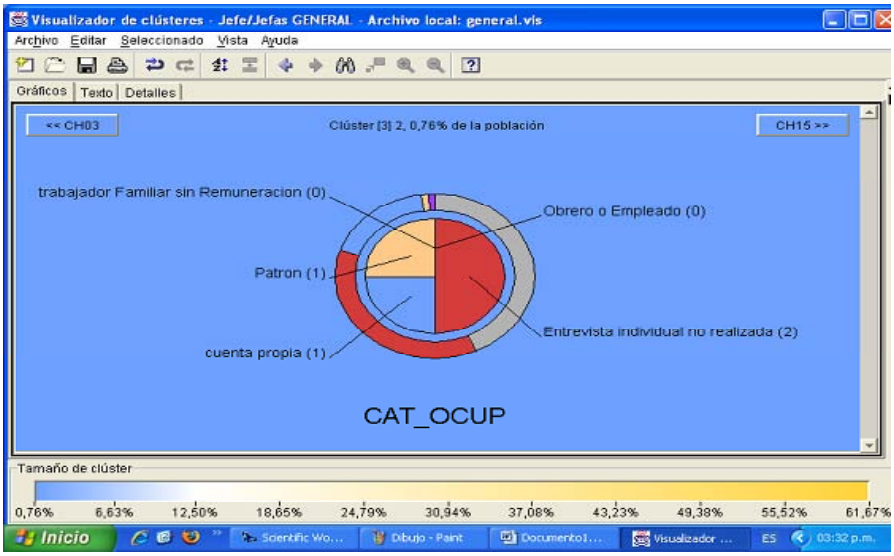

Figura 6.92: La visualización de la variable CAT_OCUP (categoría ocupacional) nos permite conocer las diferentes categorias que son predominantes.

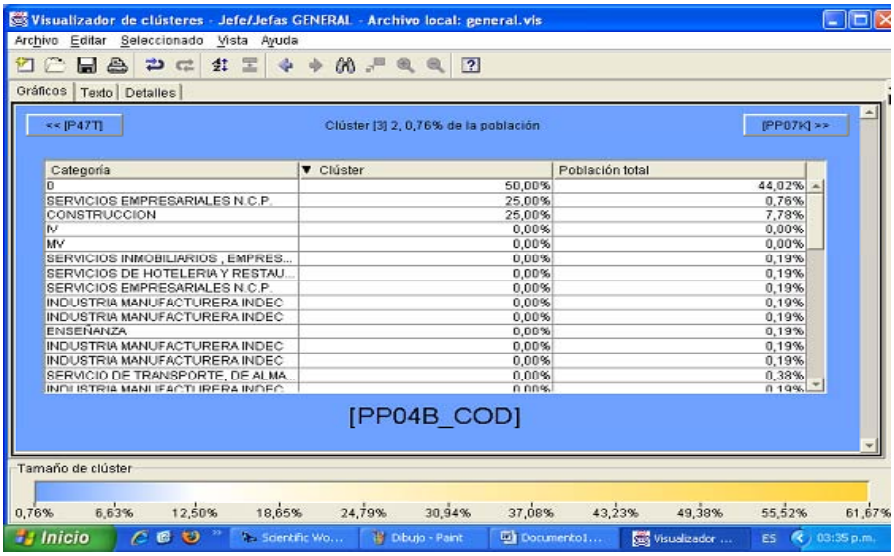

Figura 6.93: En el siguiente cuadro se puede contemplar a las opciones que contienen mayor frecuencia en la variable PP04B_COD.

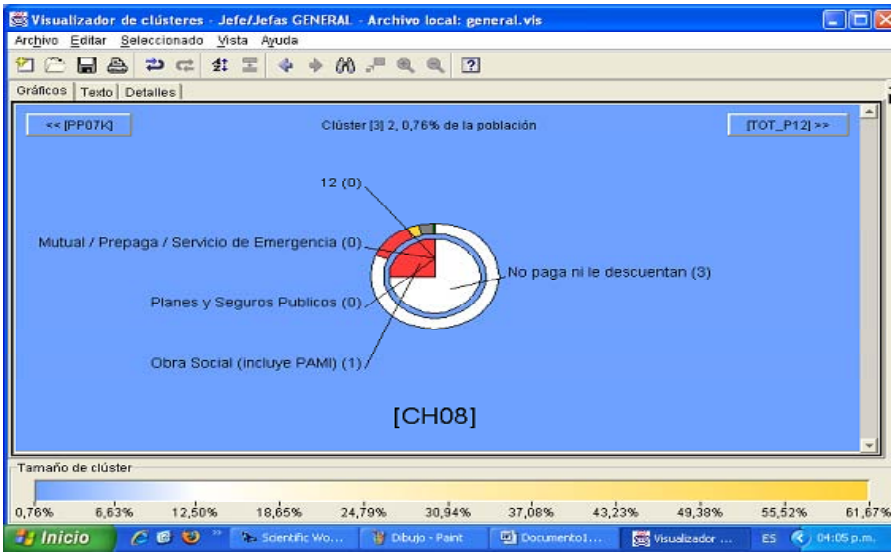

Figura 6.94: Muestreo del resultado de la variable CH08 (cobertura médica).

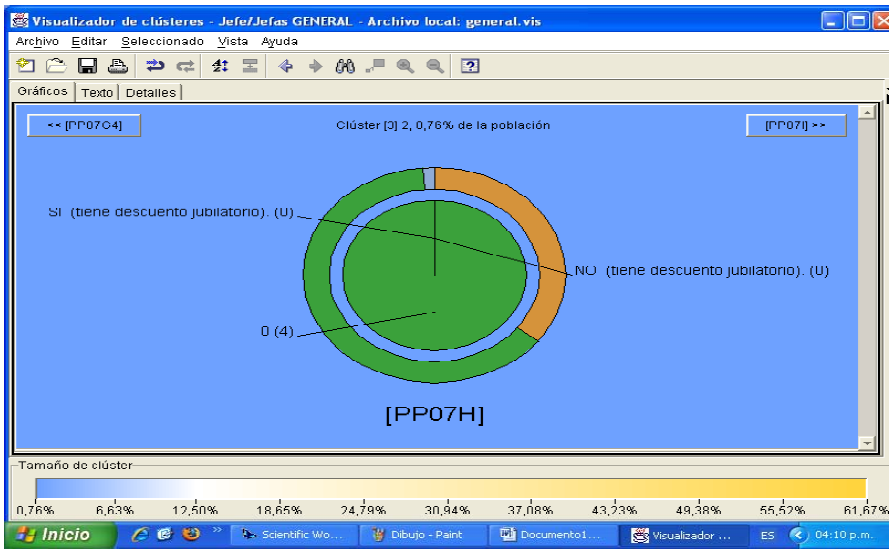

Figura 6.95: Visualización del resultado de la variable PP07H (Descuento Jubilatorio).

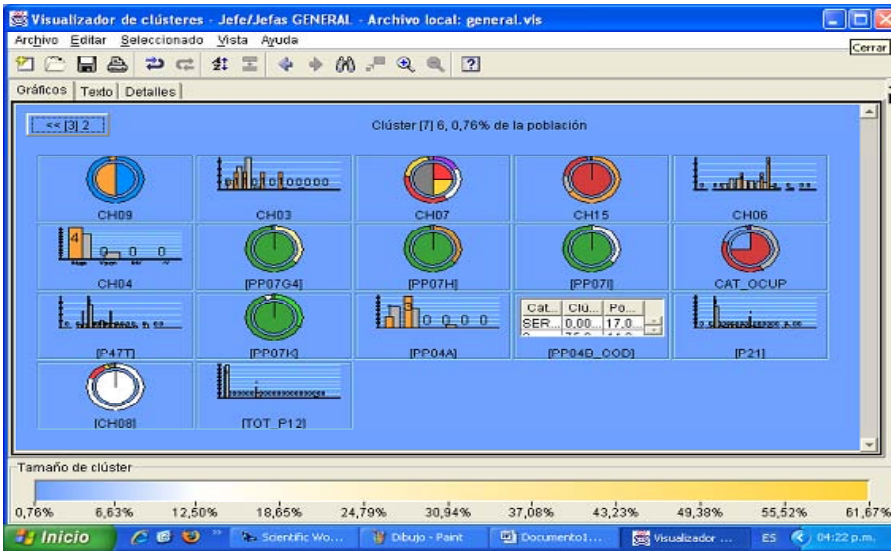

Figura 6.96: Vista general de la octava agrupación con un 0,76 % de la población total.

Básicamente, los pasos a llevar a cabo son similares al de *Conocer los Perfiles Socio Demográficos de los Planes Jefes y Jefas*, con la diferencia que en este se toman variables de educación en cambio de variables socio-demográficas (ver fig. 6.97 de la pág. 198).

Al ejecutar la función de minería, como se puede ver en la fig. 6.98 de la pág. 198, el *Intelligent Miner* proveerá en este caso un criterio de *condorcet* de **0,629** siendo que el aceptable es 0,65.

Al visualizar los objetos de resultados (ver fig. 6.99 de la pág. 199) se nota la exitencia de 8 clústers identificados por la ejecución de minería.

La primera columna contiene el nombre y el ID del cluster, la siguiente representa el tamaño de cluster en porcentaje con respecto a la muestra.

En este caso prácticamente un 97,92% de la población está representada sólo por los primeros cuatro clústeres, dividiéndose el 2,08% restante entre los demás.

La primer agrupación de 73,06 % de la población total, en ella se puede visualizar que sexo *femenido* es el predominante con un rango de edad de *[20-*

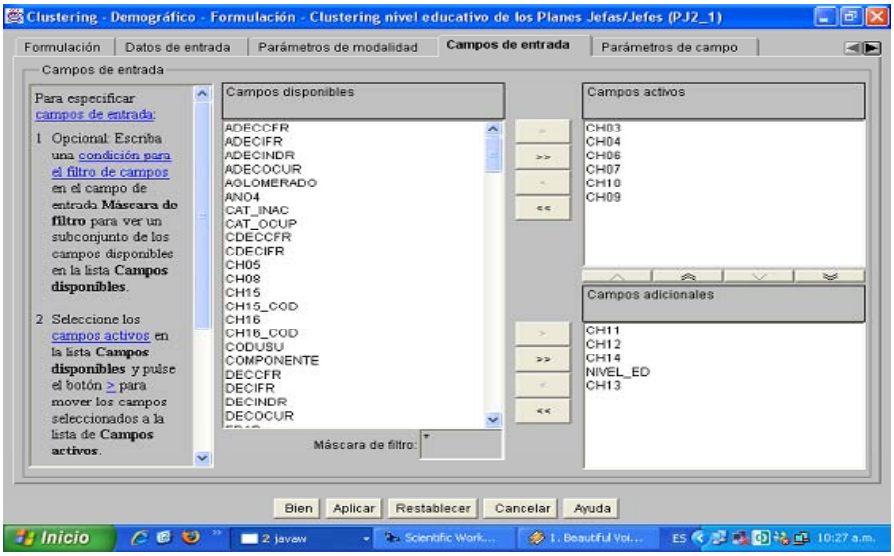

Figura 6.97: Selección de las variables de educación en los campos activos y campos adicionales.

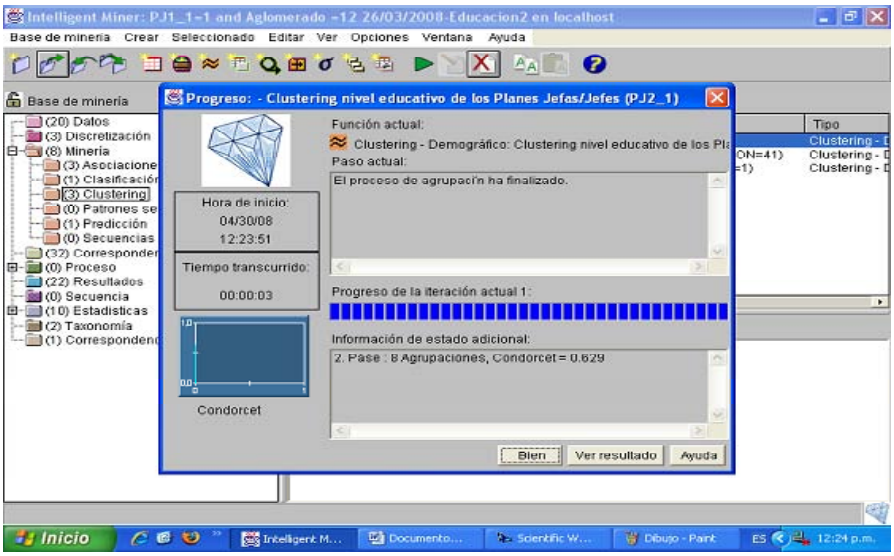

Figura 6.98: El cuadro de progreso del Intelligent Miner proveerá la siguiente información *(2 Pase: 8 Agrupaciones, Condorcet = 0,629)*.

Figura 6.99: Visualización de los diferentes clústers identificados por el *Intelligent Miner*.

25] años y de estado civil *unido o juntado* (ver fig. 6.100 de la pág. 200), (ver fig. 6.101 de la pág. 200), respectivamente (ver fig. 6.102 de la pág. 201).

Respecto a la dimensión educativa se puede observar la variable CH09 *(Analfabetismo)* que asume el valor "sabe leer y escribir" (ver fig. 6.103 de la pág. 201).

El nivel educativo predominante de estas personas es de primaria completa como se puede visualizar en la fig. 6.104 de la pág. 202.

En la variable CH10 (Asiste o Asistió a algún establecimiento educativo) se puede ver la opción sobresaliente de *"no asiste, pero asistió"* (ver fig. 6.105 de la pág. 202).

El nivel más alto que cursan o cursaron estas personas puede observarse en la fig. 6.106 de la pág. 203 que es el *"nivel primario"*. Como se puede observar en la fig. 6.107 de la pág. 203 se puede contemplar a un elevado número de personas que finalizaron dicho nivel.

Y para finalizar con el análisis de este cluster se visualizará la variable CH14 (¿Cuál fue el último año que aprobó?), donde se puede observar que el

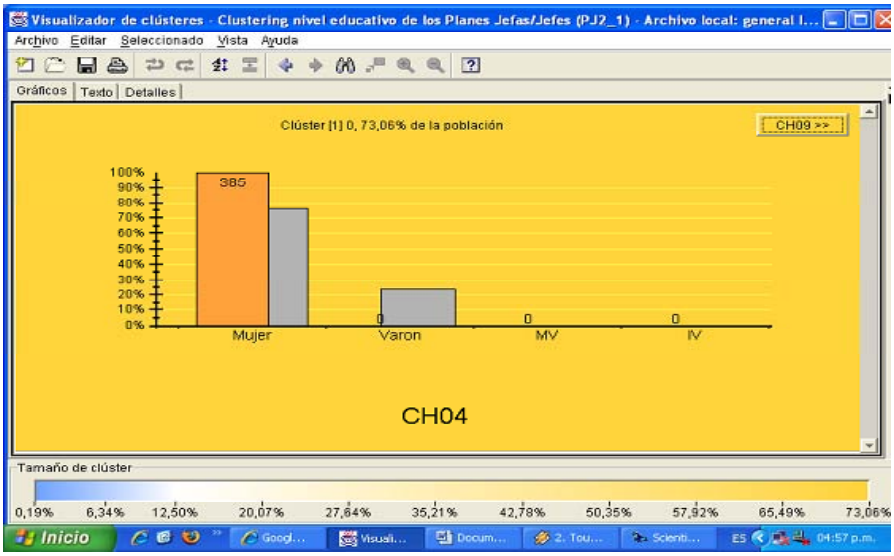

Figura 6.100: Visualización de la variable CH04 (sexo).

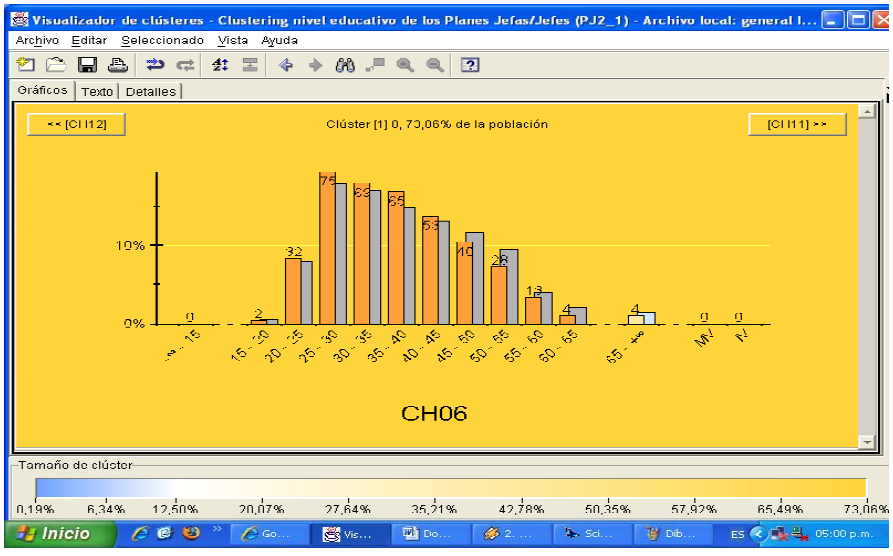

Figura 6.101: Muestreo del contenido de la variable CH06 (años).

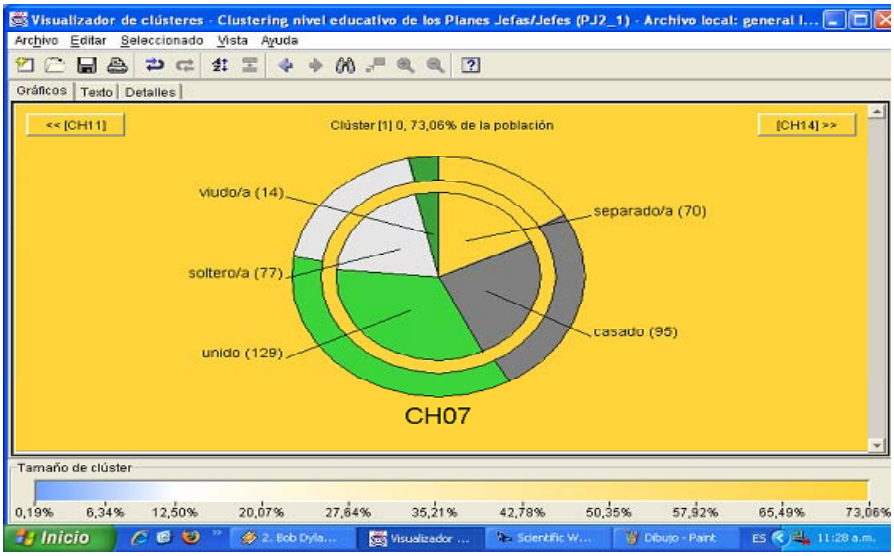

Figura 6.102: La opción "unido" es la predominante en la variable CH07 (Estado Civil).

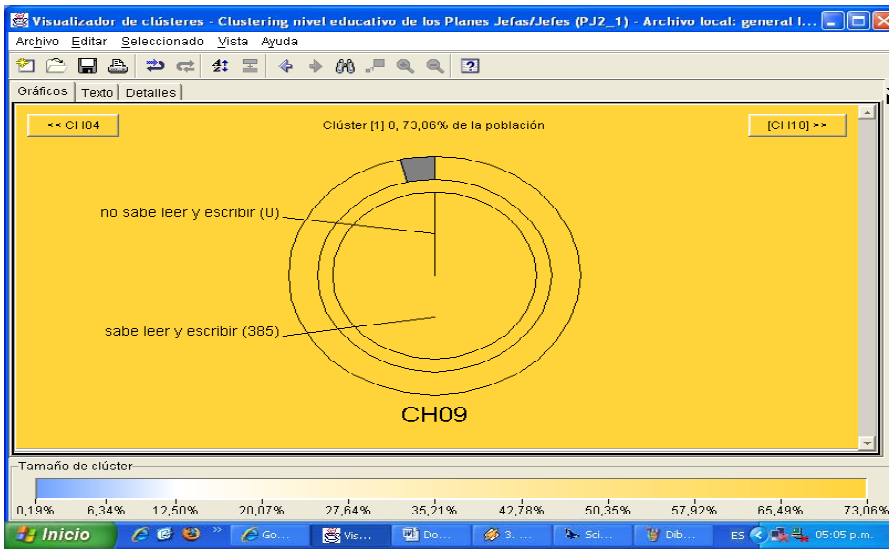

Figura 6.103: Visualización de la variable CH09 (analfabetismo).

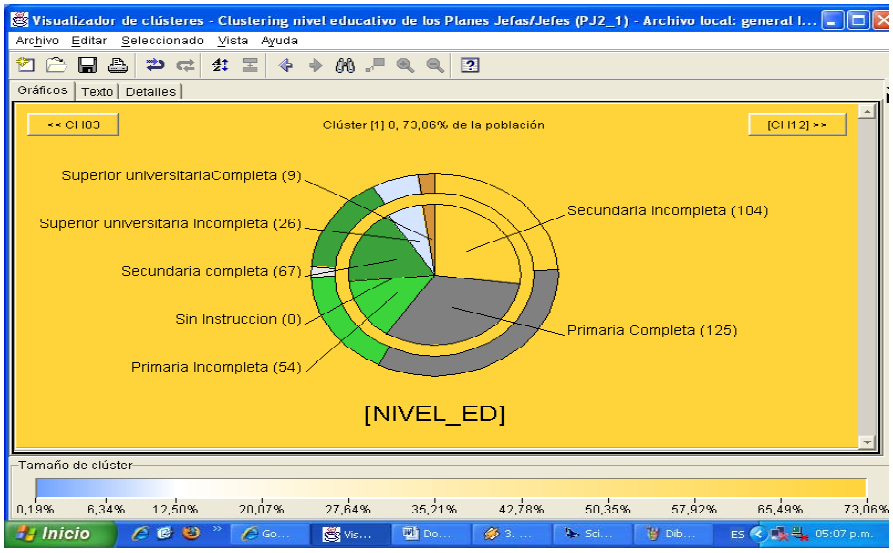

Figura 6.104: El nivel educativo predominante es *"Primaria Completa"*.

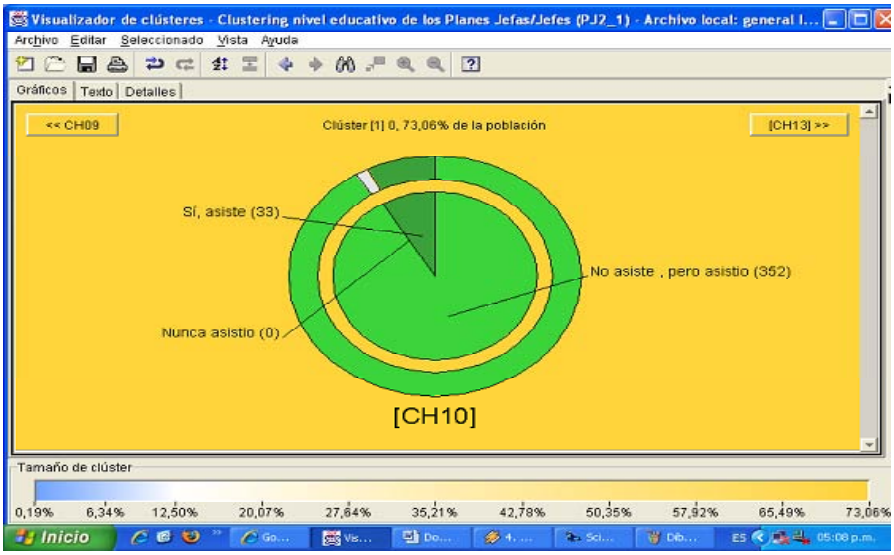

Figura 6.105: Visualización de la variable CH10 (asistencia a algun establecimiento educativo).

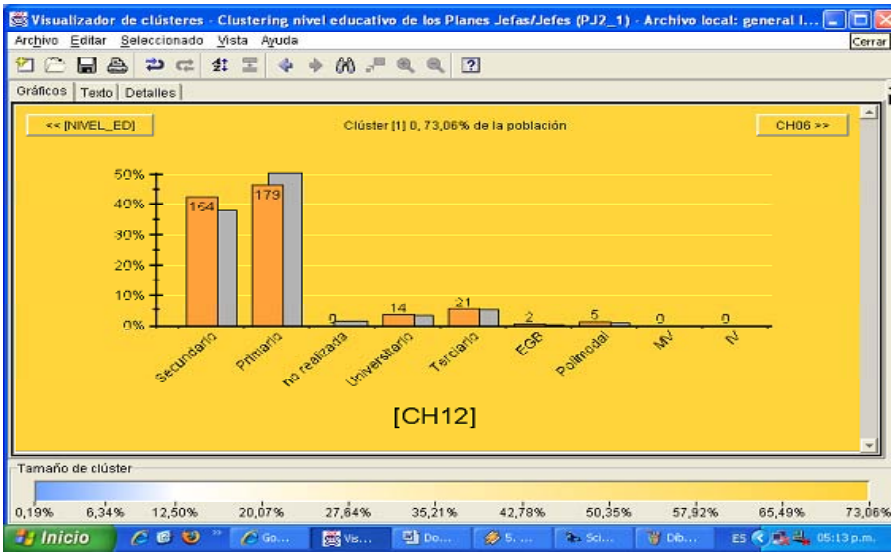

Figura 6.106: Visualización de la variable CH12 (¿Cuál es el nivel más alto que cursa o cursó?).

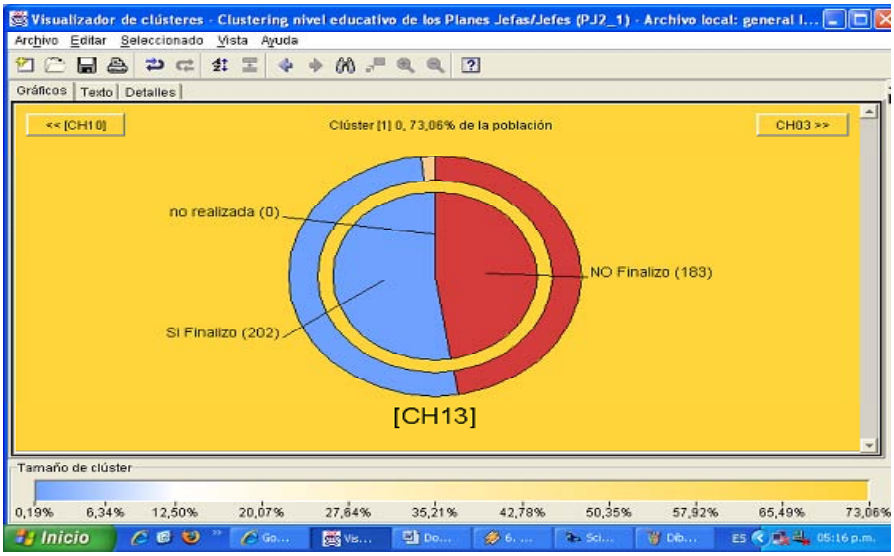

Figura 6.107: Visualización de la variable CH13 (si finalizo el nivel más alto alcanzado o cursado).

máximo año aprobado por estos individuos es el *segundo* año (ver fig. 6.108 de la pág. 204).

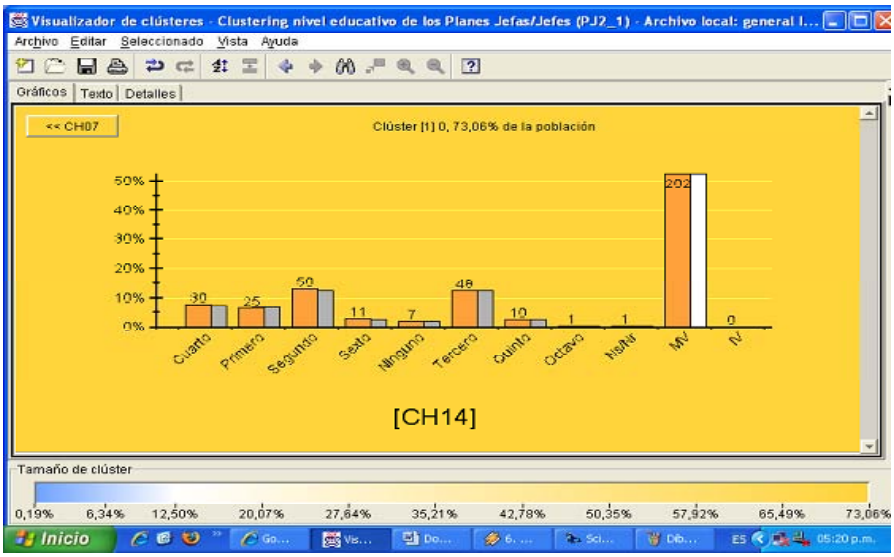

Figura 6.108: En la variable CH14 (¿Cual fue el último año que aprobó?), la opción predominante es *"Segundo año"*.

En el segundo cluster (20,68% de la población total) se puede visualizar que el sexo predominante es el *varón* con un estado civil de *unido o juntado* y con la edad que varia de los 30 a 35 años (ver fig. 6.109 de la pág. 205), (ver fig. 6.110 de la pág. 205).

En la fig. 6.111 de la pág. 206 se puede observar que el índice de analfabetismo tiene **3,67%** "No sabe, leer y escribir " y un **96,33%** "Sabe, leer y escribir".

El nivel educativo que resulta ser predominante es *"primaria completa"* como se puede ver en la fig. 6.112 de la pág. 206.

El nivel más alto que cursaron estas personas es el *"nivel primario"* como se puede observar en la fig. 6.113 de la pág. 207. También se puede observar en la fig. 6.114 de la pág. 207 la existencia de un elevado número de personas que no han finalizado dicho nivel. Teniendo a la opción "segundo año" como la predominante de la variable CH14 (¿Cuál fue el último año que aprobó?) (ver fig. 6.115 de la pág. 208).

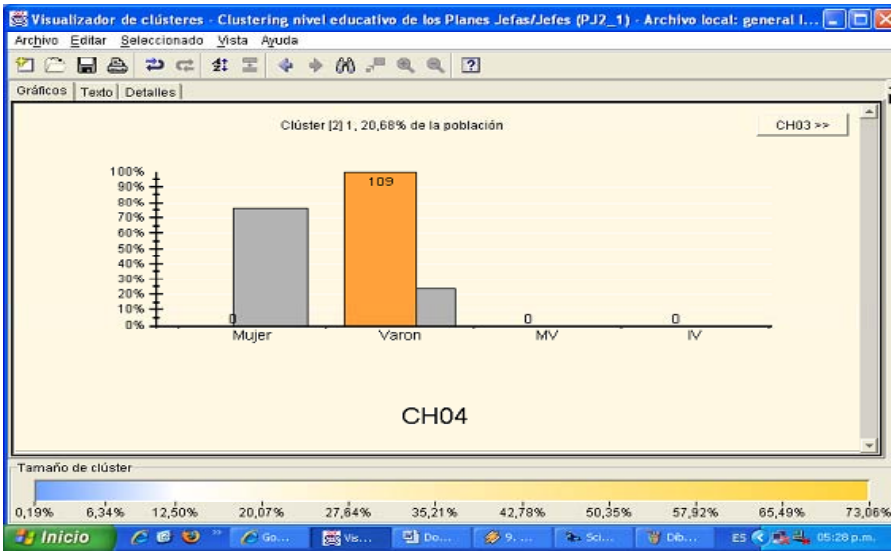

Figura 6.109: Visualización del segundo cluster (20,68 % de la población) donde el sexo predominante es el varón.

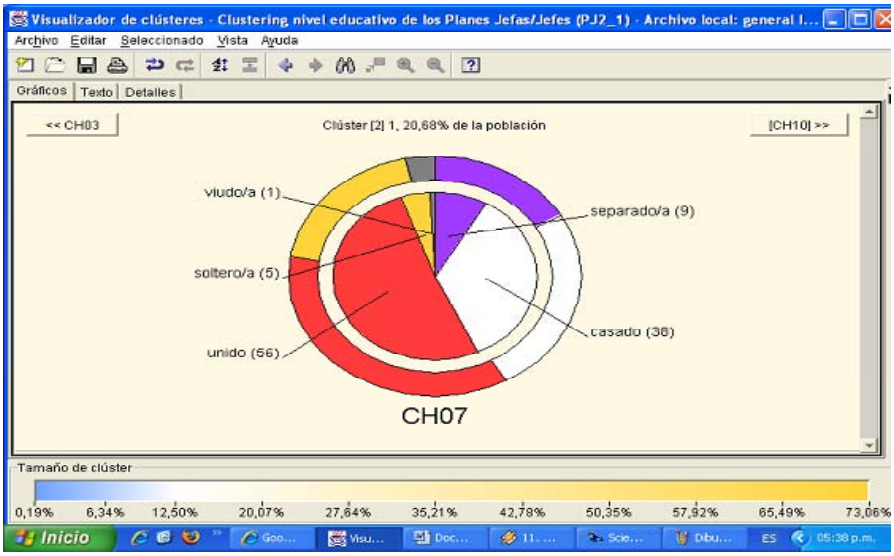

Figura 6.110: Visualización del segundo cluster (20,68 % de la población) con un estado civil de unido o juntado.

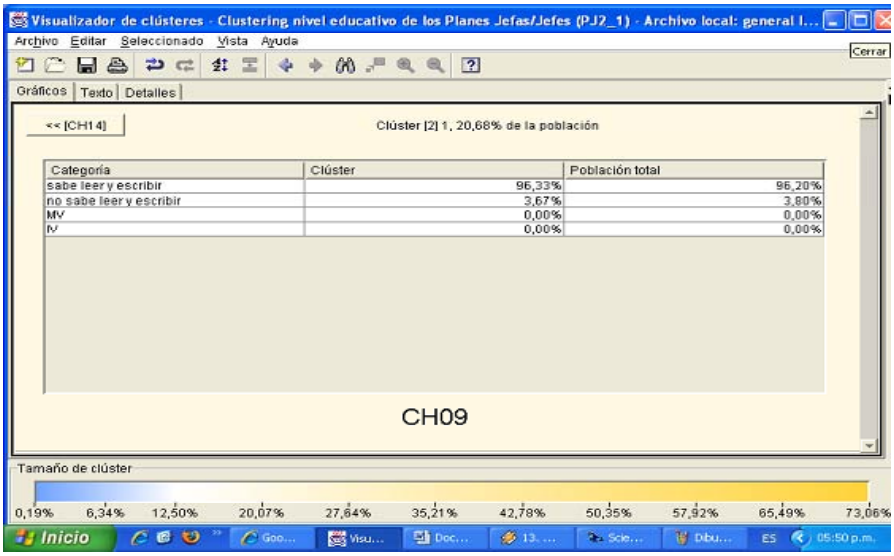

Figura 6.111: Muestreo del los ìndices de analfabetismo obtenidos de la variable CH09 (Analbafetismo).

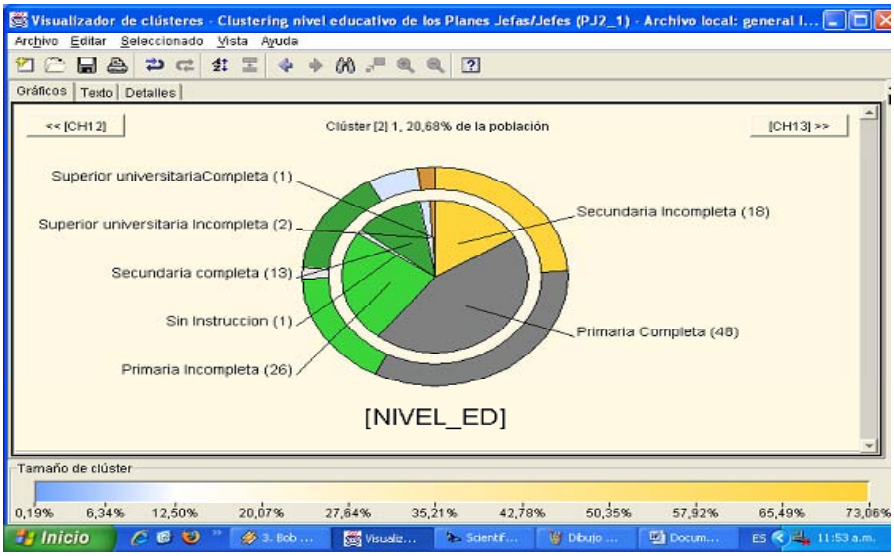

Figura 6.112: El nivel educativo predominante es de “Primaria completa”.

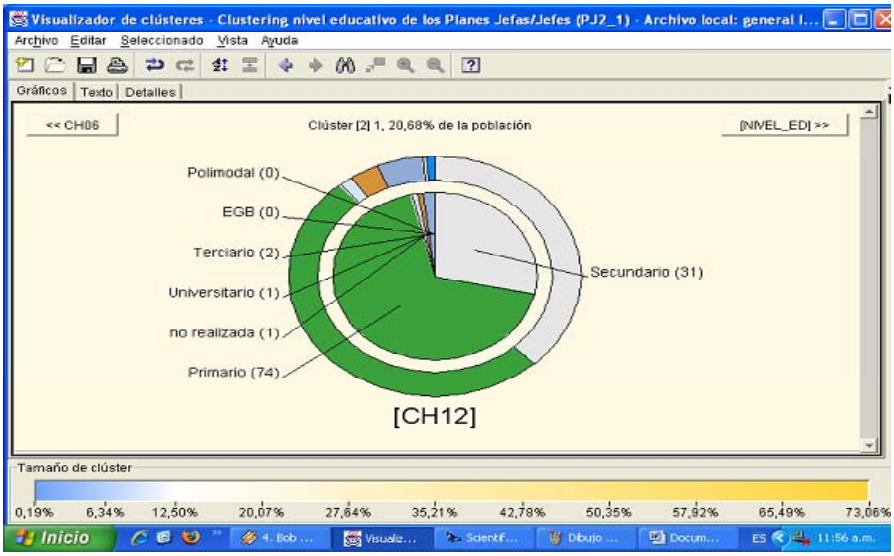

Figura 6.113: El nivel más alto que cursaron estas personas fue el "Nivel Primario".

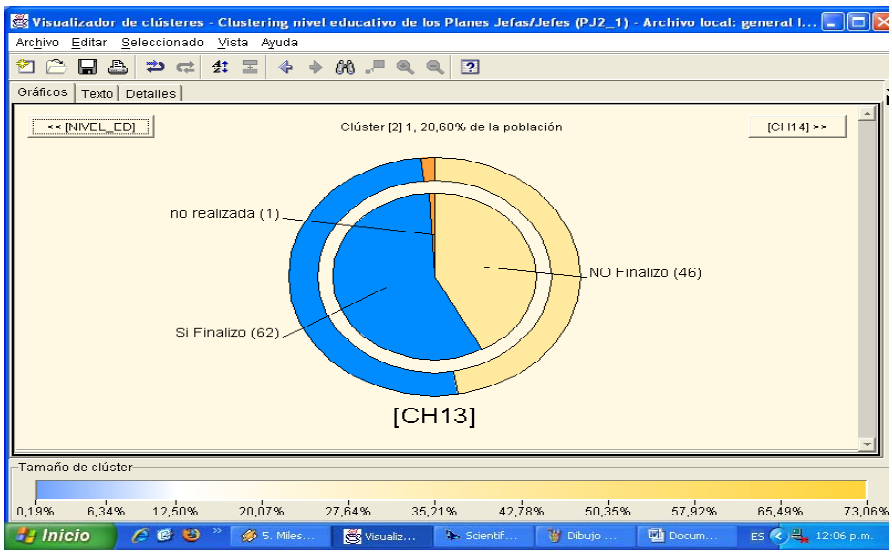

Figura 6.114: Visualización de la variable CH13 (¿Finalizó ese nivel?).

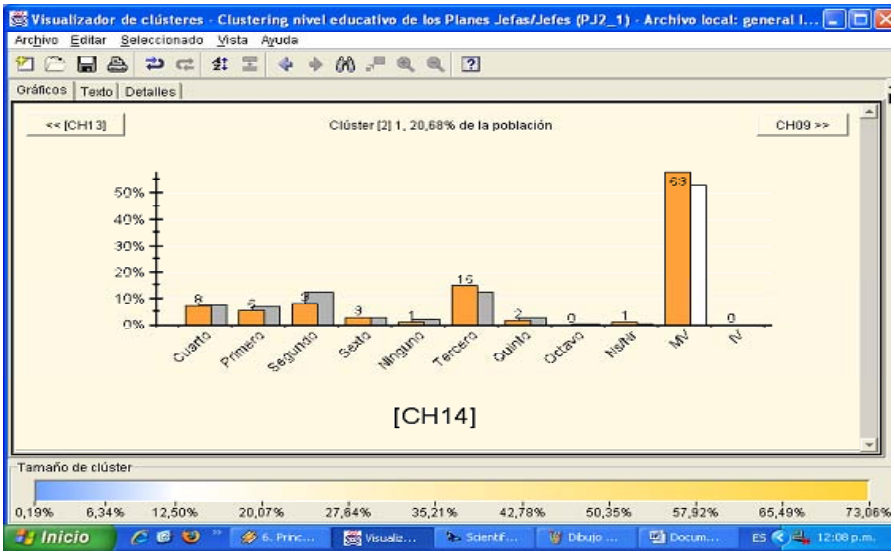

Figura 6.115: El "segundo año" es la opción con más representación en la variable CH14 (¿Cuál fue el último año que aprobó?).

En el tercer cluster de 2,47 % de la población total se puede encontrar que la variable CH04 (sexo) posee a la sexo masculino como el predominante (ver fig. 6.116 de la pág. 209).

También se puede observar en la fig. 6.116 de la pág. 209 el resultado de la variable CH07 (estado civil).

En la fig. 6.117 de la pág. 209 se puede comprobar que la opcion "No asiste, pero asistió" es la predominante en la variable CH10 (¿Asiste o asistió a algún establecimiento educativo?).

Se puede apreciar en la fig. 6.118 de la pág. 210 que no todos los establecimientos educativos a los que estas personas recurren son "Públicos", si no que también se puede apreciar la existencia de los "Privados".

En la fig. 6.119 de la pág. 210 se puede observar que la variable asume las opciones de nivel secundario como las del nivel universitario.

En la fig. 6.120 de la pág. 211 se puede apreciar la cantidad de personas que abandonaron estos niveles educativos.

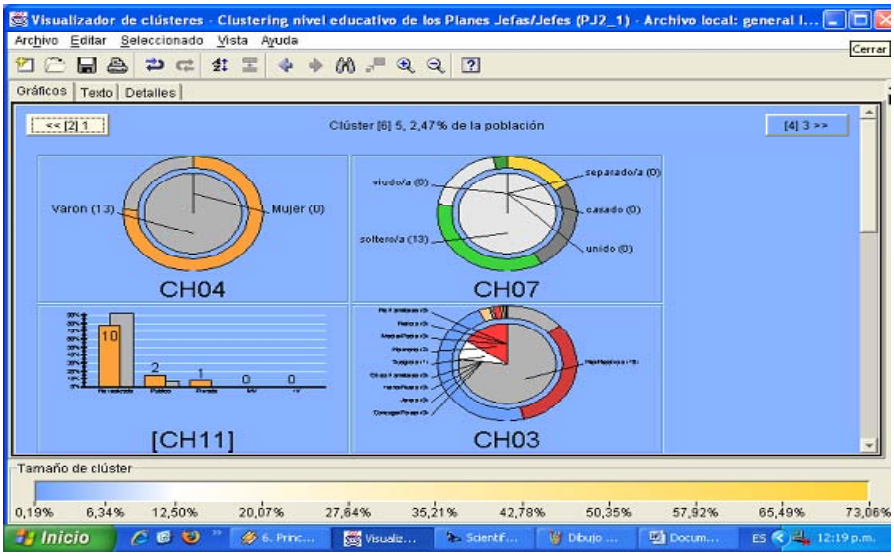

Figura 6.116: Visualización de las variables CH04 (sexo) y CH07 (estado civil).

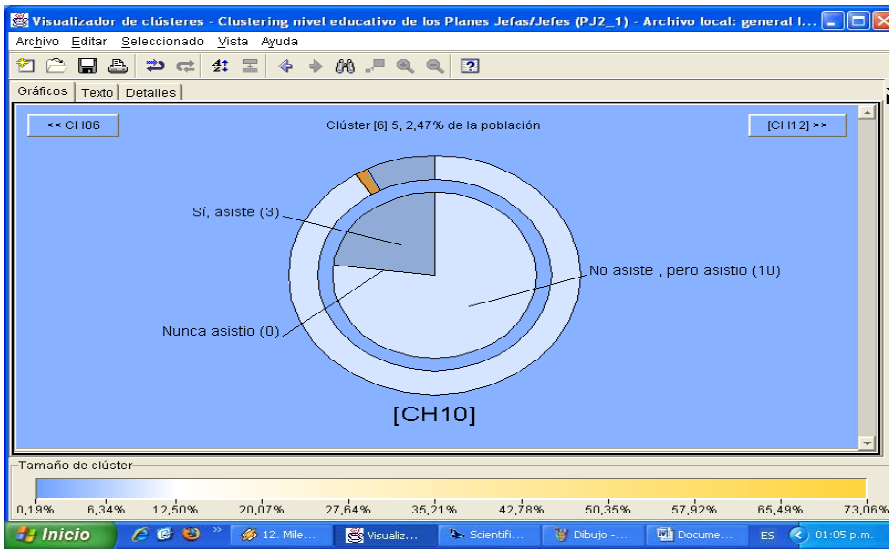

Figura 6.117: Visualización de la variables CH10 (¿Asiste o asistió a algún establecimiento educativo?).

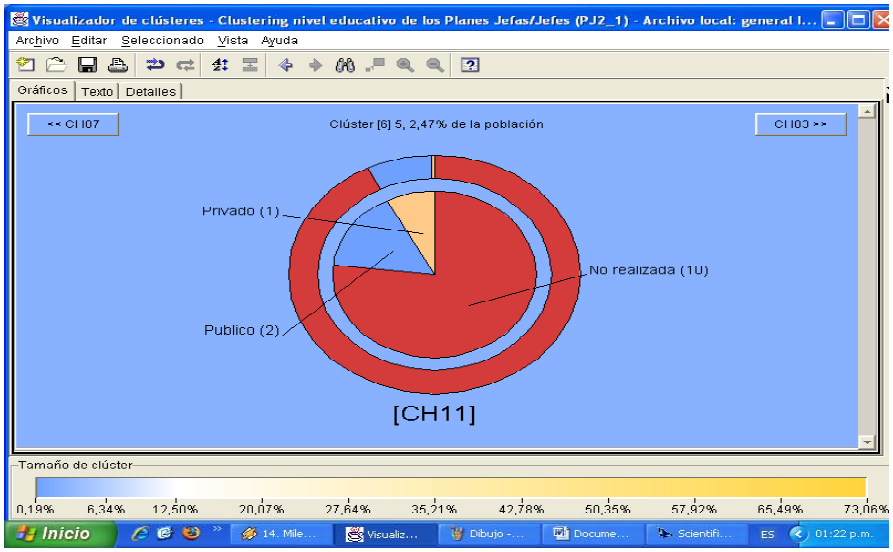

Figura 6.118: Visualización de la variable CH11 (Tipo de establecimiento educativo).

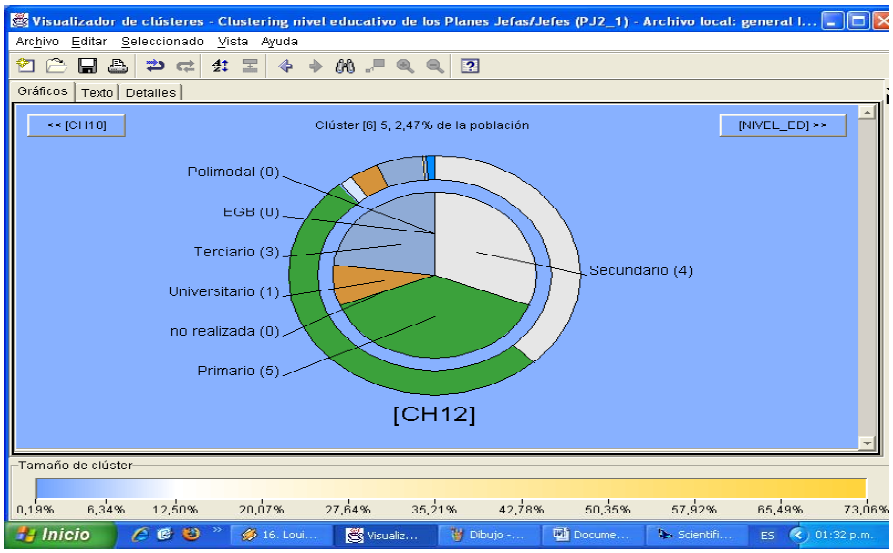

Figura 6.119: Muestreo de la variable CH12 (¿Cuál es el nivel más alto que cursa o cursó?).

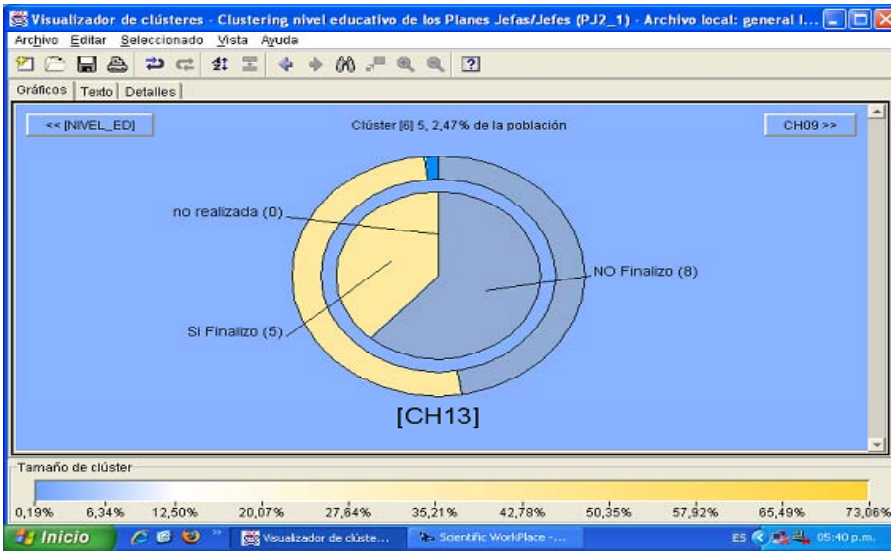

Figura 6.120: Visualización de la variable CH13 (*¿Finalizó ese nivel?*).

Siendo el nivel educativo alcanzado por estas personas los de "Superior Universitario Incompleta" y "Secundaria Incompleta", como se puede vizualizar en el fig. 6.121 de la pág. 212.

En el cuarto cluster 2,47 % de la población total se puede encontrar que el sexo predominante es el *femenino* con una distribución de la variable edad con las misma frecuencia para el rango [35-40] como para [50-55] y sucediendo lo mismo pero en este caso con la variable CH07 (estado civil) donde esta asume las opciones de "separado/a" y de "unido" (ver fig. 6.122 de la pág. 212), (ver fig. 6.123 de la pág. 213), respectivamente (ver fig. 6.124 de la pág. 213) .

Sobre su educación se puede observar en la fig. 6.125 de la pág. 214 que *"no sabe leer, ni escribir"* es la opción que predomina la variable CH09 (Analfabetismo).

En el quinto clúster de 0,76 % de la población total se puede encontrar que la variable CH04 (sexo) posee a la sexo femenino como el predominante (ver fig. 6.126 de la pág. 215).

También se puede observar en la siguiente fig. 6.127 de la pág. 215 que el estado civil que predomina en el quinto clúster es el "soltero/a", con un rango de edad de [20-25] años como se puede visulizar en la siguiente fig. 6.128 de

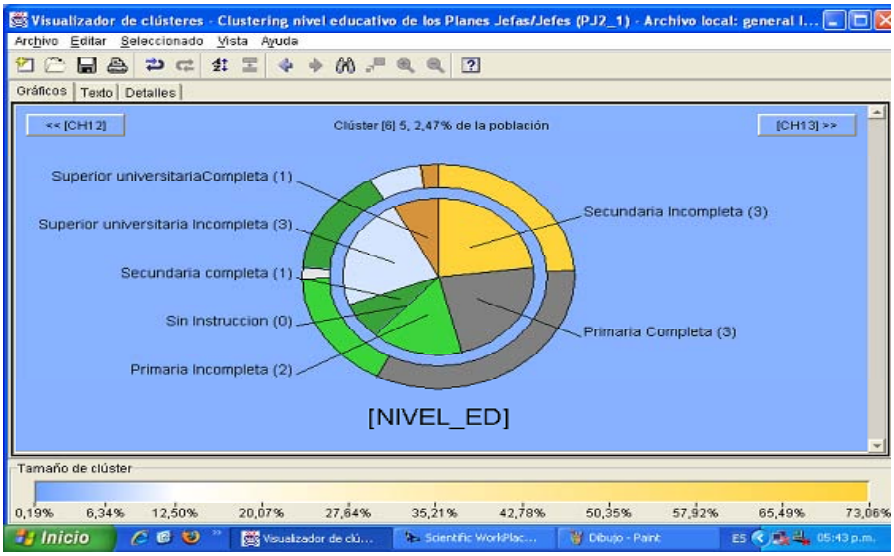

Figura 6.121: El nivel educativo alcanzado por estos individuos es *"secundaria incompleta", "primaria completa" y "superior universitaria incompleta".*

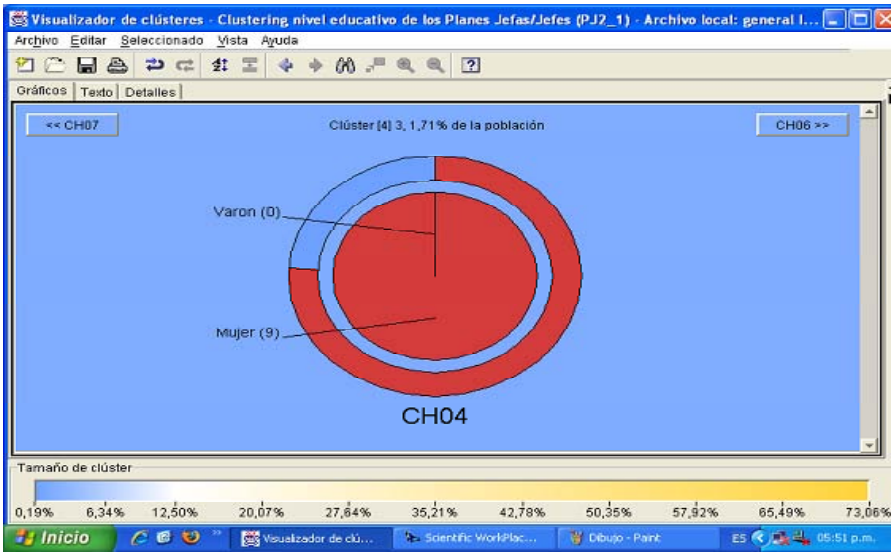

Figura 6.122: Visualización de variable CH04 (sexo).

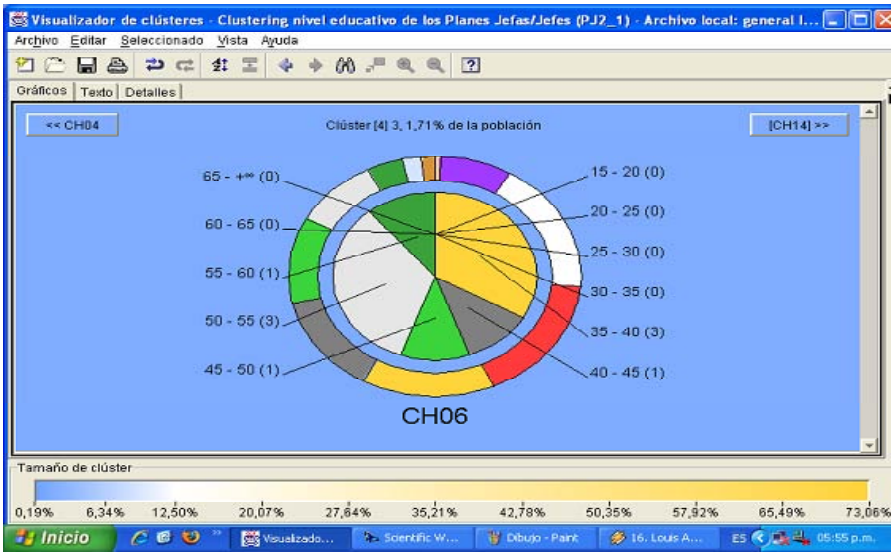

Figura 6.123: Muestreo del resultado de la variable CH06 (años).

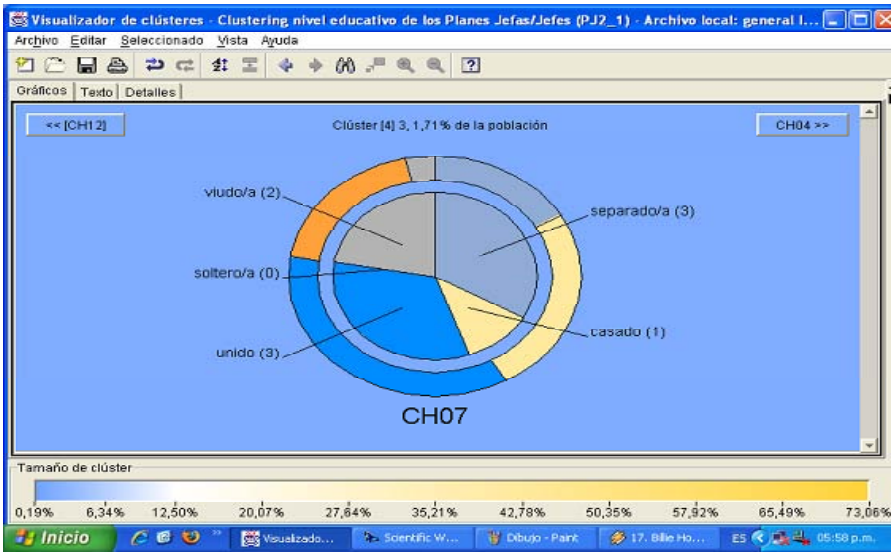

Figura 6.124: Visualización del resultado obtenido de la variable CH07 (estado civil).

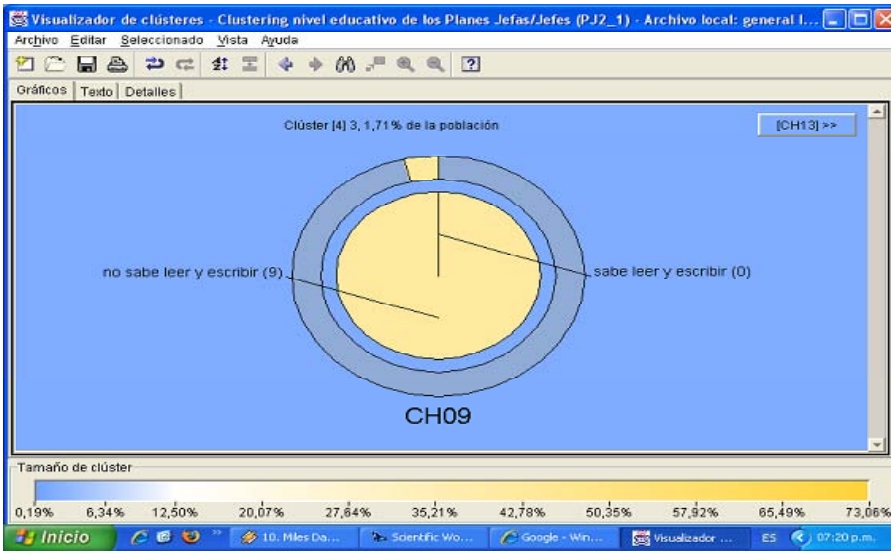

Figura 6.125: Muestreo del contenido de la variable CH09 (Analfabetismo).

la pág. 216.

A diferencia del cuarto clúster (ver fig. 6.125 de la pág. 214) en esté la opción con mayor presencia en la variable CH09 (Analfabetismo) como se puede apreciar en el siguiente (ver fig. 6.129 de la pág. 216) es la de "sabe leer y escribir".

En la fig. 6.130 de la pág. 217 se puede comprobar que la opcion "No asiste, pero asistió" es la predominante en la variable CH10 (¿Asiste o asistió a algún establecimiento educativo?).

El nivel educativo predominante en esté clúster como se puede apreciar en la siguiente fig. 6.129 de la pág. 216 es de "secundaria incompleta".

Teniendo como nivel máximo cursado por estos individuos el nivel "secundario" como se puede visualizar en la siguiente fig. 6.129 de la pág. 216.

Por último se puede apreciar en la siguiente fig. 6.129 de la pág. 216 que la opción "primer" año es que posee mayor representación en la variable CH14 (¿Cuál fue el último año que aprobó?).

La sexta agrupación de 0,57 % de la población total, en ella se puede

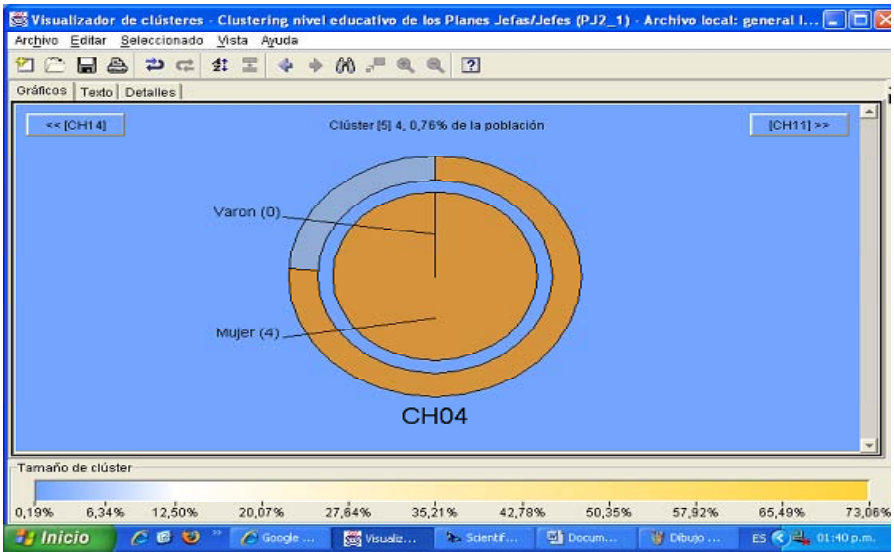

Figura 6.126: El sexo femenino es el predominante en el quinto clúster (0,76 % de la población total).

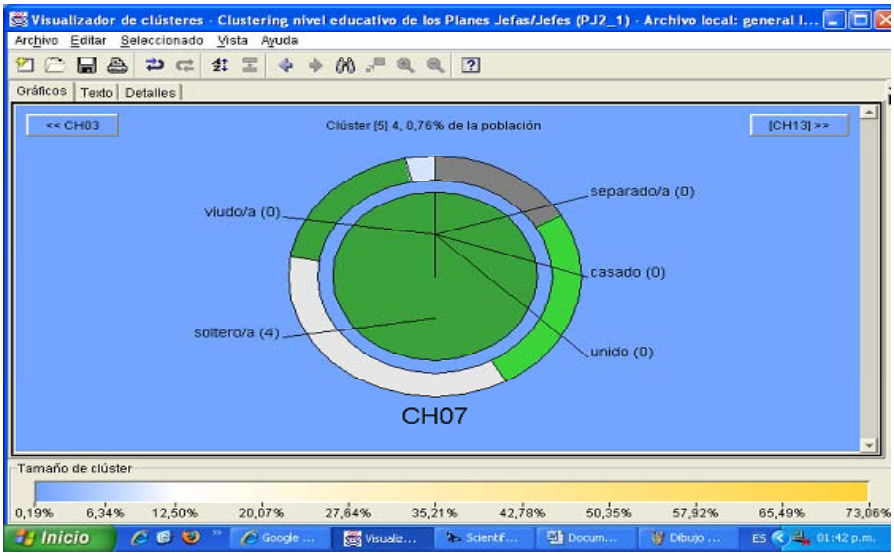

Figura 6.127: Muestreo del resultado de la variable CH07 (estado civil).

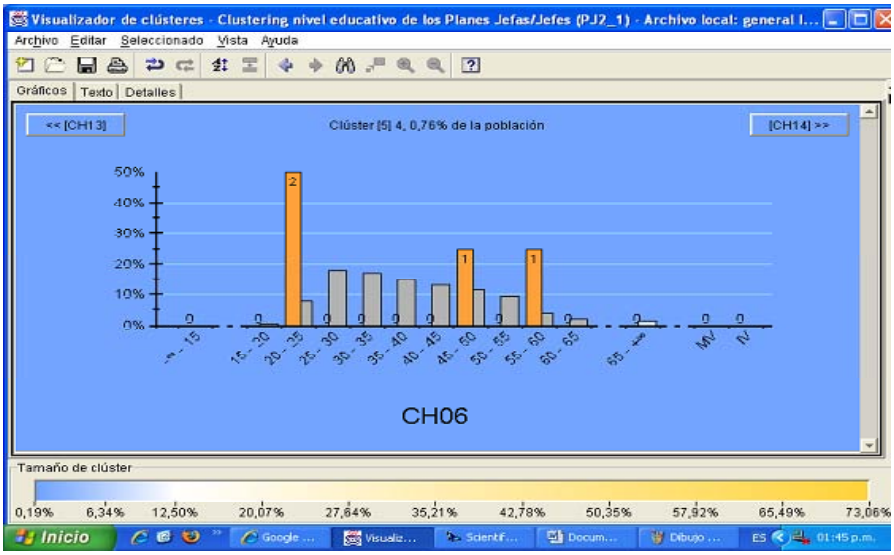

Figura 6.128: Visualización de los rangos de edades del quinto clúster (0,76 % de la población total).

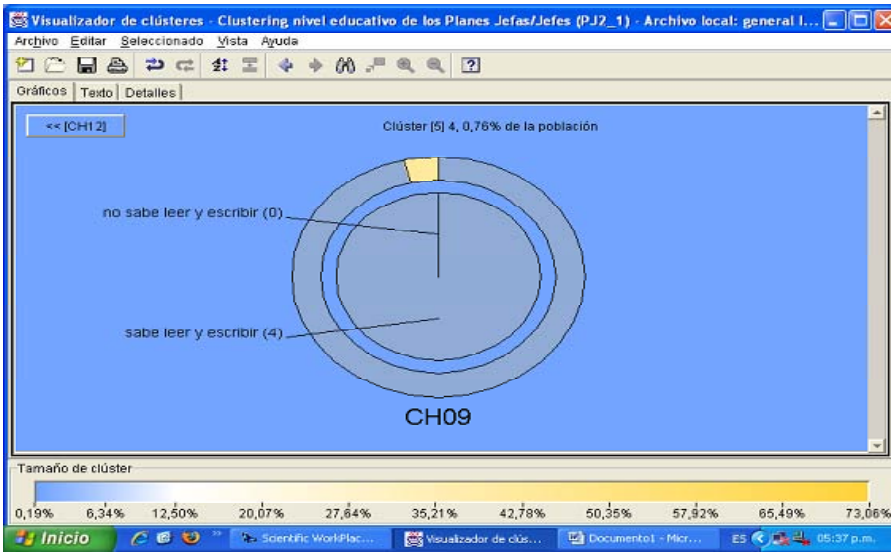

Figura 6.129: La opción "sabe leer y escribir" es la de mayor frecuencia en la variable CH09 (Analfabetismo) a diferencia del clúster Nº4 que posee un elevado índice de analfabetismo.

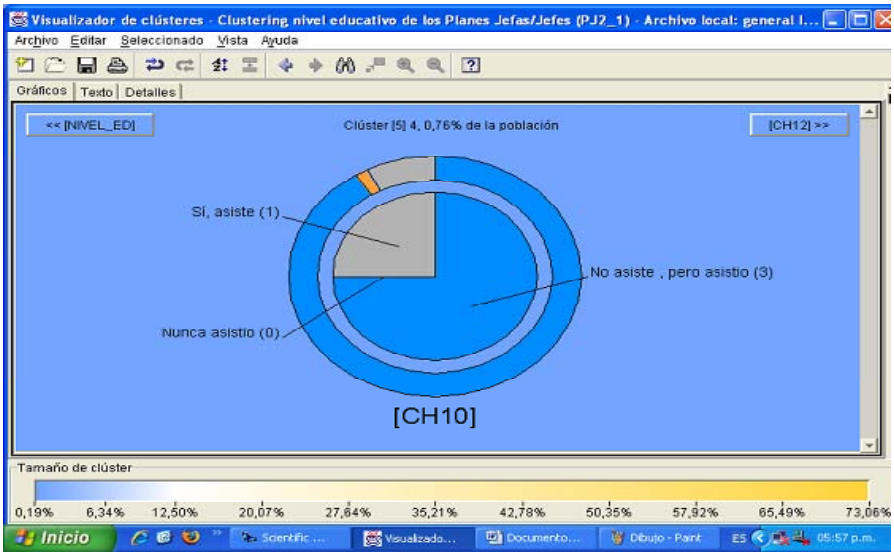

Figura 6.130: La opcion "No asiste, pero asistió" es la predominante en la variable CH10 (¿Asiste o asistió a algún establecimiento educativo?).

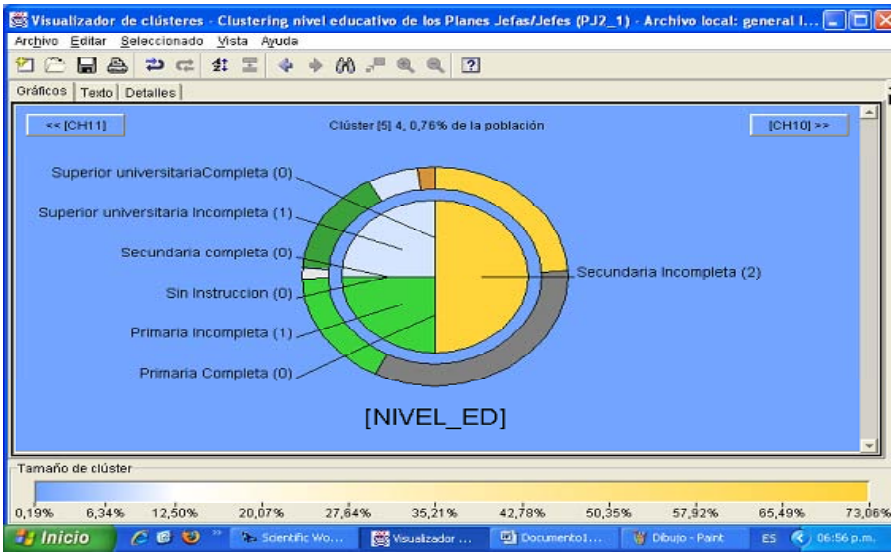

Figura 6.131: Secundaria incompleta es el nivel educativo predominante en el clúster numero N^{o}5.

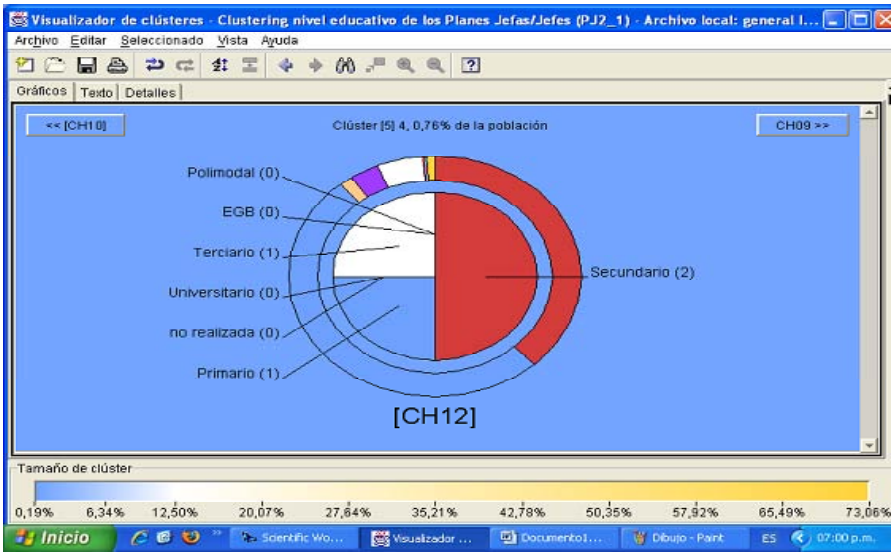

Figura 6.132: Visualización del resultado de la variable CH12 (¿Cuál es el nivel más alto que cursa o cursó?).

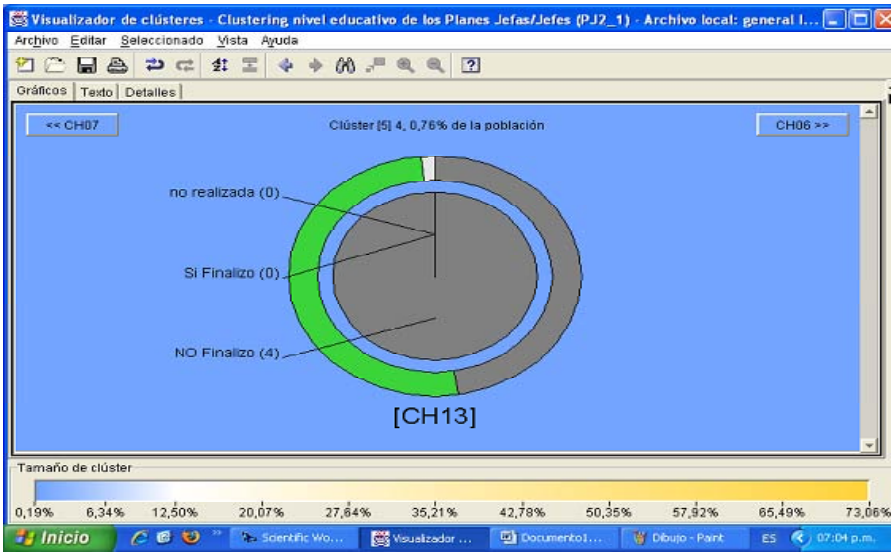

Figura 6.133: Muestreo del resultado de la variable CH13 (¿Finalizó ese nivel?).

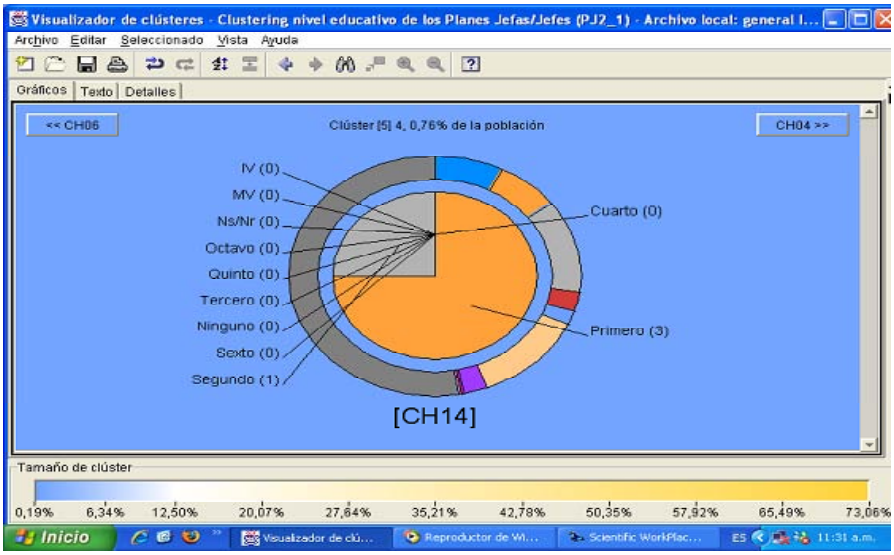

Figura 6.134: La opción "primer" año es que posee mayor representación en la variable CH14 (¿Cuál fue el último año que aprobó?).

visualizar que sexo *masculino* es el predominante con un estado civil *separado* con los respectivos rangos de edades [45-50], [60-65] y [65-∞+] (ver fig. 6.135 de la pág. 220), (ver fig. 6.136 de la pág. 220), respectivamente (ver fig. 6.137 de la pág. 221).

En cuanto a la formación educativa de estas personas se puede apreciar en la siguiente (ver fig. 6.138 de la pág. 221) que poseen un elevado índice de analfabetismo y un nivel educativo sin instrucción (ver fig. 6.139 de la pág. 222).

En la variable CH10 ¿Asiste o Asistió a algún establecimiento educativo? (colegio, escuela, universidad) se puede observar a la opción sobresaliente de *"Nunca asistió"* (ver fig. 6.140 de la pág. 222).

En la séptima agrupación también con un 0,57 % de la población total como se puede observar la fig. 6.141 de la pág. 223 el sexo *femenino* es predominante con un estado civil *soltero/a* con un rango de edad *[20-25]* (ver fig. 6.142 de la pág. 223), (ver fig. 6.143 de la pág. 224), respectivamente (ver fig. 6.144 de la pág. 224).

Al igual que la formación académica de la anterior agrupación (ver fig.

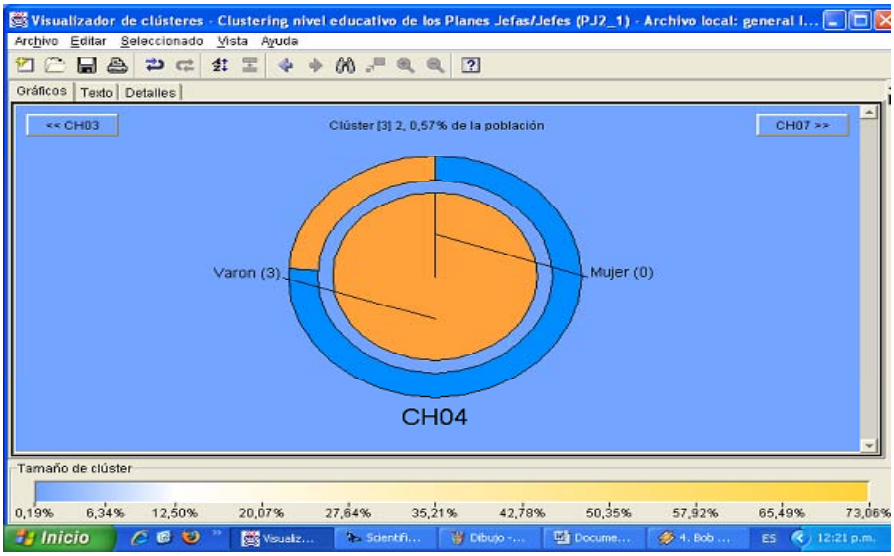

Figura 6.135: En sexo predominante es el *masculino* en el sexto clúster de 0,57 % de la población total.

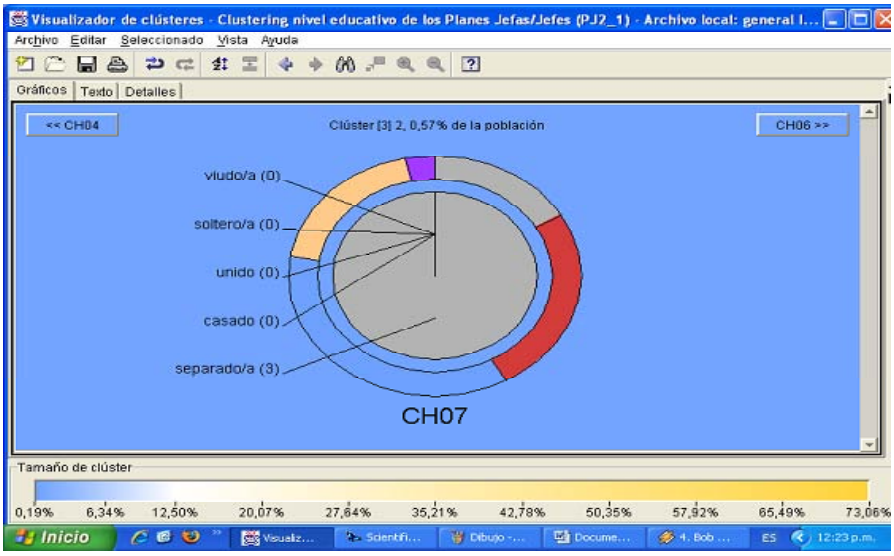

Figura 6.136: La opción separado es la que posee mayor representación en la variable CH07 (estado civil).

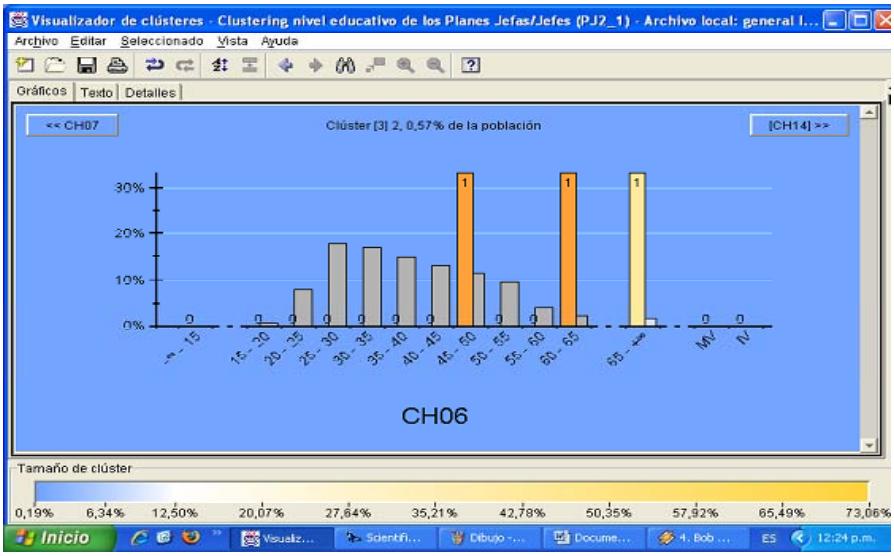

Figura 6.137: Visualización del resultado de la variable CH06 (años) en formato histograma.

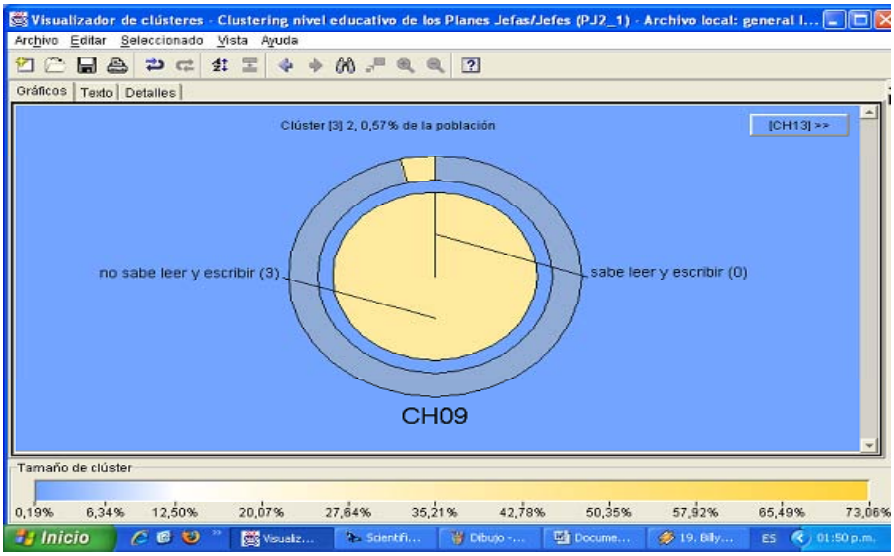

Figura 6.138: Visualización de la opción “No sabe leer y escrivir” es la predominante en este clúster.

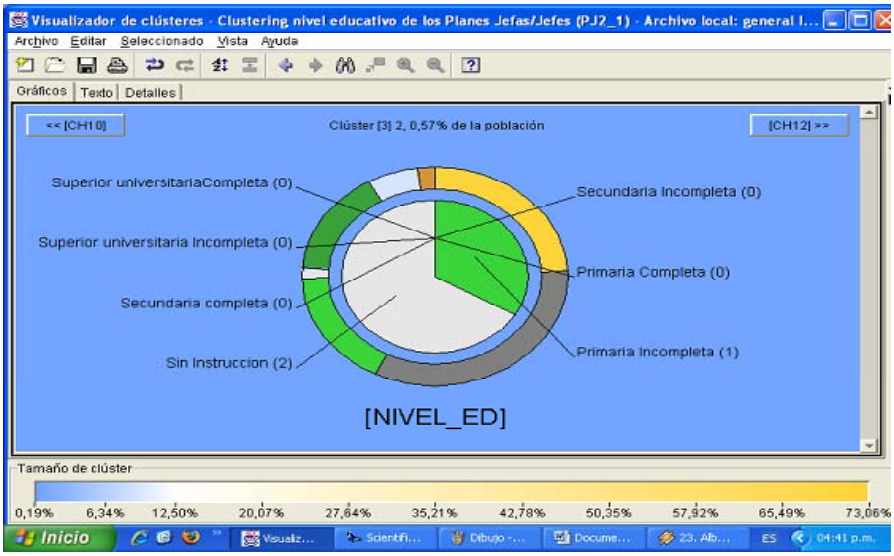

Figura 6.139: Visualización del nivel educativo "sin instrucción" en la variable NIVEL_ED (Nivel Educativo).

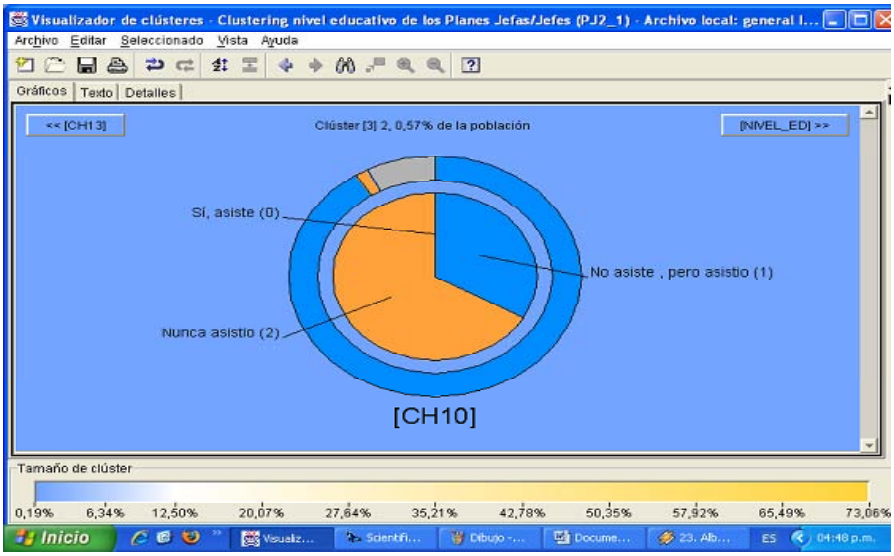

Figura 6.140: En el clúster Nº6 se puede observar que estos individuos no poseen instrucción educativa.

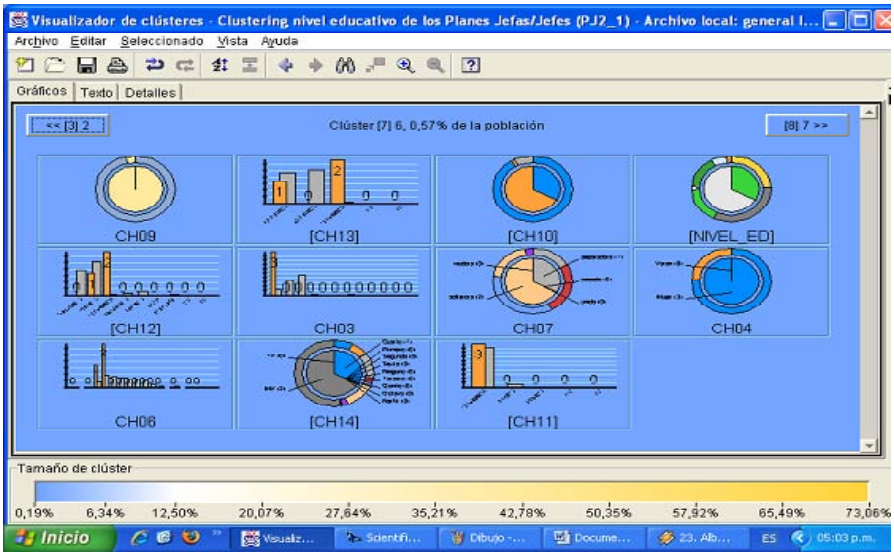

Figura 6.141: Visualización general del séptimo clúster con un 0,57 % de la población total.

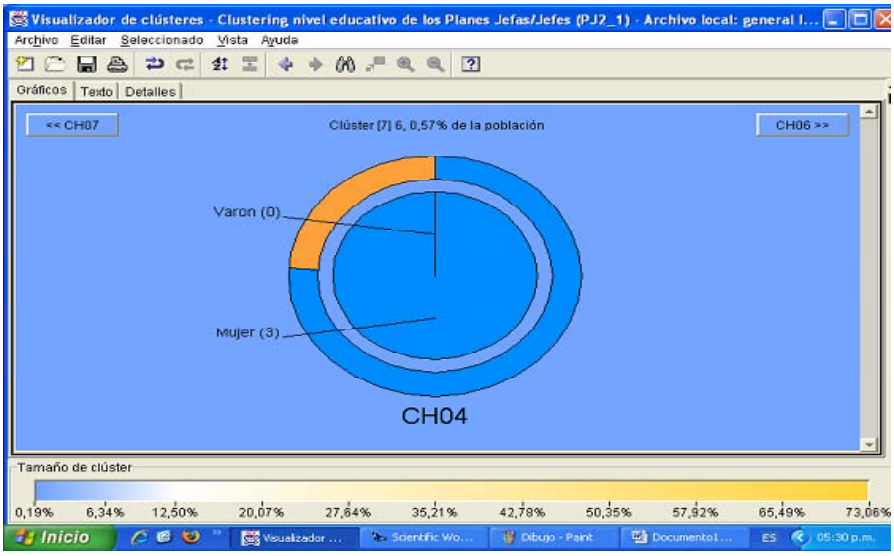

Figura 6.142: La opción “mujer” posee mayor presencia en la variable CH04 (sexo) del clúster N^{o}7.

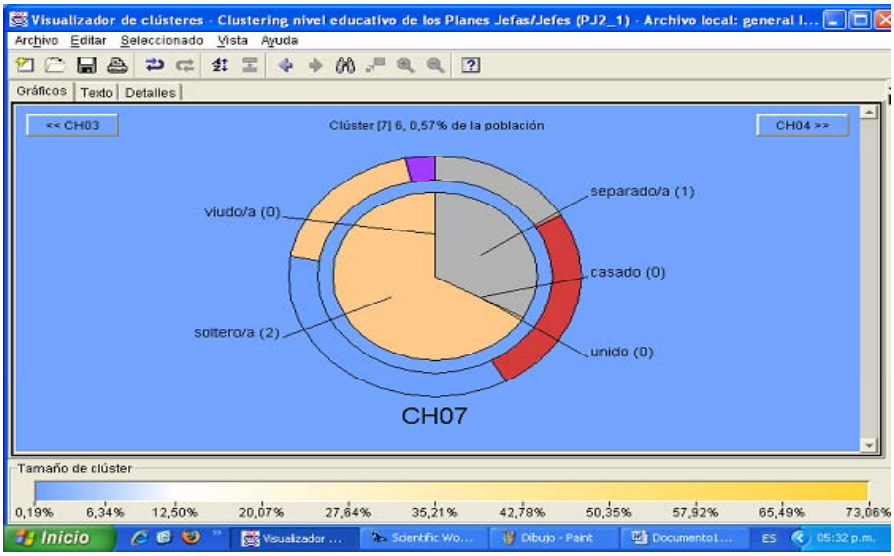

Figura 6.143: Visualización del resultado en formato de diagrama circular de la varible CH07(estado civil).

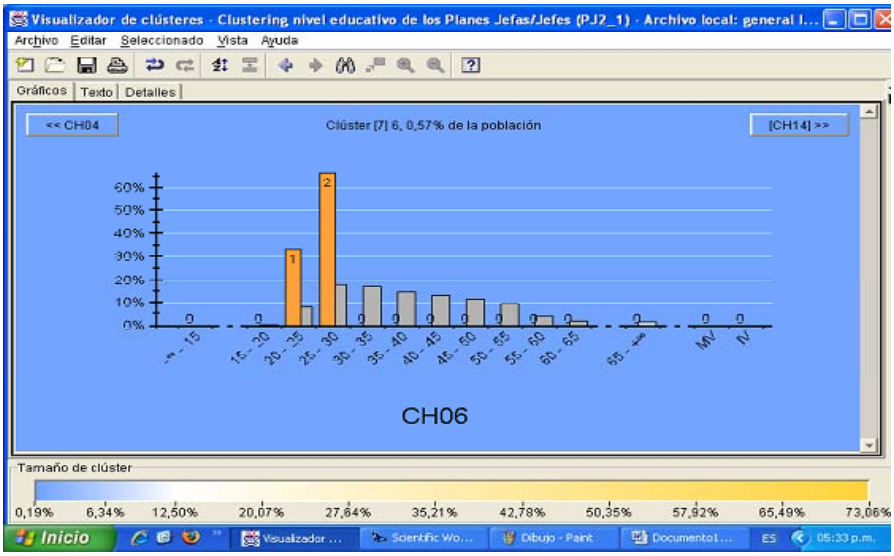

Figura 6.144: El rango de edad [20-25] años es el predominante en la variable CH06 (años) de la séptima agrupación.

6.138 de la pág. 221) estás personas *no saben leer ni escribir* ya que nunca han asistido a un establecimiento educativo teniendo un nivel educativo sin instrucción (ver fig. 6.145 de la pág. 225), (ver fig. 6.146 de la pág. 226), respectivamente (ver fig. 6.147 de la pág. 226).

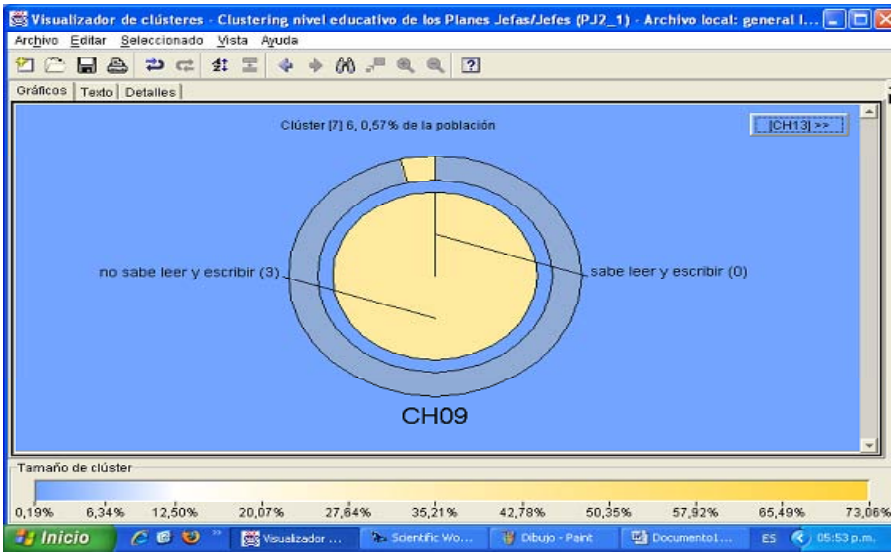

Figura 6.145: Muestreo del diagrama circular de la variable CH09 (Analfabetismo) con su correspondiente numero de analfabetos.

En la octava y última agrupación (0,19% de la población total) en ella se puede visualizar al sexo *masculino* que es el predominante con un estado civil *separado* con el rango de edad *[45-50] años* (ver fig. 6.148 de la pág. 227), (ver fig. 6.149 de la pág. 227), respectivamente (ver fig. 6.150 de la pág. 228).

En ésta la última agrupación se puede observar que posee el nivel educativo más elevado de todos los clúster antes vistos como se puede comprobar en la fig. 6.151 de la pág. 228.

Asimismo se puede observar en la fig. 6.152 de la pág. 229 que en esta agrupación se posee el máximo año aprobado con respecto a los demás clúster.

Siendo el nivel educativo más elevado que cursó esta persona el *universitario* como se puede apreciar en la siguiente fig. 6.153 de la pág. 229.

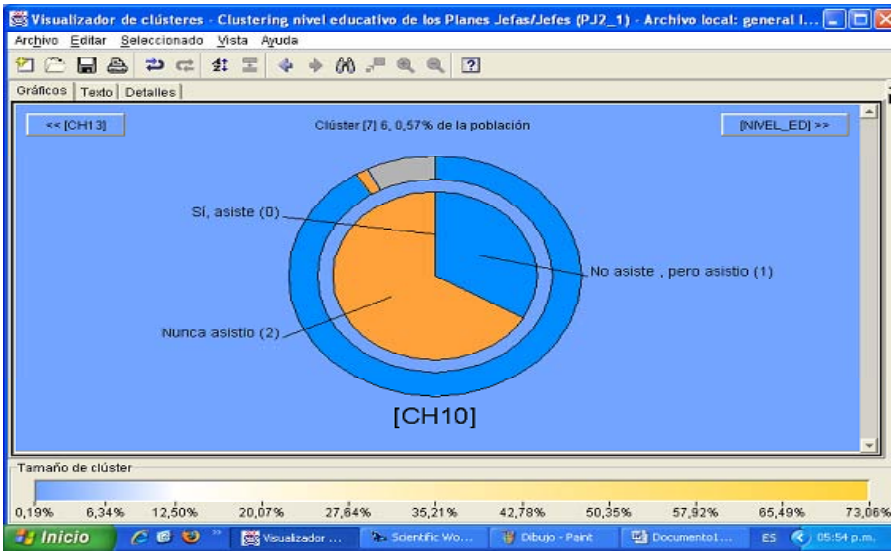

Figura 6.146: La opción "nunca asistió" es la de mayor representación en la variable CH10 (¿Asiste o asistió a algún establecimiento educativo: colegio, escuela, universidad?)

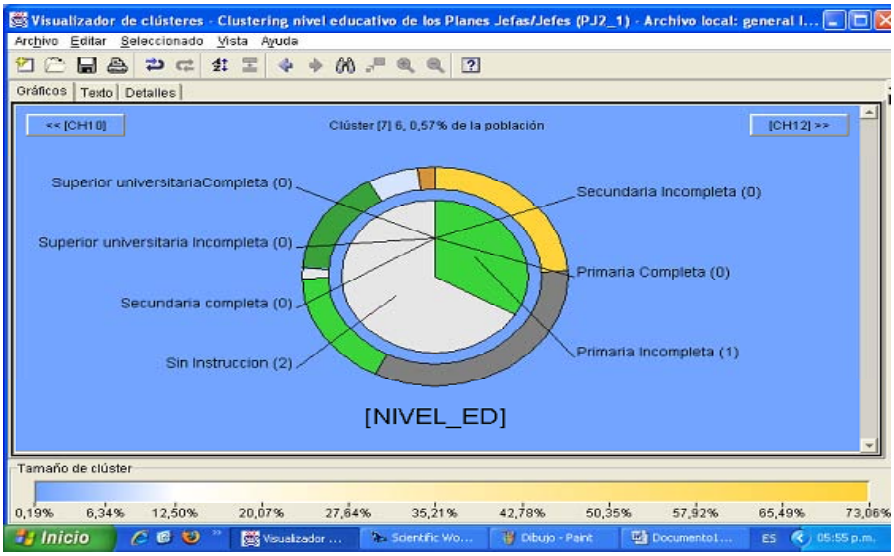

Figura 6.147: El nivel educativo en la séptima agrupación posee un nivel de *sin instrucción.*

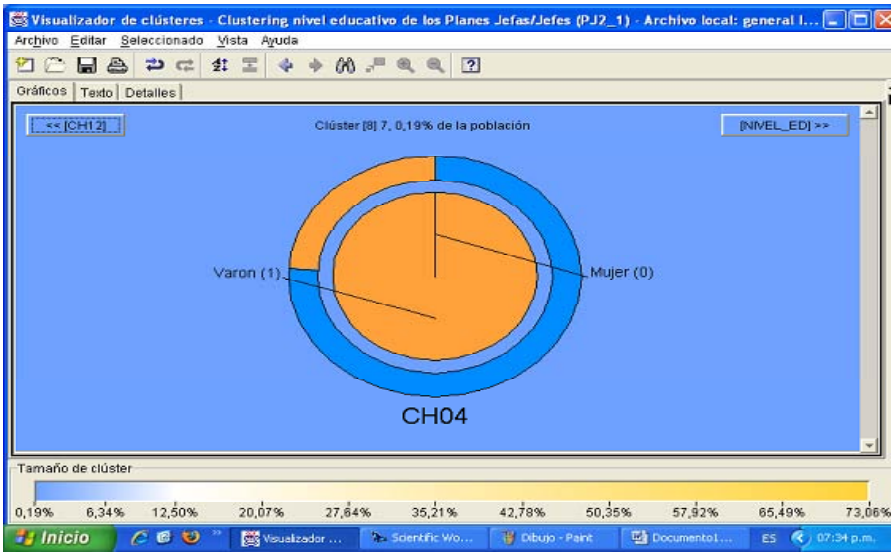

Figura 6.148: Muestreo del resultado en formato de diagrama circular de la variable CH04 (sexo).

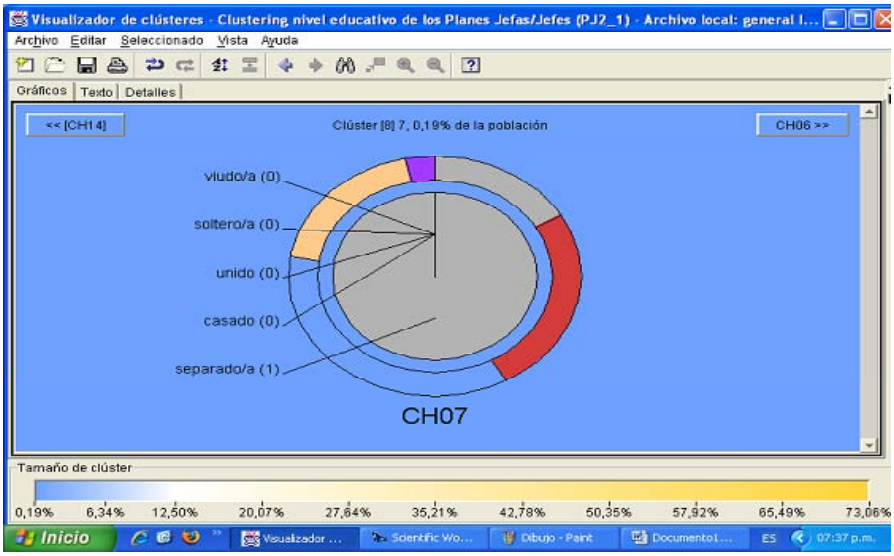

Figura 6.149: La opcíon “separado/a” es de mayor predomino en la variable CH07 (estado civil).

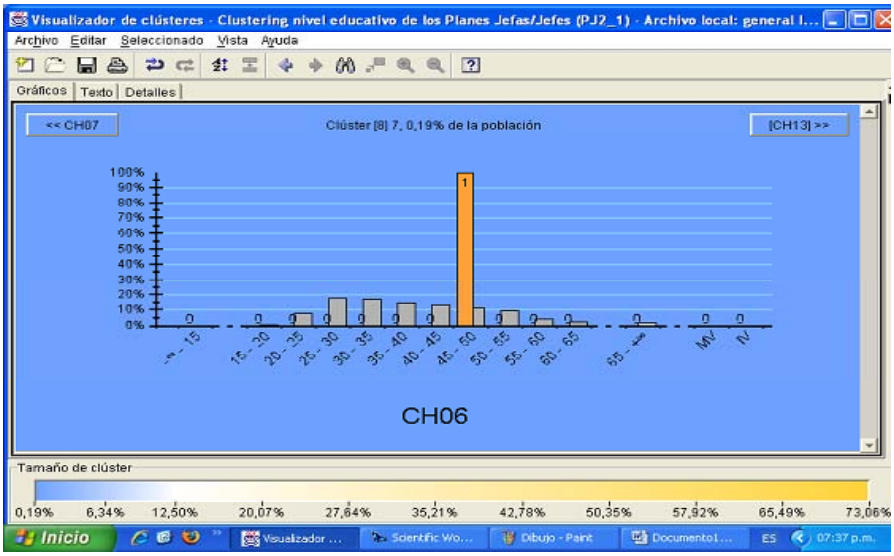

Figura 6.150: Histograma que representa la distribución de las edades en el clúster N^{o}8.

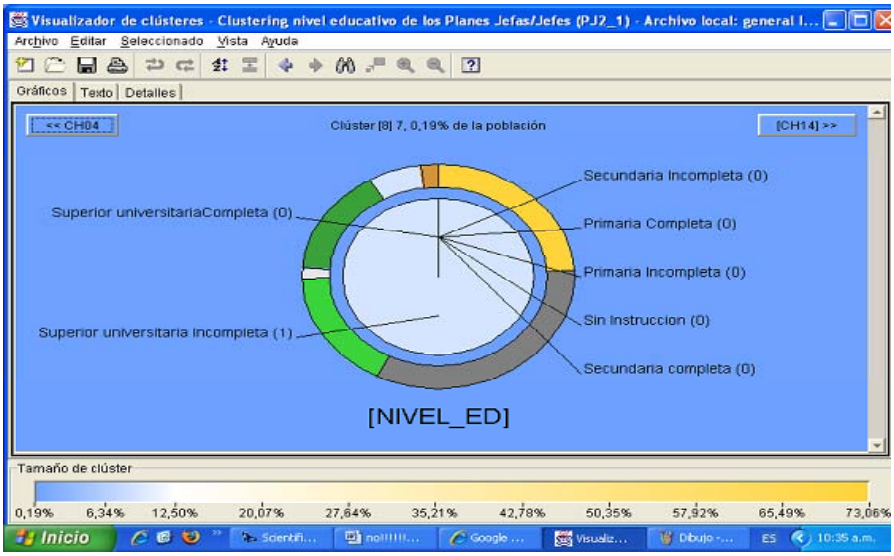

Figura 6.151: El nivel educativo en la octava y última agrupación (0,19% de la población total), posee un nivel de superior universitaria incompleta.

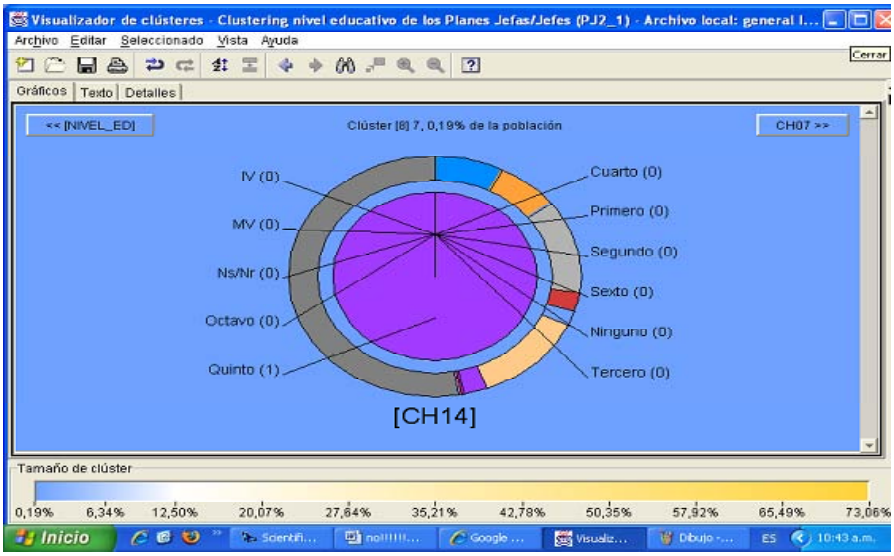

Figura 6.152: Muestreo del resultado de la variable CH14 (¿Cuál fue el último año que aprobó?).

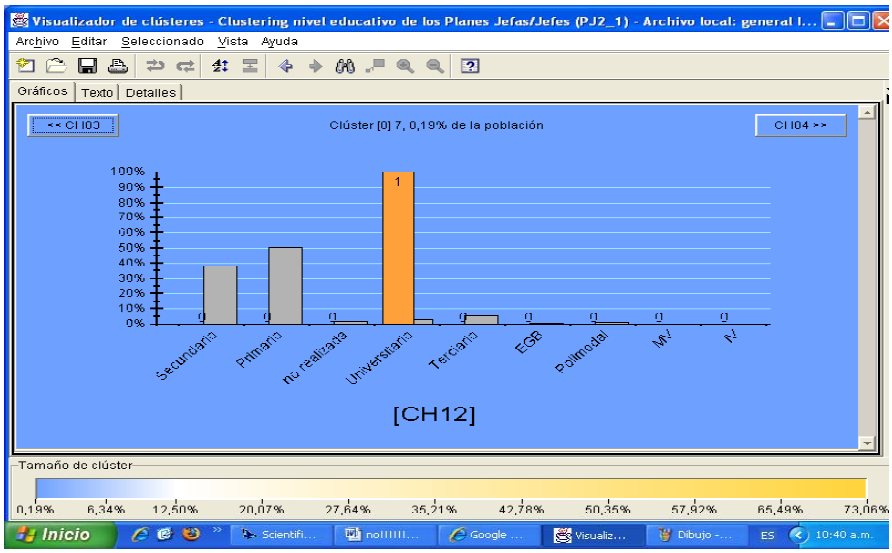

Figura 6.153: Visualización del resultado obtenido de la variable CH14 (¿Cuál fue el ultimo año que aprobo?) del octavo clúster.

Clasificación del Ingreso de Cada Individuo, en Base a sus Principales Características Sociodemográficas

Luego de obtener una visión detallada de 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.

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*.

En el apartado *"Introducción de Intelligent Miner for Data"* se describe con mayor precisión dicha técnica.

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), como se puede observar en la siguiente fig. 8.8 de la pág. 327.

Como se puede observar en la siguiente fig. 8.8 de la pág. 327, en cada nodo del árbol de decisión se evalúa un atributo.

Existe una rama por cada valor del atributo cuando los atributos son discretos y una rama por rango de valores cuando los atributos son continuos.

Nótese que en cada nivel, la rama que deriva a la izquierda es si, la derecha no.

A continuación se visualizarán las siguientes reglas extraídas del árbol de decisión (ver fig. 6.155 de la pág. 232).

En la fig.6.155 de la pág. 232 se puede visualizar así como la regla, también el valor de la etiqueta que es el ingreso total individual que en este caso posee un valor de *448.11*, como el número de registros *131* que cumplen con las características de dichos atributos.

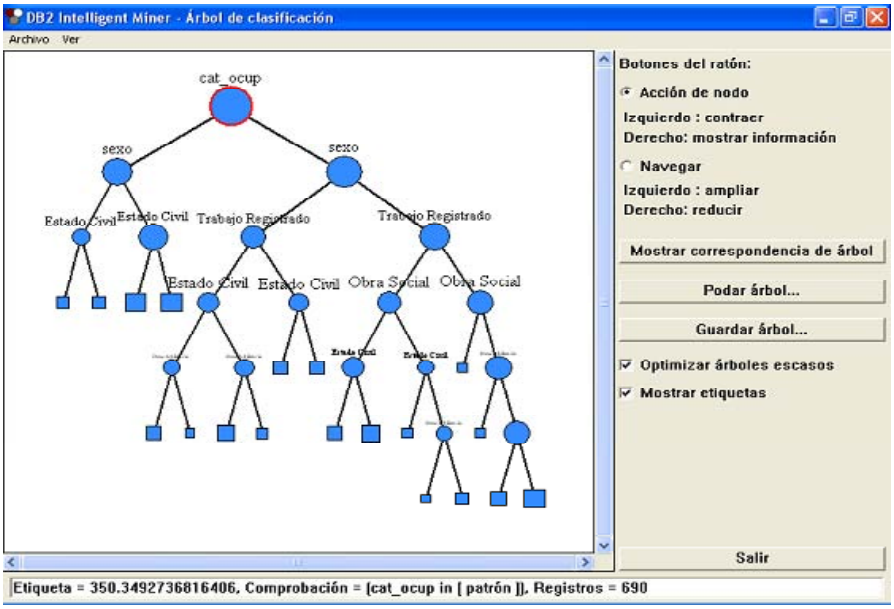

Figura 6.154: Visualización de las diecinueve reglas de que identifican los distintos nodos de del árbol.

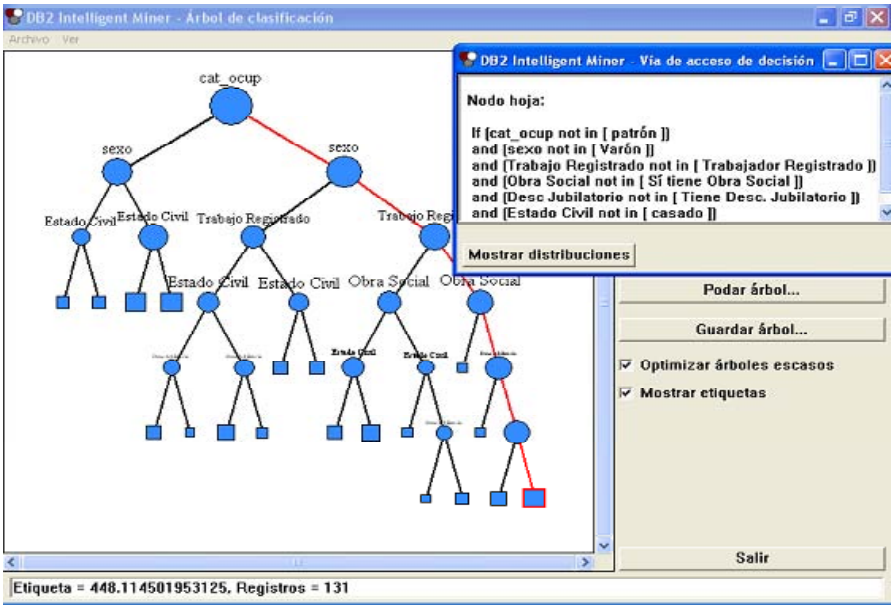

Figura 6.155: Regla N^o 1 Si el individuo de estudio es de sexo femenino, no es patrón, no tiene trabajo registrado, ni obra social, ni descuento jubilatorio y su estado civil no es el casados entonces el ingreso total individual es de 448.11.

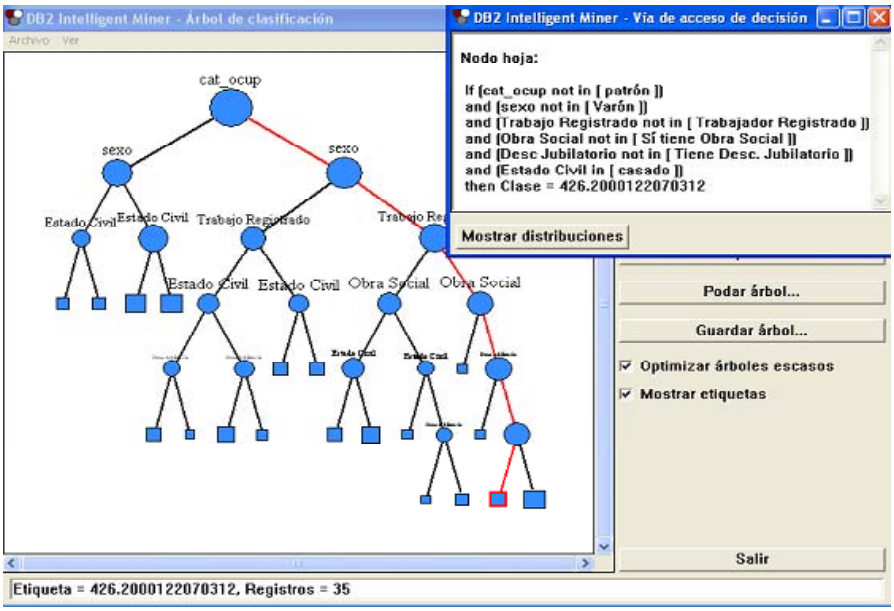

Figura 6.156: Regla Nº 2 Si el individuo de estudio es de sexo femenino, no es patrón, no tiene trabajo registrado, ni obra social, ni descuento jubilatorio y su estado civil es el casados entonces el ingreso total individual es de 426.20.

Como se puede apreciar en la fig.6.156 de la pág. 233, el valor de la etiqueta es *426.20* siendo este el ingreso total individual, también se puede observar que el número de registros involucrados en dicha regla es de *35.*

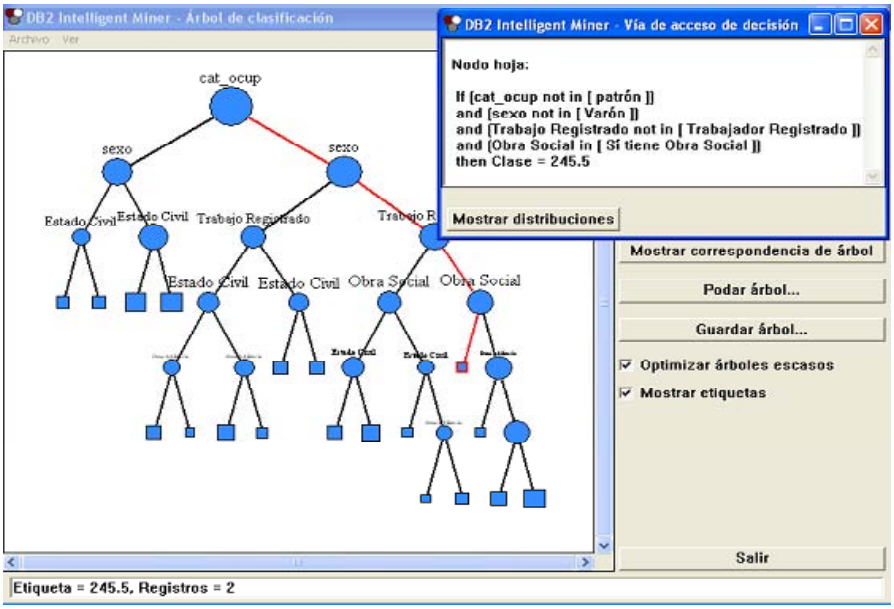

Figura 6.157: Regla N^o 3 Si el individuo de estudio es de sexo femenino, no es patrón, no tiene trabajo registrado, pero sí posee obra social, entonces el ingreso total individual es de 245.5.

El la fig.6.157 de la pág. 234 se puede observar que esta rama no posee dos niveles, esto significa que en la regla faltarán dos atributos *Estado Civil* y *Descuento Jubilatorio* precisamente.

Puede observarse en la fig.6.158 de la pág. 235 que la rama del árbol en el nivel N^{o}3 tiene hacia la izquierda lo que implicará la afirmación de dicho atributo en este caso trabajo registrado.

El la fig.6.159 de la pág. 236 se puede apreciar los 150 que es el valor del ingreso total individual, también se puede observar que existe un único registro que cumple con dichas características.

En la fig. 6.160 de la pág. 237 se puede visualizar que la rama no pasa por el nodo del atributo descuento jubilatorio y por ende se obtiene como resultado una regla más pequeña.

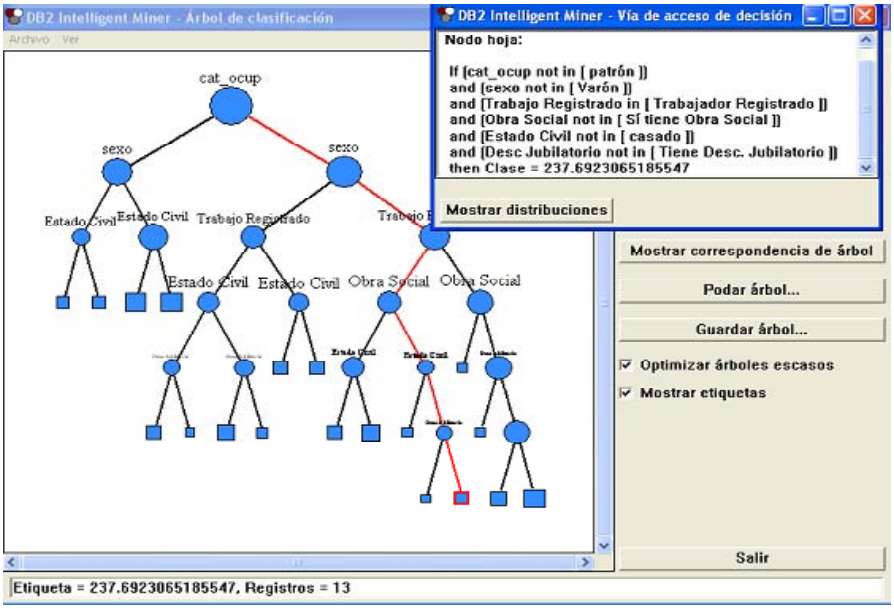

Figura 6.158: Regla N^o 4 Si el individuo de estudio es de sexo femenino, goza de un trabajo registrado, no es patrón, no posee obra social, ni descuento jubilatorio y su estado civil no es casados entonces el ingreso total individual es de 237.69.

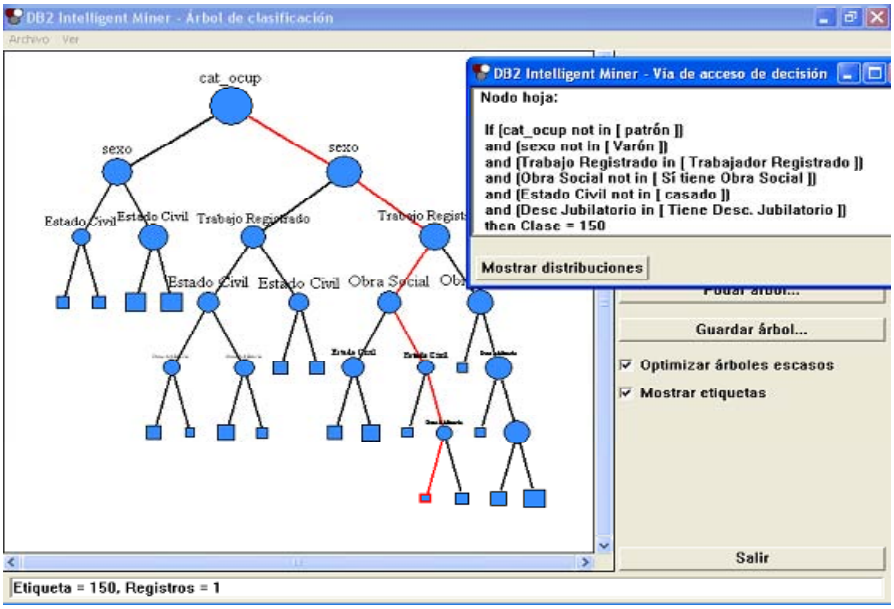

Figura 6.159: Regla Nº 5 Si el individuo de estudio es de sexo femenino, goza de un trabajo registrado, pero no posee descuento jubilatorio, ni posee obra social a su vez no es patrón y su estado civil no es casado entonces el ingreso total individual es de 150.

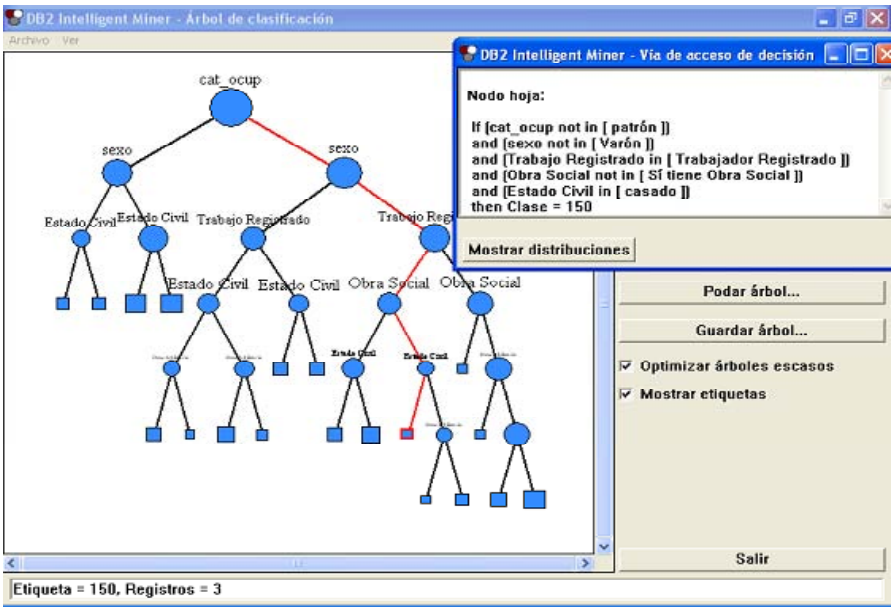

Figura 6.160: Regla N^o 6 Si el individuo de estudio es de sexo femenino, goza de un trabajo registrado, pero no posee obra social a su vez no es patrón y su estado civil es el de casado entonces el ingreso total individual es de 150.

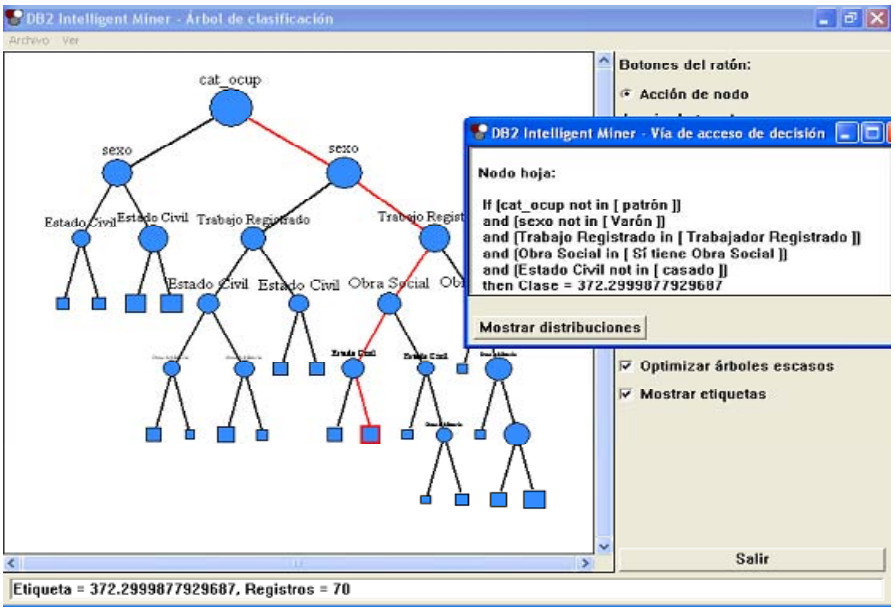

Figura 6.161: Regla Nº 7 Si el individuo de estudio es de sexo femenino, goza de un trabajo registrado y posee obra social a su vez no es patrón y su estado civil no es el de casado entonces el ingreso total individual es de 372.30.

Además de apreciar en la fig. 6.161 de la pág. 238 el valor del ingreso total 372.30 también se puede visualizar el número total de registros que en este caso son 70.

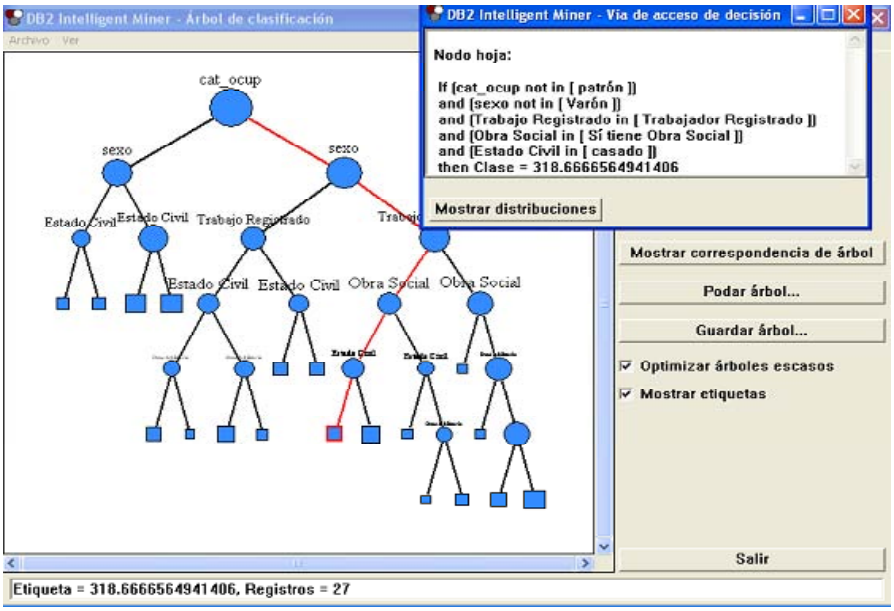

Figura 6.162: Regla N^o 8 Si el individuo de estudio es de sexo femenino, goza de un trabajo registrado y posee obra social a su vez no es patrón y su estado civil es el de casado entonces el ingreso total individual es de 318.66.

En el siguiente en la fig. 6.162 de la pág. 239 se puede visualizar así como la regla el valor de la etiqueta que es el ingreso total individual, como el número de registros que cumplen con las características de dichos atributos.

En este caso como se puede apreciar en la 6.163 de la pág. 240, el ingreso total individual es de 594.86 con los 28 registros que cumplen con esas condiciones.

A diferencia de la anterior fig.6.163 de la pág. 240 en la 6.164 de la pág. 241 se puede apreciar que la rama del árbol en el ultimo nodo atributo tiende hacia la derecha lo que implica que el individuo es casado.

En la fig.6.165 de la pág. 242 se puede apreciar además de las reglas obtenidas el valor del ingreso total individual que es de 502.5 y el numero de registros involucrados en dicha relación que en este caso son únicamente 4.

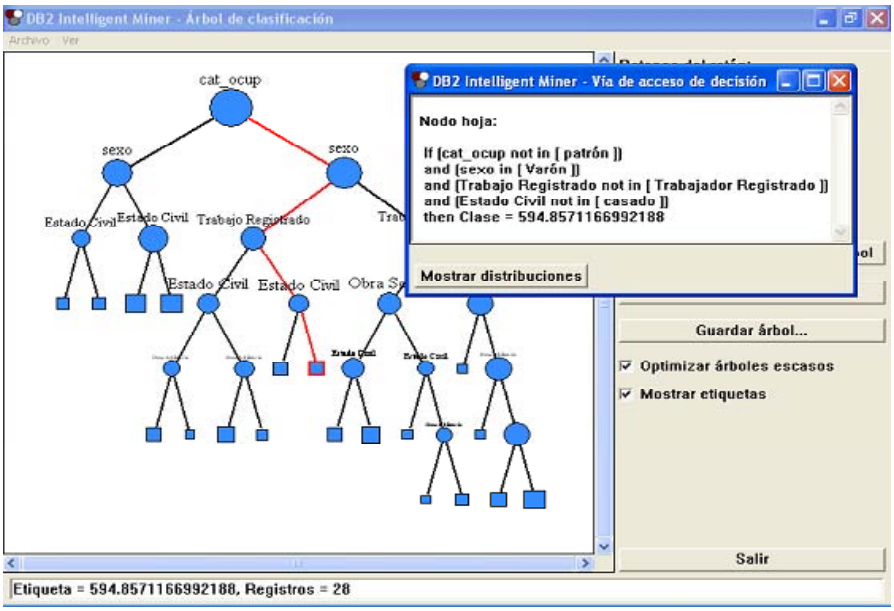

Figura 6.163: Regla Nº 9 Si el individuo de estudio es de sexo masculino, no es patrón, no goza de un trabajo registrado y su estado civil no es casado entonces el ingreso total individual es de 594.86.

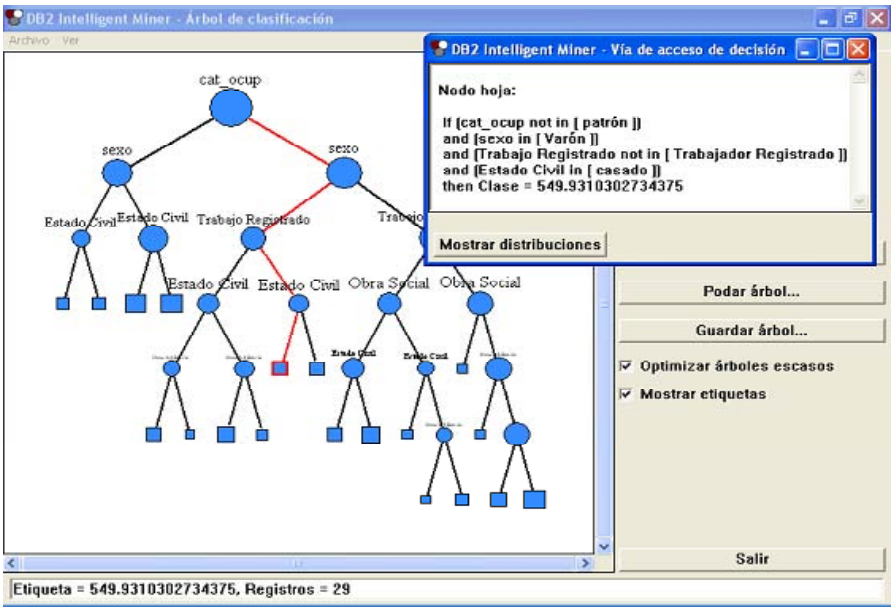

Figura 6.164: Regla N^o 10 Si el individuo de estudio es de sexo masculino, no es patrón, no posee un trabajo registrado y su estado civil es el de casado entonces el ingreso total individual es de 549.93.

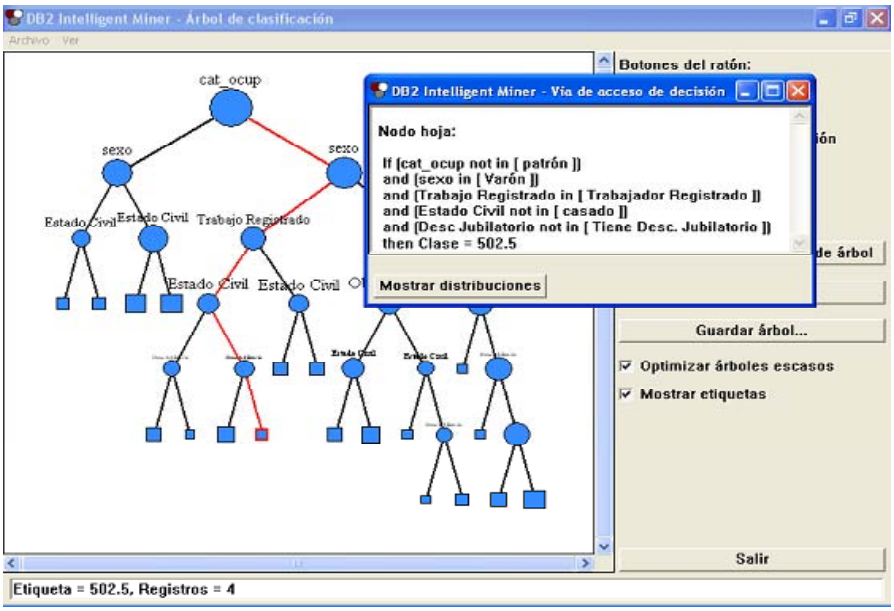

Figura 6.165: Regla Nº 11 Si el individuo de estudio es de sexo masculino, no es patrón, posee un trabajo registrado, pero no goza de descuento jubilatorio y su estado civil es el de casado entonces el ingreso total individual es de 502.5.

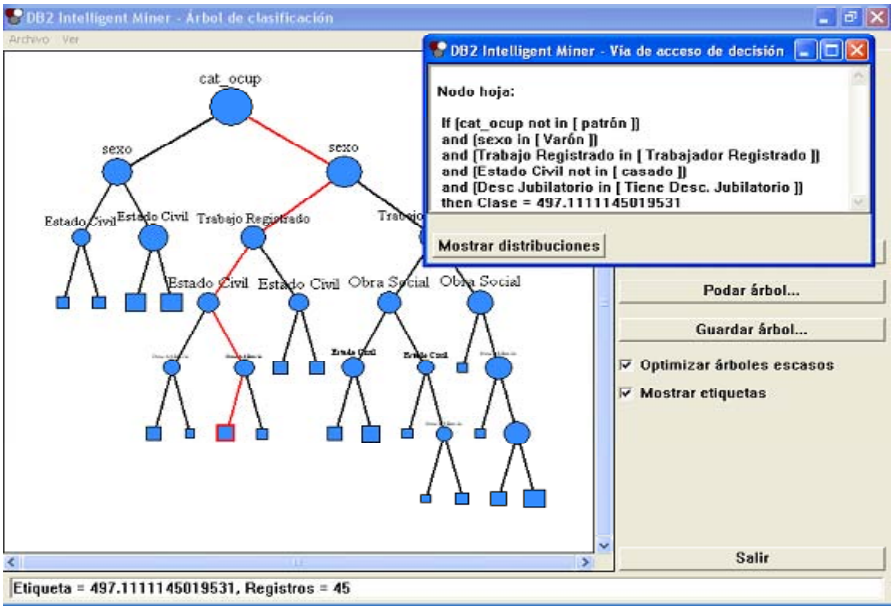

Figura 6.166: Regla N^o 12 Si el individuo de estudio es de sexo masculino, no es patrón, posee un trabajo registrado, goza de descuento jubilatorio y el estado civil no es el de casado entonces el ingreso total individual es de 497.11.

En la anterior fig.6.166 de la pág. 243 se pude visualizar el valor de la etiqueta que en este caso es el ingreso total individual tienen un monto de 497.11 con un total de 45 registros involucrados en dicha regla.

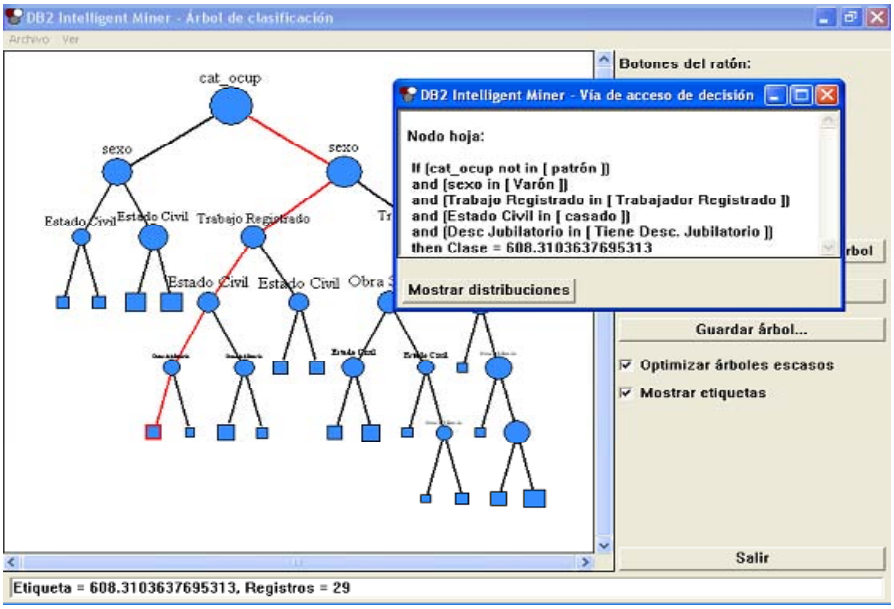

Figura 6.167: Regla N^o 14 Si el individuo de estudio es de sexo masculino, no es patrón, posee un trabajo registrado, goza de descuento jubilatorio y el estado civil es el de casado entonces el ingreso total individual es de 608.31.

Como se puede observar en la siguiente este caso es el ingreso total individual tienen un monto de 608.31 siendo esta hoja la de mayor jerarquía de todas reglas antes vistas.

Los valores obtenidos en este caso como se puede observar en la fig.6.168 de la pág. 245 son los siguientes, 203.79 para el monto total individual y 128 el número de registros que se obtienen como resultado a dicha relación.

Como podemos observar en la fig.6.168 de la pág. 245 los valores producto de dicha relación son, 170.87 para el monto total individual con la cantidad de 111 registros que cumplen con dicha regla.

En la la fig.6.168 de la pág. 245 se puede observar que el monto total individual es de 259.52 teniendo 21 registros involucrados en dicha regla.

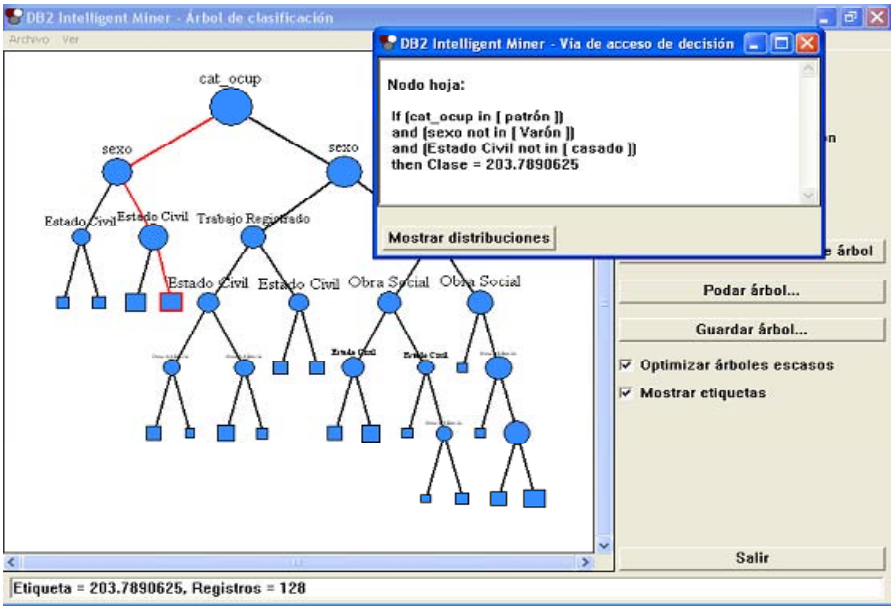

Figura 6.168: Regla N^o 15 Si el individuo de estudio es de sexo masculino, es patrón y su estado civil es el de casado entonces el ingreso total individual es de 203.79.

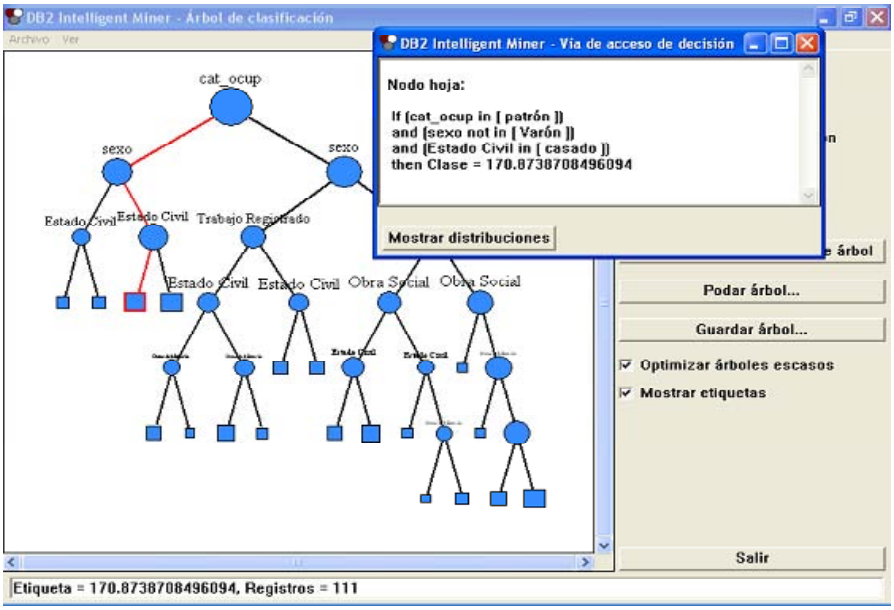

Figura 6.169: Regla Nº 16 Si el individuo de estudio es de sexo femenino, es patrón y su estado civil es el de casado entonces el ingreso total individual es de 170.87.

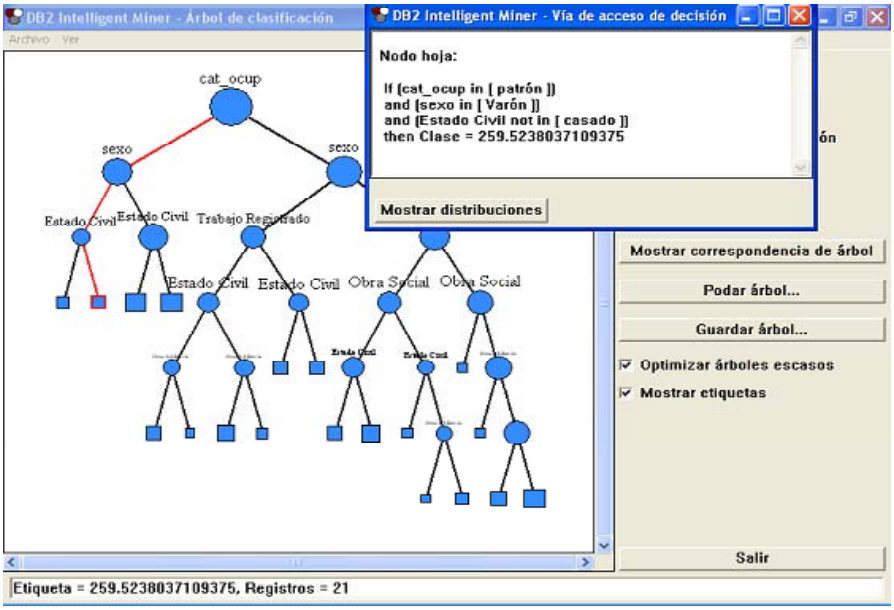

Figura 6.170: Regla N^o 17 Si el individuo de estudio es de sexo masculino, es patrón y su estado civil no es el de casado entonces el ingreso total individual es de 259.52.

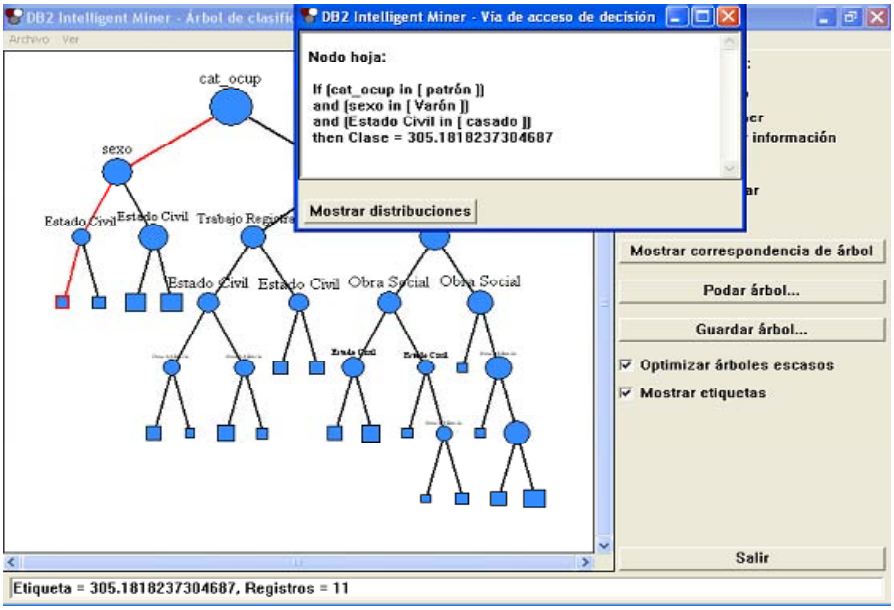

Figura 6.171: Regla N^o 18 Si el individuo de estudio es de sexo masculino, es patrón y su estado civil es el de casado entonces el ingreso total individual es de 305.18.

Como se puede observar en la siguiente de la la fig.6.168 de la pág. 245, siendo está la ultima rama extraída del árbol decisión, también se puede observar el monto total individual y su correspondientes registros involucrados.

Clasificación del Ingreso de Cada Individuo, en Base a sus Principales Características Educacionales

La principal diferencia con el anterior objeto de estudio, es que en este se tomarán a la dimensión educacional en vez de las características socioeconómica de cada individuo.

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

Se puede visualizar en la siguiente fig. 8.9 de la pág. 327 que se identifican treintaidos reglas que explican el perfil de estos individuos, determinadas por los nodos de desarrollo del árbol *(mayor cantidad de individuos y mayor pureza).*

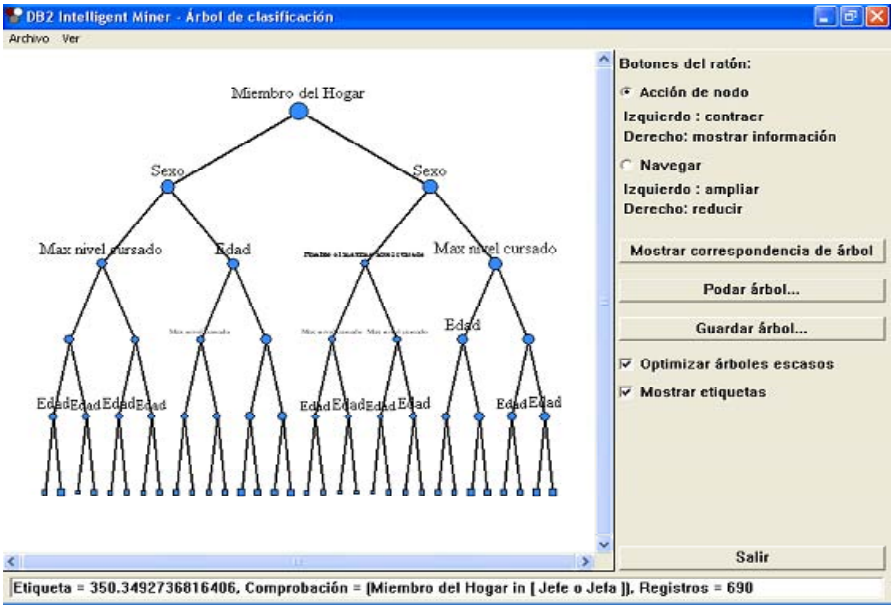

Figura 6.172: Visualización del Árbol de Decisión *"Clasificación del Ingreso de Cada Individuo, en Base a sus Principales Características Educacionales"*.

Como se puede observar en la siguiente fig. 8.8 de la pág. 327, este árbol cuenta con cuatro niveles comenzando por el nodo raíz que es el atributo miembro del hogar.

A continuación se visualizarán las siguientes reglas extraídas del árbol de decisión (ver fig. 6.173 de la pág. 250).

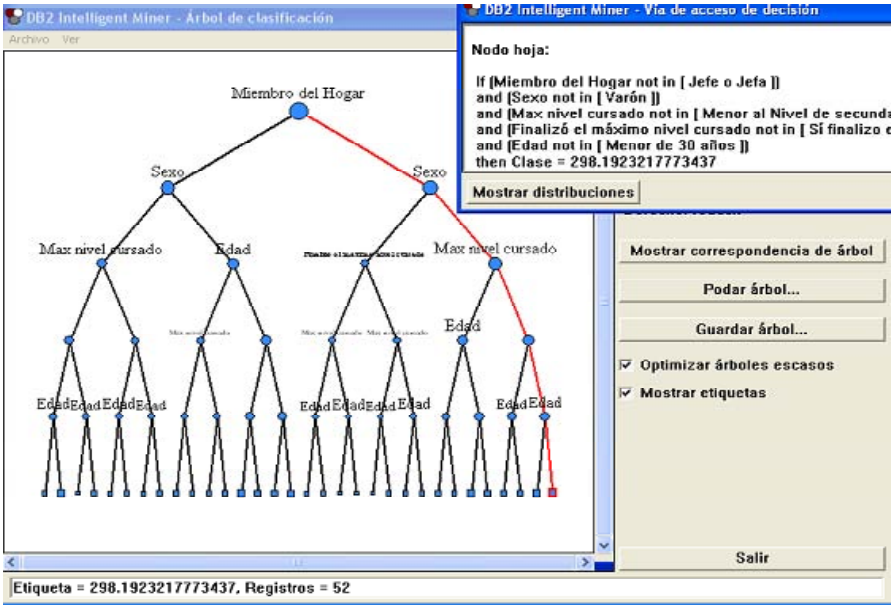

Figura 6.173: Visualizácion de la regla N^{o}1 con sus con sus respectiva rama del árbol de decisión involucrada en dicha relación.

Como se puede observar en la fig. 6.173 de la pág. 250 la regla obtenida es la siguiente, si es no jefe/jefa del hogar donde habita y no es masculino y no es menor a 30 años y no su máximo nivel cursado no es inferior al secundario y no finalizo el máximo nivel cursado entonces el ingreso es de *298.19*.

En la fig.6.174 de la pág. 251 se puede visualizar así como la siguiente regla. Si no es jefe/jefa del hogar donde habita, no es masculino es menor a 30 años su máximo nivel cursado no es inferior al secundario y no finalizó el máximo nivel cursado entonces el ingreso es de *276.13*.

También el valor de la etiqueta que es el ingreso total individual que en este caso posee un valor de *276.13*, como el número de registros *54* que cumplen con las características de dichos atributos.

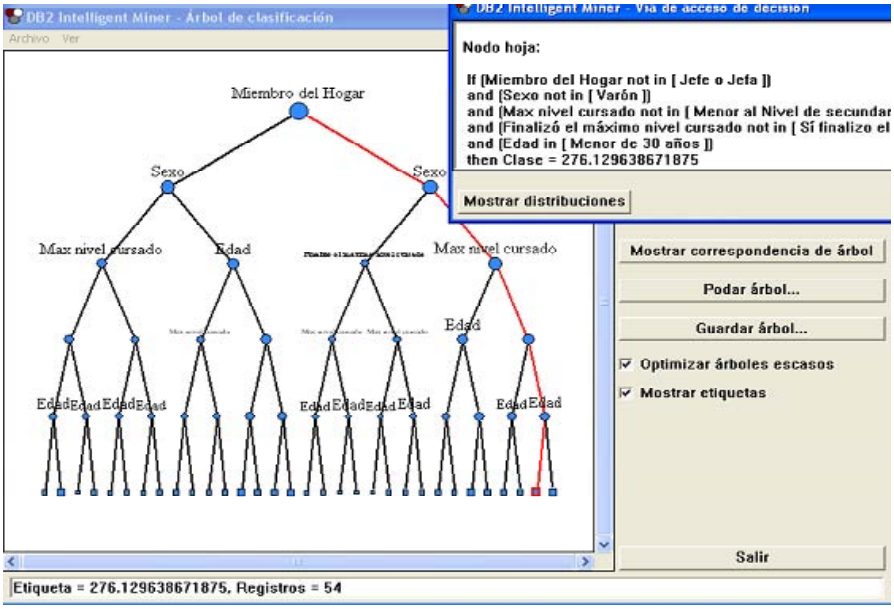

Figura 6.174: Visualizácion de la regla N^{o}2 del árbol de decisión, como así tambíen el numero de registro que cumplen con esas características.

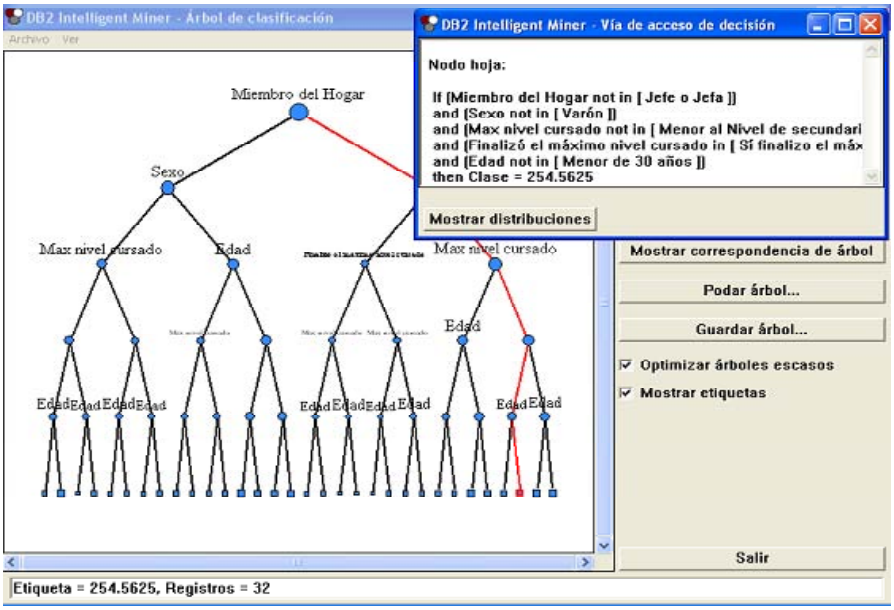

Figura 6.175: Visualizácion de la regla Nº3 del árbol de decisión, como así tambíen el numero de registro y el ingreso total individual que cumplen con esas características.

Así como se puede observar en la fig.6.175 de la pág. 252 la etiqueta que en este caso no es mas que el ingreso total individual, teniendo un valor de 254.56 con total de 32 de registros involucrados en dicha relación.

También podemos visualizar la regla N°3 que es la siguiente. Si no es jefe/jefa del hogar donde habita no es masculino, no es menor a 30 años su máximo nivel cursado no es inferior al secundario y finalizó el máximo nivel cursado entonces el ingreso es de *254.56*.

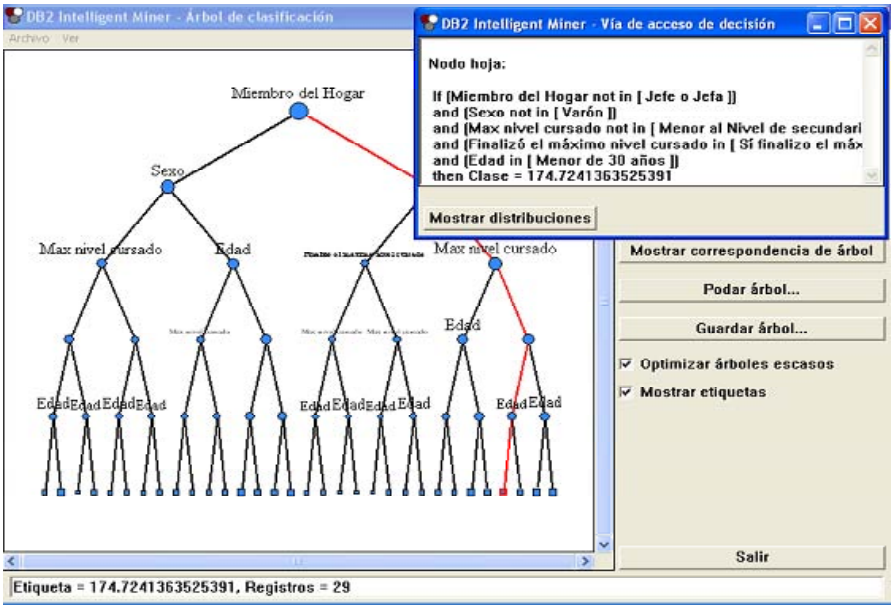

Figura 6.176: Visualizácion de la regla N°4 del árbol de decisión, como así tambíen el ingreso total individual que es de 174.72 y el numero de 29 que son los registros que cumplen con esas características.

La regla extraída del siguiente árbol de decisión como puede visualizarse en la fig. 6.176 de la pág. 253 cumple con las siguientes condiciones:

Si no es jefe/jefa del hogar donde habita, no es masculino, es menor a 30 años su máximo nivel cursado no es inferior al secundario y finalizó el máximo nivel cursado entonces el ingreso es de 174.72.

La regla extraída del siguiente árbol de decisión en este caso cumple con las siguientes condiciones (ver fig.6.177 de la pág. 254):

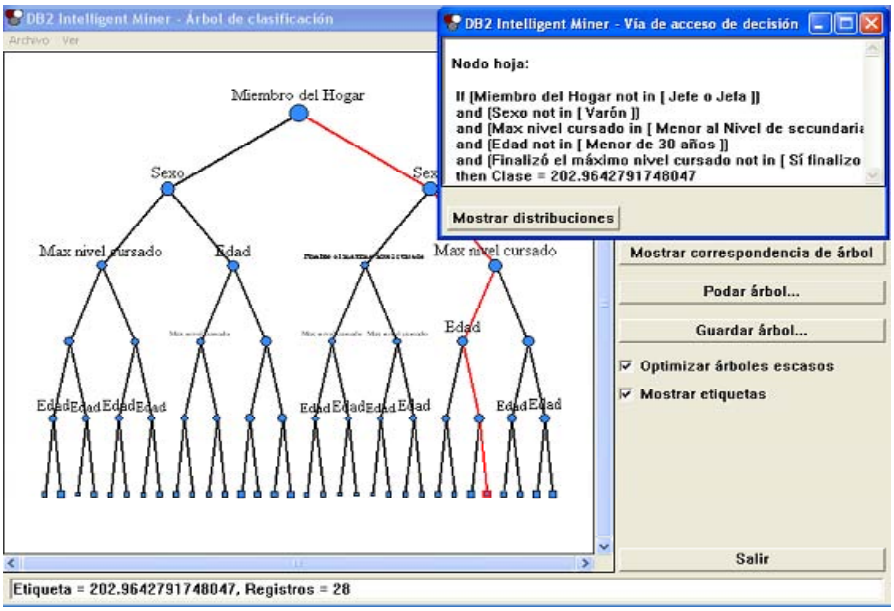

Figura 6.177: Visualizácion de la regla Nº5 del árbol de decisión, como así tambíen el ingreso total individual que es de 202.97 y el numero de 28 que son los registros que cumplen con esas características.

Si el individuo de estudio es de sexo femenino, no es la jefa del hogar donde habita, posee mas de 30 años edad su máximo nivel cursado es inferior al secundario y el mismo no lo a finalizado entonces el ingreso es de *202.96*.

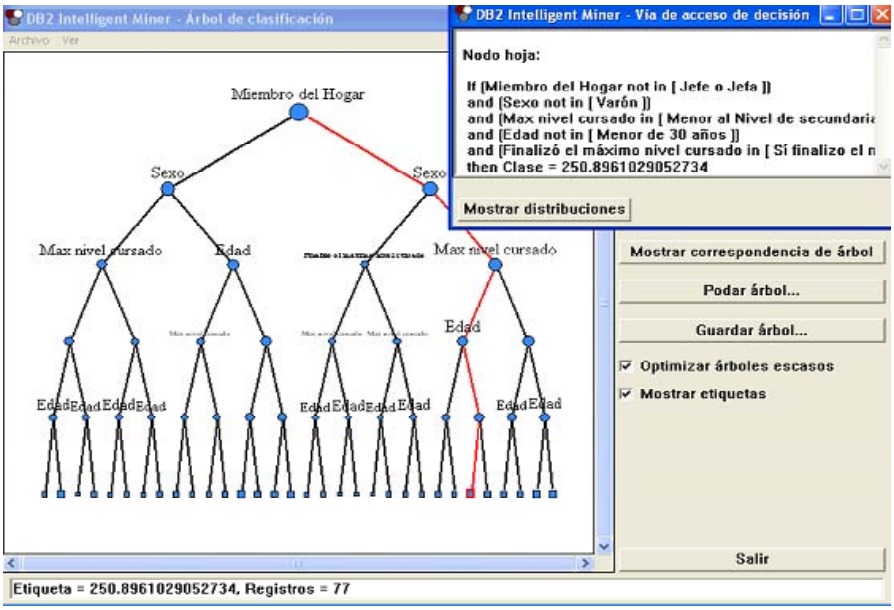

Figura 6.178: Visualizácion de la regla N°6 del árbol de decisión con su respectiva rama involucrada, como así también sus correspondiente etiqueta y numero de registros.

Así como se puede observar en la fig.6.178 de la pág. 255 la regla obtenida, también se visualiza al ingreso total individual, teniendo un valor de 250.89 con total de 77 de registros.

La regla cumple con las siguientes condiciones (ver fig.6.178 de la pág. 255):

Si el individuo de estudio es de sexo femenino, no es la jefa del hogar donde habita, posee menos de 30 años edad su máximo nivel cursado es inferior al secundario y el mismo lo a finalizado entonces el ingreso es de *250.89*.

Como se puede observar en la fig. 6.179 de la pág. 256 la regla obtenida es la siguiente:

Si el individuo de estudio es de sexo femenino, no es la jefa del hogar donde

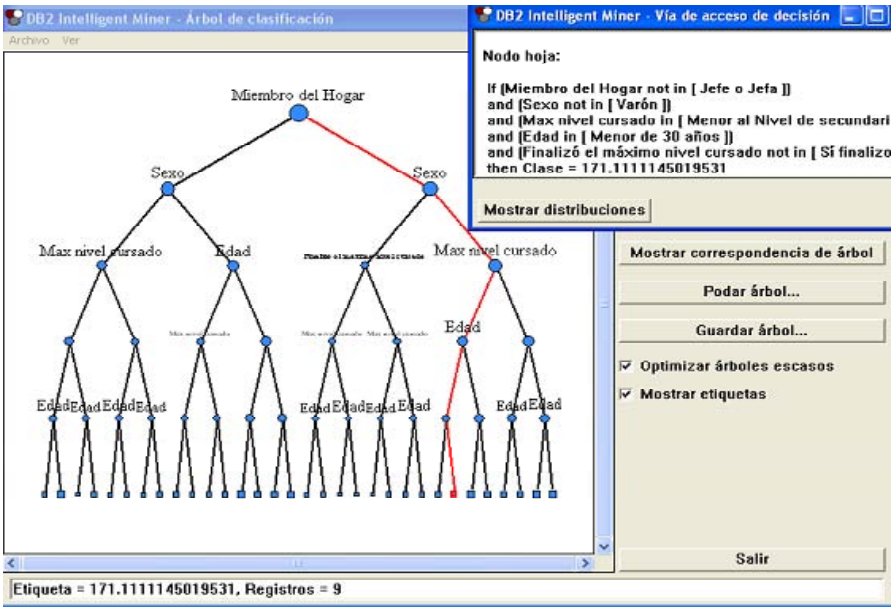

Figura 6.179: Visualizácion de la regla N^{o}7 del árbol de decisión con su respectiva rama involucrada, como así también sus correspondiente etiqueta y numero de registros.

habita, posee menos de 30 años edad su máximo nivel cursado es inferior al secundario y el mismo no lo a finalizado entonces el ingreso es de *171.11*.

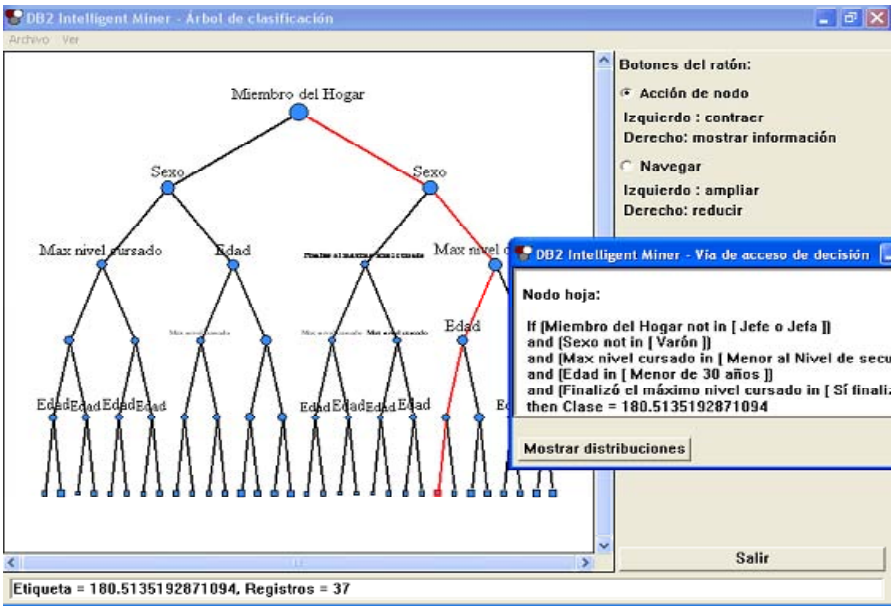

Figura 6.180: Visualizácion de la regla Nº8 del árbol de decisión con su respectiva rama involucrada, como así también sus correspondiente etiqueta y numero de registros.

En la fig.6.180 de la pág. 257 se puede visualizar así como la siguiente regla:

Si el individuo de estudio es de sexo femenino, no es la jefa del hogar donde habita, posee menos de 30 años edad, su máximo nivel cursado es inferior al secundario y el mismo lo a finalizado a el mismo, entonces el ingreso es de *180.51*.

También el valor del ingreso total individual que en este caso posee un valor de *180.51*, como el número de registros *37* que cumplen con las características de dichos atributos.

La regla extraída del árbol de decisión en este caso es la siguiente (ver la fig. 6.181 de la pág. 258):

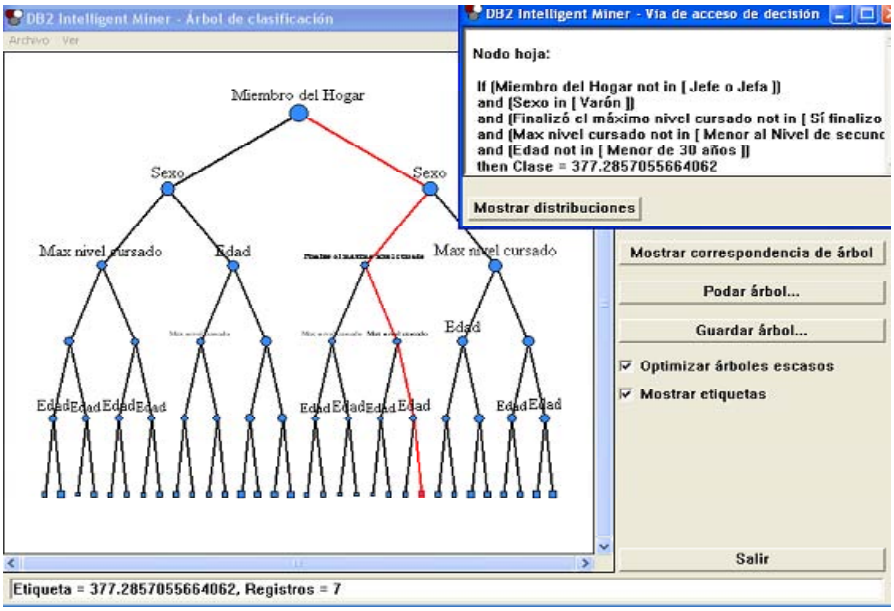

Figura 6.181: Visualizácion de la regla N°9 del árbol de decisión con su respectiva rama involucrada, como así también sus correspondiente etiqueta y numero de registros.

Si el individuo de estudio es de sexo masculino, no es el jefa del hogar donde habita, posee mas de 30 años edad, su máximo nivel cursado es inferior al secundario y al mismo no lo a finalizado, entonces el ingreso es de *377.28*.

A diferencia de las ramas y reglas vistas anteriormente, esta tiende en el nodo atributo sexo asía la izquierda lo que implicará que todas reglas obtenidas tendrán como resultado final a un individuo masculino en vez de femenino.

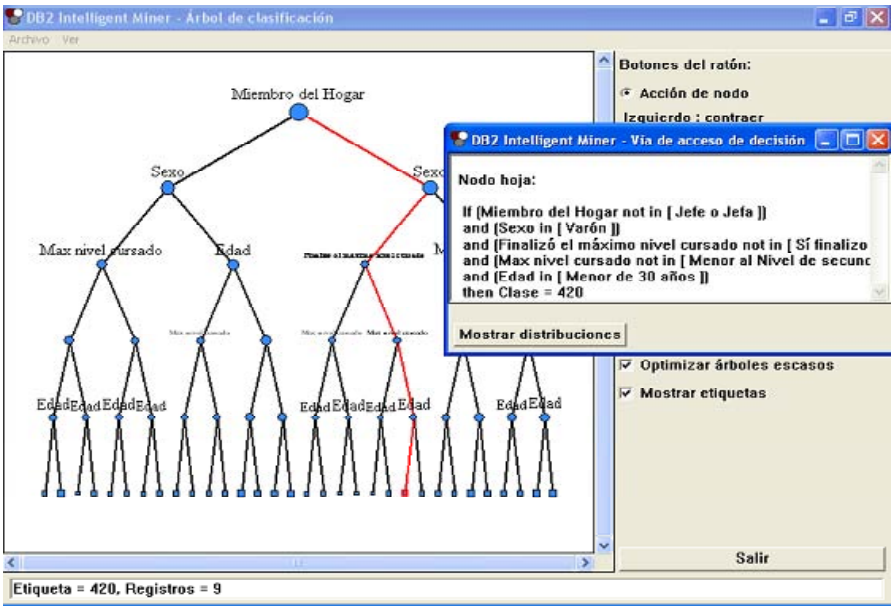

Figura 6.182: Visualizácion de la regla N^{o}10 del árbol de decisión con su respectiva rama involucrada, como así también sus correspondiente etiqueta y numero de registros.

Se puede observar en la fig.6.182 de la pág. 259 la regla obtenida, como así también al ingreso total individual, teniendo un valor de *420* con total de *9* de registros involucrados.

Dicha regla cumple con las siguientes condiciones (ver fig.6.182 de la pág. 259):

Si el individuo de estudio es de sexo masculino, no es el jefe del hogar donde habita, posee menos de 30 años edad, su máximo nivel cursado es inferior al secundario y al mismo no lo a finalizado, entonces el ingreso es de *420*.

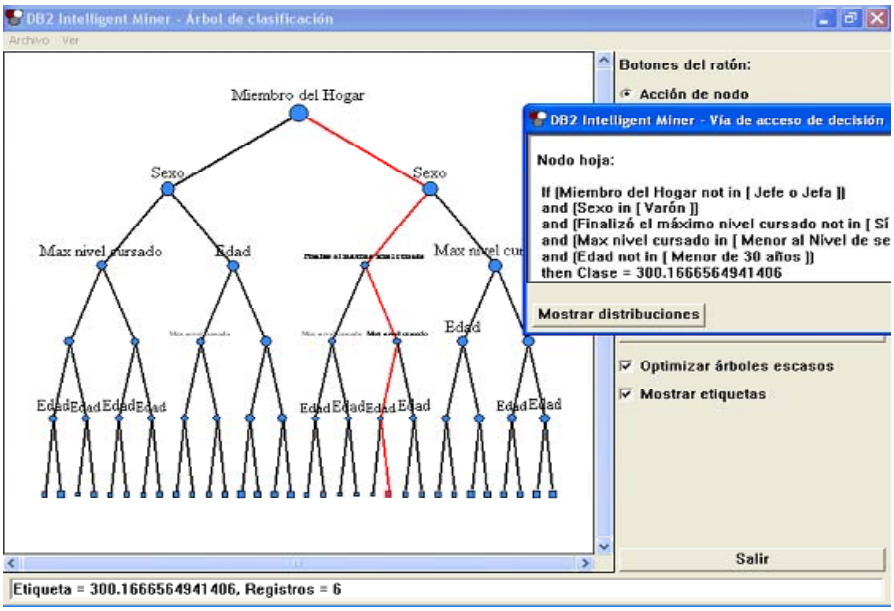

Figura 6.183: Visualizácion de la regla N°11 del árbol de decisión con su respectiva rama involucrada, como así también sus correspondiente etiqueta y numero de registros.

En la fig.6.183 de la pág. 260 se puede visualizar así como la siguiente regla: Si el individuo de estudio es de sexo masculino, no es el jefe del hogar donde habita, no posee menos de 30 años edad, su máximo nivel cursado es inferior al secundario y al mismo no lo a finalizado, entonces el ingreso es de *300.16*.

También el valor de la etiqueta que es el ingreso total individual que en este caso posee un valor de *300.16*, como el número de registros *6* que cumplen con las características de dichos atributos.

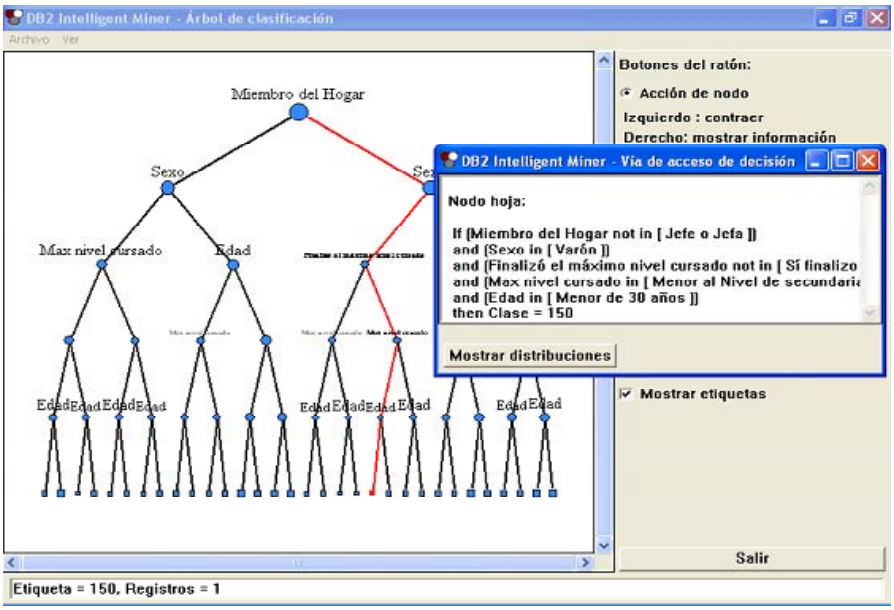

Figura 6.184: Visualizácion de la regla N^{o}12 del árbol de decisión con su respectiva rama involucrada, como así también sus correspondiente etiqueta y numero de registros.

Como se puede observar en la fig. 6.184 de la pág. 261 la regla obtenida es la siguiente:

Si el individuo de estudio es de sexo masculino, no es el jefe del hogar donde habita, posee menos de 30 años edad, su máximo nivel cursado es inferior al secundario y al mismo no lo a finalizado, entonces el ingreso es de *150.*

Con respecto al valor de la etiqueta y al numero de registros como puede visualizarse en la fig.6.184 de la pág. 261.

El valor del ingreso total individual que es de *150* siendo el menor significancia con respecto a los anteriores resultados, en cuanto al número de registros en este caso solamente hay *1* solo que cumple con estas características.

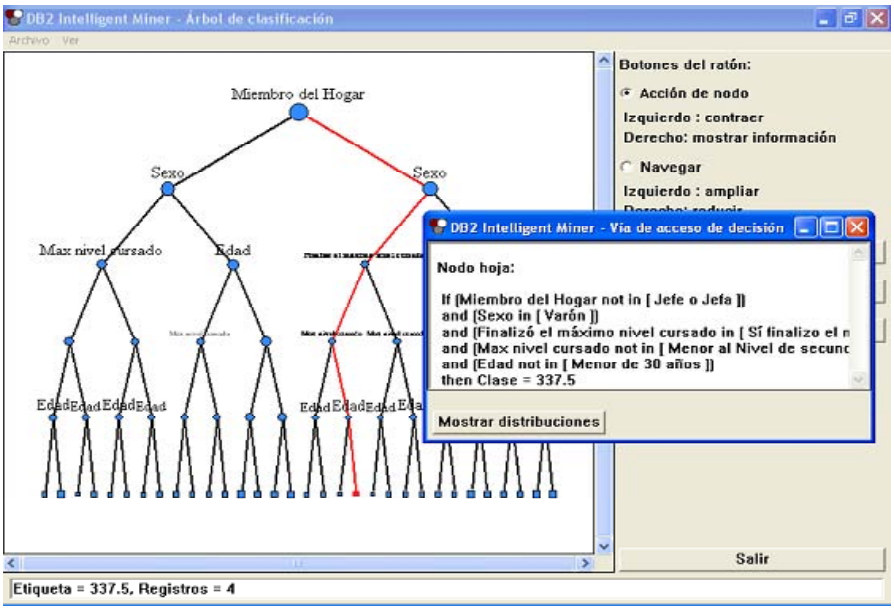

Figura 6.185: Visualizácion de la regla N°13 del árbol de decisión con su respectiva rama involucrada, como así también sus correspondiente etiqueta y numero de registros.

Como se puede observar en la fig. 6.185 de la pág. 262 la regla obtenida es la siguiente:

Si el individuo de estudio es de sexo masculino, no es el jefe del hogar donde habita, posee mas de 30 años edad, su máximo nivel cursado no es inferior al secundario y al mismo lo a finalizado, entonces el ingreso es de *337.5.*

Se puede observar en la fig.6.186 de la pág. 263 la regla obtenida, como así también al ingreso total individual, teniendo un valor de *367.5* con total de *2* de registros involucrados.

Dicha regla cumple con las siguientes condiciones (ver fig.6.186 de la pág. 263):

Si el individuo de estudio es de sexo masculino, no es el jefe del hogar donde

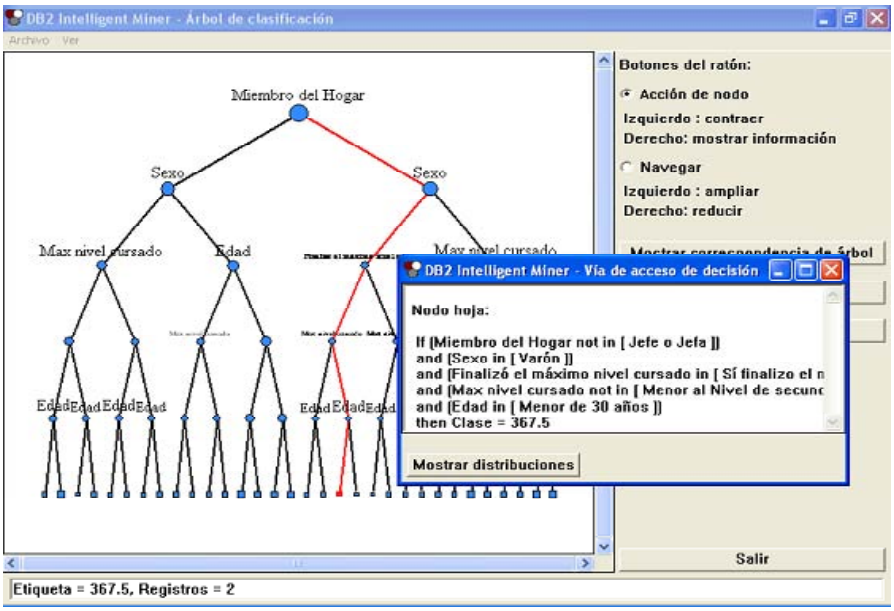

Figura 6.186: Visualizácion de la regla N°14 del árbol de decisión con su respectiva rama involucrada, como así también sus correspondiente etiqueta y numero de registros.

habita, posee menos de 30 años edad, su máximo nivel cursado no es inferior al secundario y al mismo lo a finalizado, entonces el ingreso es de *367.5.*

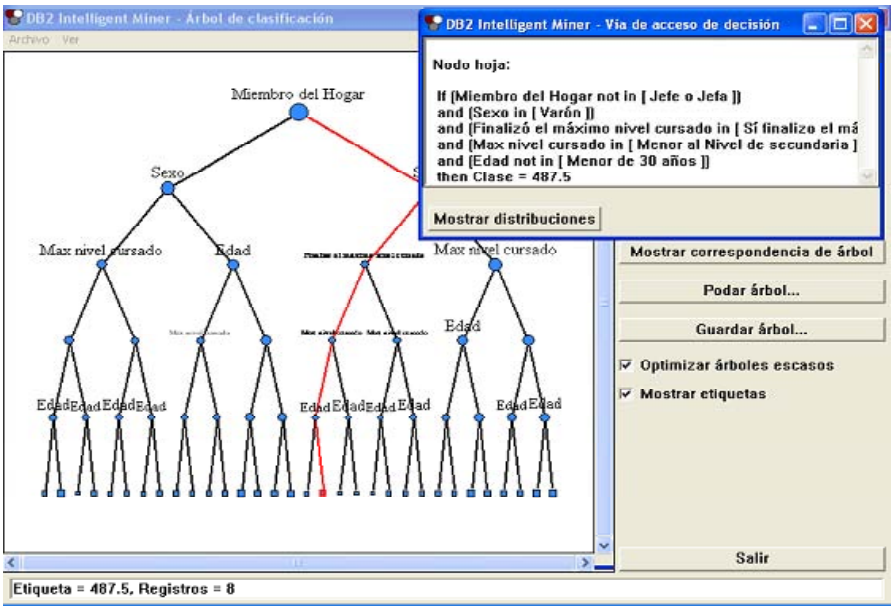

Figura 6.187: Visualizácion de la regla N°15 del árbol de decisión con su respectiva rama involucrada, como así también sus correspondiente etiqueta y numero de registros.

En la fig.6.187 de la pág. 264 se puede visualizar así como la siguiente regla:

Si el individuo de estudio es de sexo masculino, no es el jefe del hogar donde habita, posee mas de 30 años edad, su máximo nivel cursado es inferior al secundario y al mismo no lo a finalizado, entonces el ingreso es de *487.5.*

También el valor de la etiqueta que es el ingreso total individual que en este caso posee un valor de *487.5*, como el número de registros *8* que cumplen con las características de dichos atributos.

La regla extraída del árbol de decisión en este caso es la siguiente (ver la fig. 6.188 de la pág. 265):

Si el individuo de estudio es de sexo masculino, no es el jefe del hogar donde habita, el mismo posee menos de 30 años edad, su máximo nivel cursado es

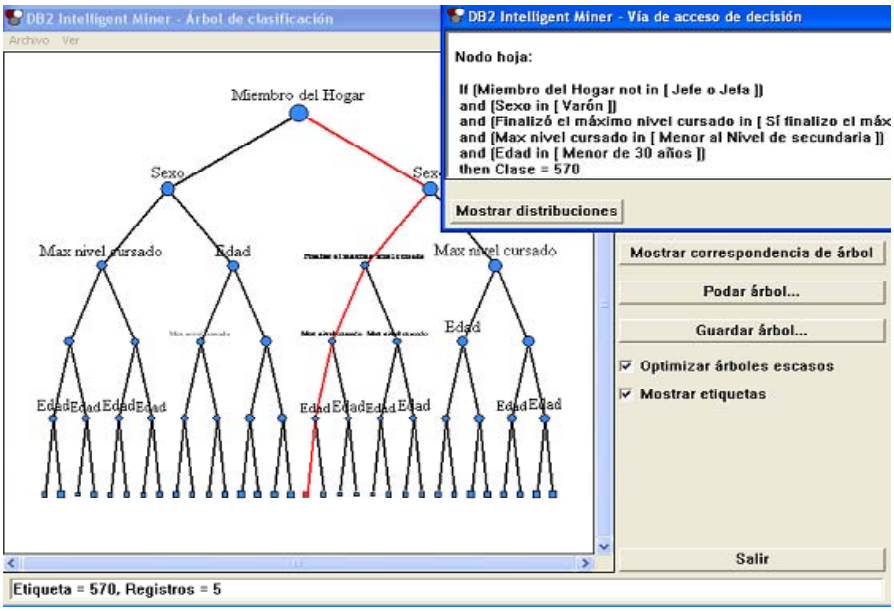

Figura 6.188: Visualizácion de la regla N^{o}16 del árbol de decisión con su respectiva rama involucrada, como así también sus correspondiente etiqueta y numero de registros.

inferior al secundario y al mismo lo a finalizado, entonces el ingreso es de *570*.

Con respecto al valor de la etiqueta y al numero de registros como puede visualizarse en la fig.6.188 de la pág. 265.

El valor del ingreso total individual que es de *150 siendo* el menor significancia con respecto a los anteriores resultados, en cuanto al número de registros en este caso solamente hay *1* solo que cumple con estas características.

Al observar en la fig. 6.188 de la pág. 265 se puede visualizar que esta ha sido la hoja con mayor ingreso total individual.

También fue la última rama con el atributo padre *"miembro del hogar = no es jefe/jefa"*.

Lo que implica que todas las reglas extraídas en las próximas ramas del árbol de decisión contendrán a los jefes o jefas del hogar en cuestión.

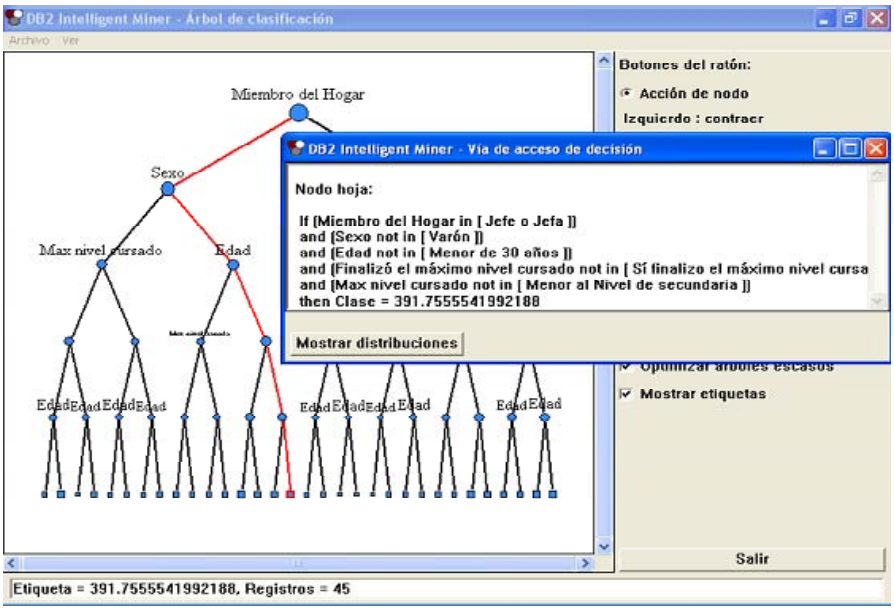

Figura 6.189: Visualizácion de la regla N°17 del árbol de decisión con su respectiva rama involucrada.

En la fig.6.189 de la pág. 266 se puede visualizar así como la siguiente regla:

Si el individuo de estudio es de sexo femenino, es el jefe o jefa del hogar donde habita, posee más de 30 años edad, su máximo nivel cursado no es inferior al secundario y al mismo no lo a finalizado, entonces el ingreso es de *391.75.*

También el valor de la etiqueta que es el ingreso total individual que en este caso posee un valor de *391.75*, como el número de registros *45* que cumplen con las características de dichos atributos.

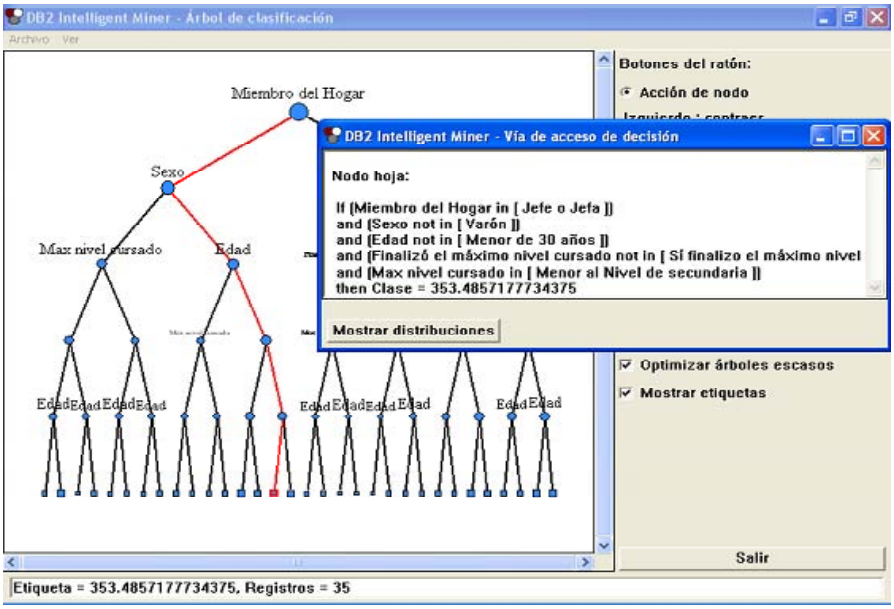

Figura 6.190: Visualizácion de la regla N^{o}18 del árbol de decisión con su respectiva rama involucrada.

Como se puede observar en la fig. 6.190 de la pág. 267 la regla obtenida es la siguiente:

Si el individuo de estudio es de sexo femenino, es el jefe o jefa del hogar donde habita, posee más de 30 años edad, su máximo nivel cursado es inferior al secundario y al mismo no lo a finalizado, entonces el ingreso es de *353.48.*

Con respecto al valor de la etiqueta esta posee un valor de *353.48,* con un total de *27* registros involucrados dicha regla (ver la fig.6.190 de la pág. 267).

Como se puede observar en la fig. 6.191 de la pág. 268 la regla obtenida

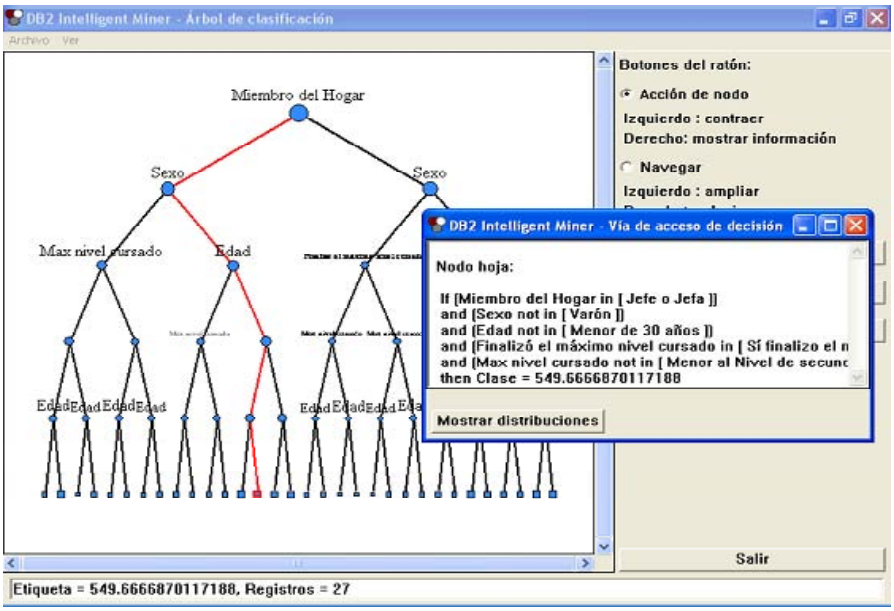

Figura 6.191: Visualizácion de la regla N°19 del árbol de decisión con su respectiva rama involucrada, como así también sus correspondiente etiqueta con un valor de *549.66* y con *27* registros involucrados.

es la siguiente:

Si el individuo de estudio es de sexo femenino, es el jefe o jefa del hogar donde habita, posee más de 30 años edad, su máximo nivel cursado no es inferior al secundario y al mismo lo a finalizado, entonces el ingreso es de *549.66.*

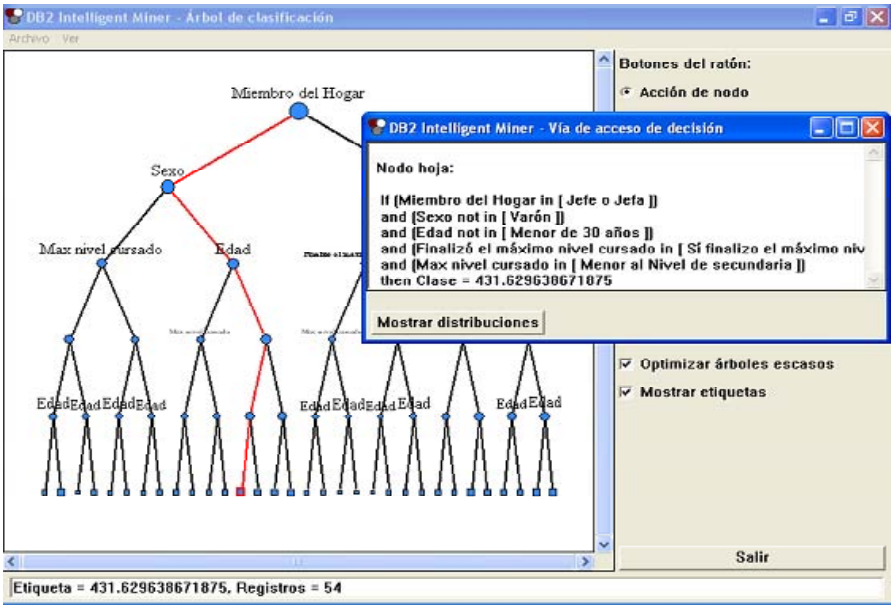

Figura 6.192: Visualizácion de la regla N^{o}20 del árbol de decisión con su respectiva rama involucrada.

En la fig.6.192 de la pág. 269 se puede visualizar así como la siguiente regla:

Si el individuo de estudio es de sexo femenino, es el jefe o jefa del hogar donde habita, posee más de 30 años edad, su máximo nivel cursado es inferior al secundario y al mismo lo a finalizado, entonces el ingreso es de *431.63.*

También el valor de la etiqueta que es el ingreso total individual que en este caso posee un valor de *431.63*, como el número de registros *54* que cumplen con las características de dichos atributos.

Se puede observar en la fig.6.193 de la pág. 270 la regla obtenida, como así también al ingreso total individual, teniendo un valor de *321.16* con total

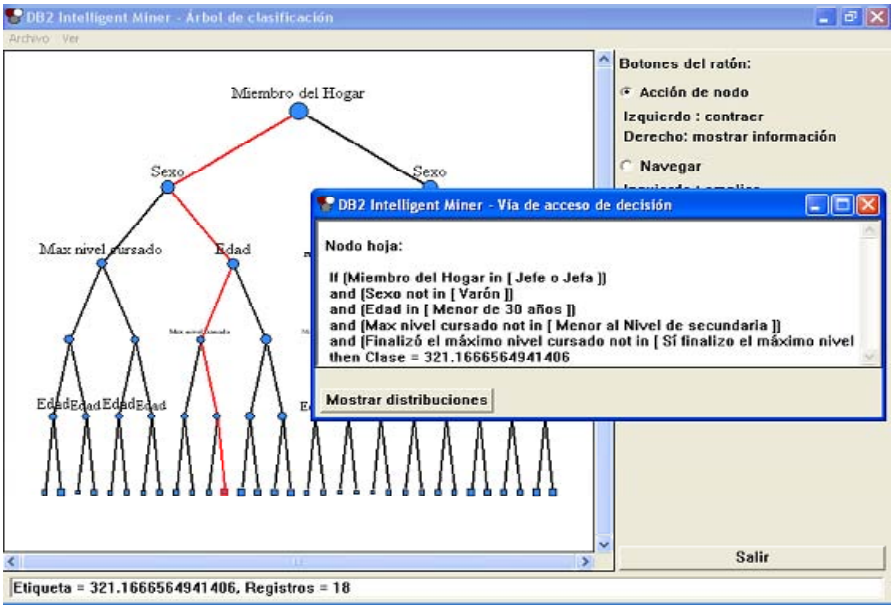

Figura 6.193: Visualizácion de la regla N°21 del árbol de decisión con su respectiva rama involucrada.

de *18* de registros involucrados.

Dicha regla cumple con las siguientes condiciones (ver fig.6.193 de la pág. 270):

Si el individuo de estudio es de sexo femenino, es el jefe o jefa del hogar donde habita, posee menos de 30 años edad, su máximo nivel cursado no es inferior al secundario y al mismo no lo a finalizado, entonces el ingreso es de *321.16.*

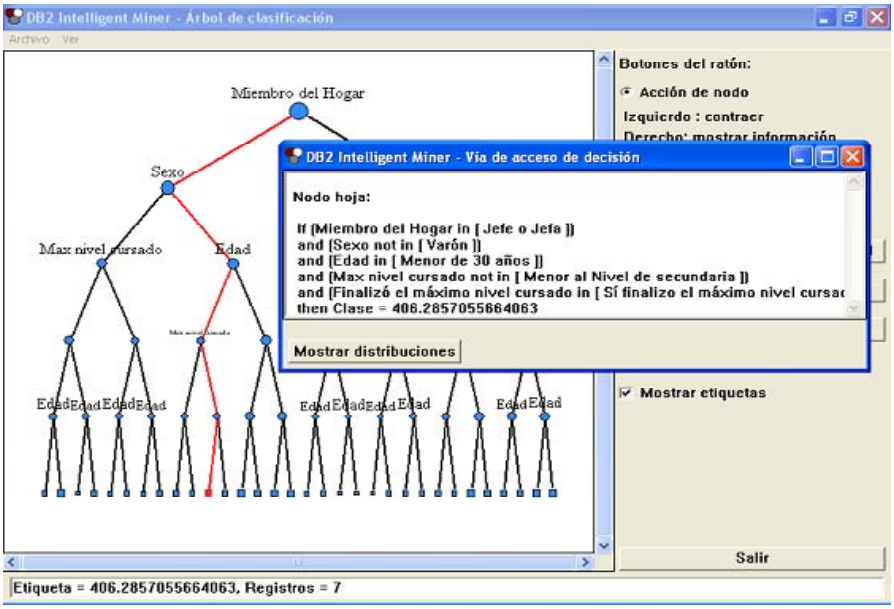

Figura 6.194: Visualizácion de la regla N^{o}22 del árbol de decisión con su respectiva rama involucrada.

En la fig.6.194 de la pág. 271 se puede visualizar así como la siguiente regla:

Si el individuo de estudio es de sexo femenino, es el jefe o jefa del hogar donde habita, posee menos de 30 años edad, su máximo nivel cursado no es inferior al secundario y al mismo lo a finalizado, entonces el ingreso es de *406.28.*

También el valor de la etiqueta que es el ingreso total individual que en este caso posee un valor de *406.28*, como el número de registros *7* que cumplen

con las características de dichos atributos.

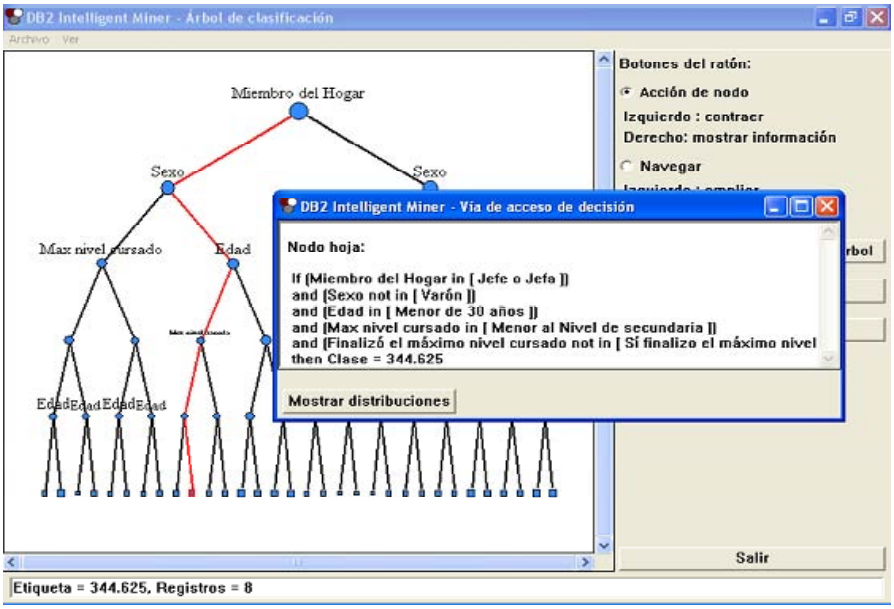

Figura 6.195: Visualizácion de la regla N°23 del árbol de decisión con su respectiva rama involucrada.

En la fig.6.195 de la pág. 272 se puede visualizar así como la siguiente regla:

Si el individuo de estudio es de sexo femenino, es el jefe o jefa del hogar donde habita, posee menos de 30 años edad, su máximo nivel cursado es inferior al secundario y al mismo no lo a finalizado, entonces el ingreso es de *344.62.*

También el valor de la etiqueta que es el ingreso total individual que en este caso posee un valor de *344.62*, como el número de registros *8* que cumplen con las características de dichos atributos.

Como se puede observar en la fig. 6.196 de la pág. 273 la regla obtenida es la siguiente:

Si el individuo de estudio es de sexo femenino, es el jefe o jefa del hogar donde habita, posee menos de 30 años edad, su máximo nivel cursado es inferior al secundario y al mismo lo a finalizado, entonces el ingreso es de *193.5.*

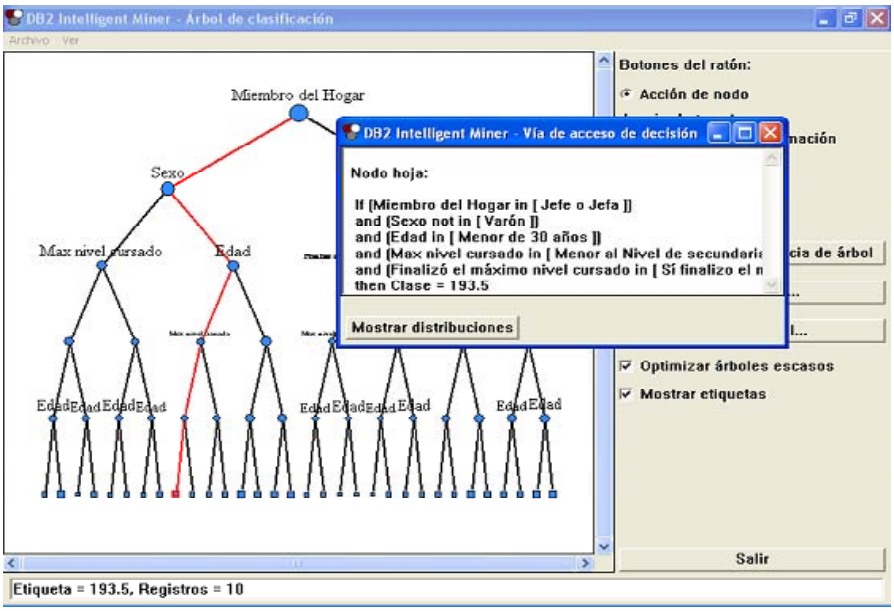

Figura 6.196: Visualizácion de la regla N^{o}24 del árbol de decisión con su respectiva rama involucrada, como así tambien los correspondientes valores del ingreso total individual y el numero de registro.

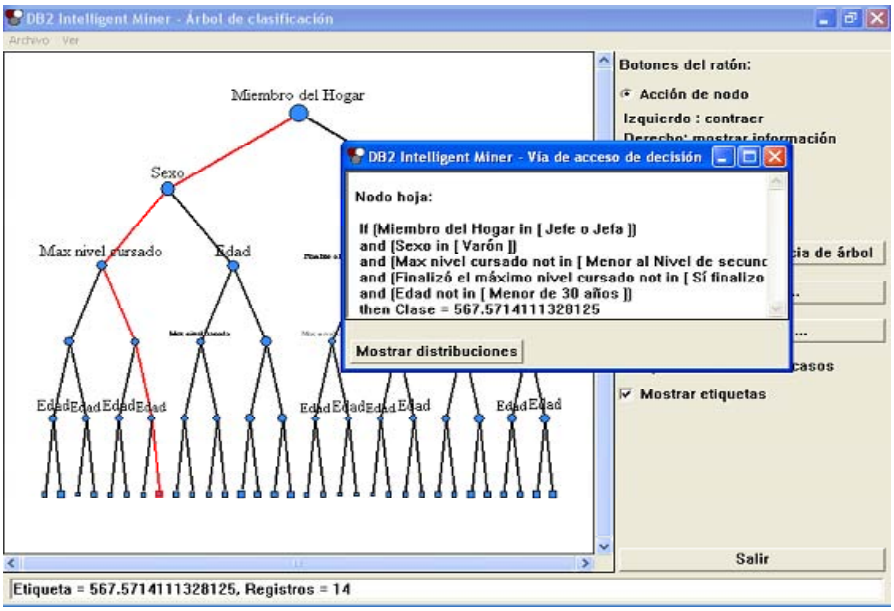

Figura 6.197: Visualizácion de la regla N°25 del árbol de decisión con su respectiva rama involucrada, como así tambien los correspondientes valores del ingreso total individual y el numero de registro.

En la fig.6.197 de la pág. 274 se puede visualizar así como la siguiente regla:

Si el individuo de estudio es de sexo masculino, es el jefe o jefa del hogar donde habita, posee mas de 30 años edad, su máximo nivel cursado no es inferior al secundario y al mismo no lo a finalizado, entonces el ingreso es de *567.57*.

También el valor de la etiqueta que es el ingreso total individual que en este caso posee un valor de *567.57*, como el número de registros *14* que cumplen con las características de dichos atributos.

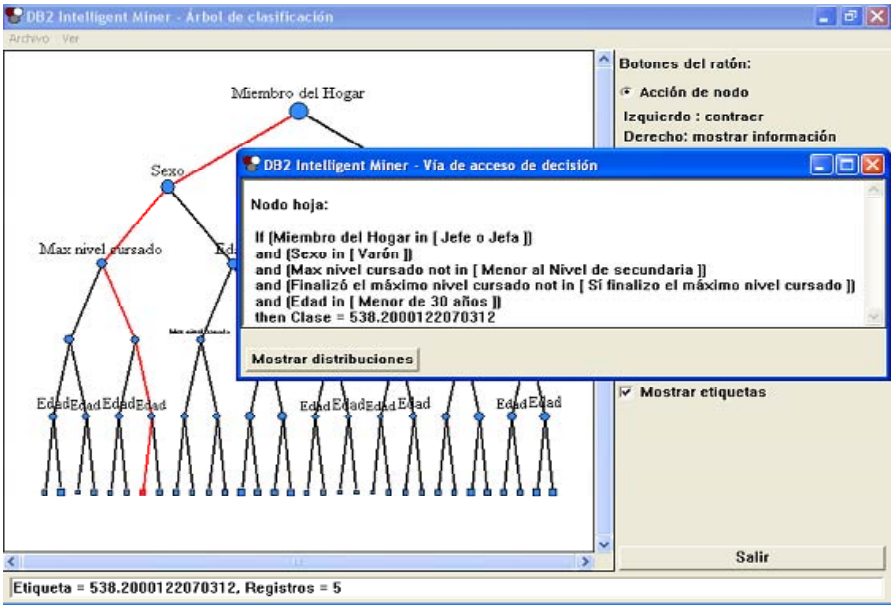

Figura 6.198: Visualizácion de la regla N^{o}26 del árbol de decisión con su respectiva rama involucrada, como así tambien los correspondientes valores del ingreso total individual y el numero de registro.

Como se puede observar en la fig. 6.198 de la pág. 275 la regla obtenida es la siguiente:

Si el individuo de estudio es de sexo masculino, es el jefe o jefa del hogar donde habita, posee menos de 30 años edad, su máximo nivel cursado no es inferior al secundario y al mismo no lo a finalizado, entonces el ingreso es de *538.20*.

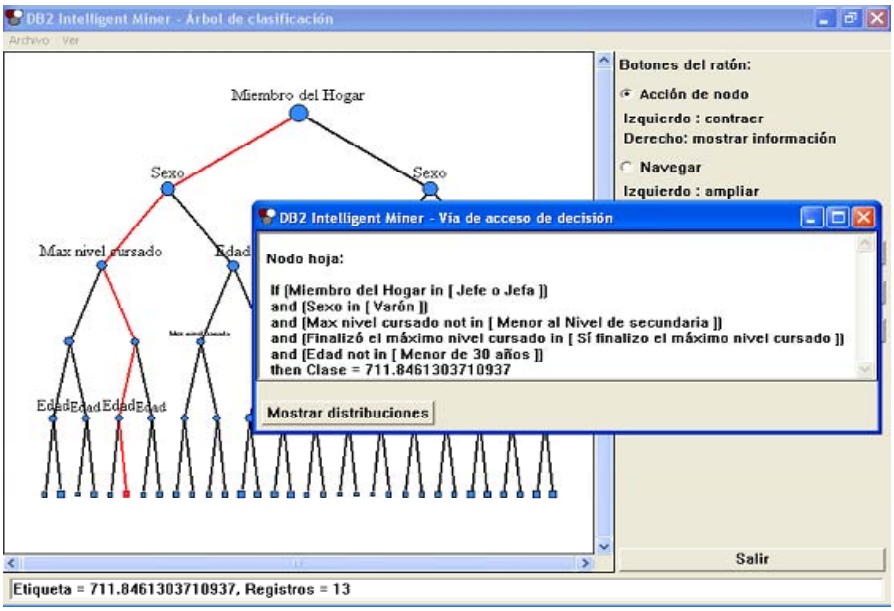

Figura 6.199: Visualizácion de la regla N°27 del árbol de decisión con su respectiva rama involucrada, como así tambien los correspondientes valores del ingreso total individual y el numero de registro.

En la fig.6.199 de la pág. 276 se puede visualizar así como la siguiente regla:

Si el individuo de estudio es de sexo masculino, es el jefe o jefa del hogar donde habita, posee más de 30 años edad, su máximo nivel cursado no es inferior al secundario y al mismo lo a finalizado, entonces el ingreso es de *711.84*.

También el valor de la etiqueta que es el ingreso total individual que en este caso posee un valor de *711.84* siendo este el de mayor valor de todas las reglas antes vistas, como el número de registros *13* que cumplen con las características de dichos atributos.

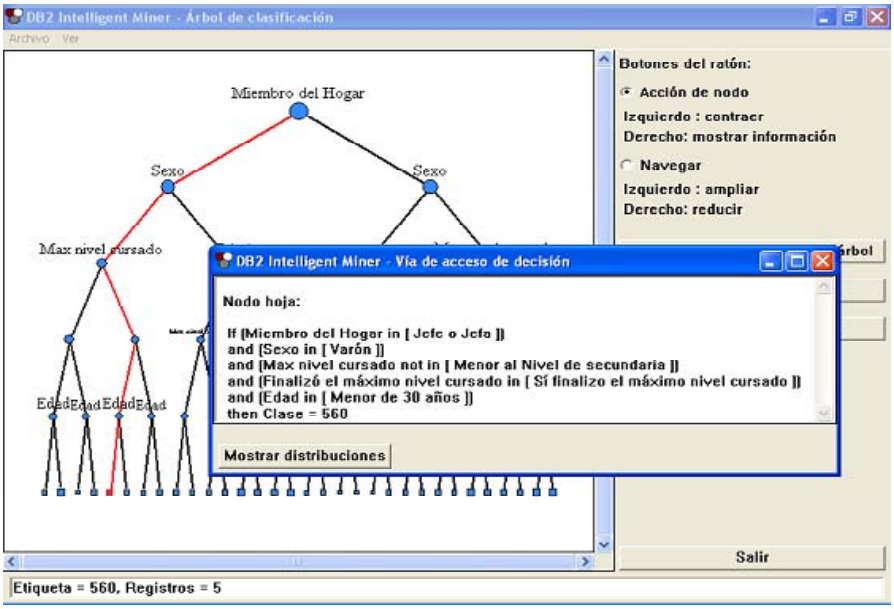

Figura 6.200: Visualizácion de la regla N°28 del árbol de decisión con su respectiva rama involucrada.

Como se puede observar en la fig. 6.200 de la pág. 277 la regla obtenida es la siguiente:

Si el individuo de estudio es de sexo masculino, es el jefe o jefa del hogar donde habita, posee menos de 30 años edad, su máximo nivel cursado no es inferior al secundario y al mismo lo a finalizado, entonces el ingreso es de *560*.

Se puede también visualizar en la fig.6.200 de la pág. 277 el ingreso total individual, teniendo un valor de *560* con total de *5* de registros involucrados.

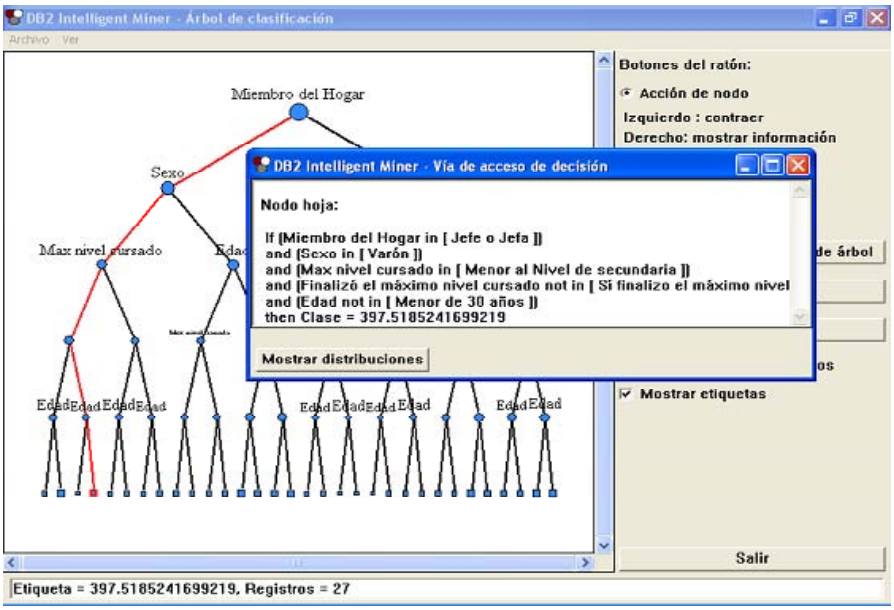

Figura 6.201: Visualizácion de la regla N°29 del árbol de decisión con su respectiva rama involucrada.

La regla extraída del árbol de decisión en este caso es la siguiente (ver la fig. 6.201 de la pág. 278):

Si el individuo de estudio es de sexo masculino, es el jefe o jefa del hogar donde habita, posee más de 30 años edad, su máximo nivel cursado es inferior al secundario y al mismo no lo a finalizado, entonces el ingreso es de *397.52.*

Como se puede apreciar en la fig. 6.201 de la pág. 278, el valor que la etiqueta posee es el de *397.52* siendo este el ingreso total individual, con un total de *27* registros involucrados en dicha relación.

En la fig.6.202 de la pág. 279 se puede visualizar así como la siguiente regla:

Si el individuo de estudio es de sexo masculino, es el jefe o jefa del hogar donde habita, posee menos de 30 años edad, su máximo nivel cursado es inferior al secundario y al mismo no lo a finalizado, entonces el ingreso es de *672.75.*

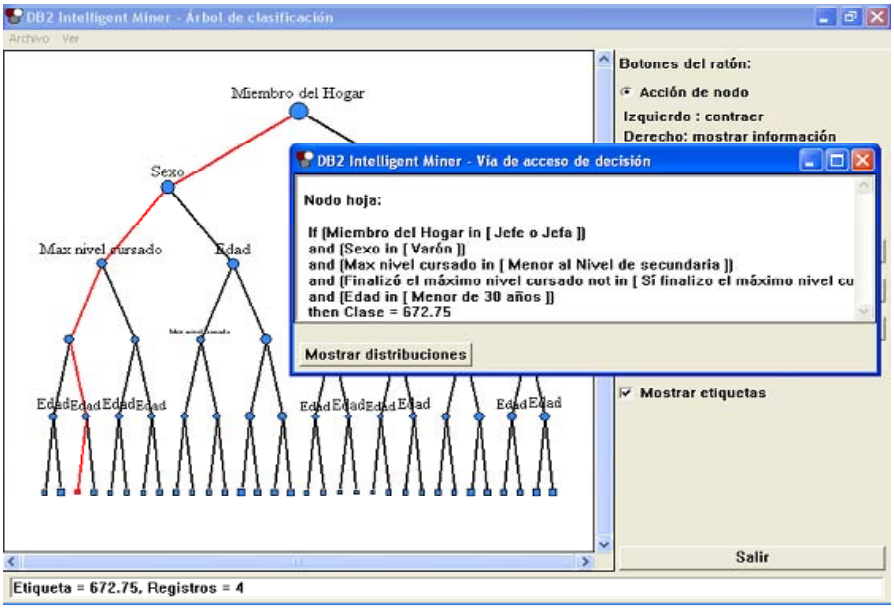

Figura 6.202: Visualizácion de la regla N^{o}30 del árbol de decisión con su respectiva rama involucrada.

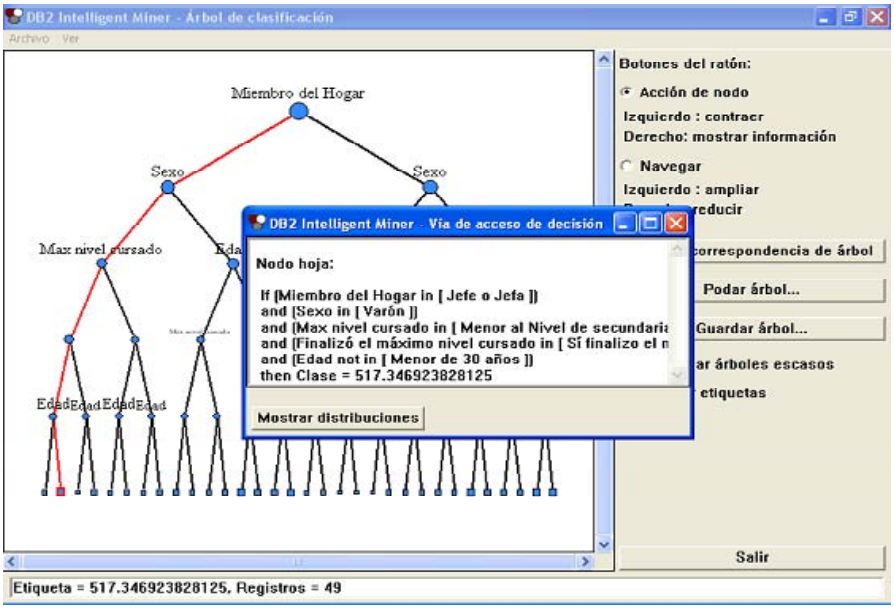

Figura 6.203: Visualizácion de la regla N°31 del árbol de decisión con su respectiva rama involucrada.

En la fig.6.203 de la pág. 280 se puede visualizar así como la siguiente regla:

Si el individuo de estudio es de sexo masculino, es el jefe o jefa del hogar donde habita, posee menos de 30 años edad, su máximo nivel cursado es inferior al secundario y al mismo lo a finalizado, entonces el ingreso es de *517.34*.

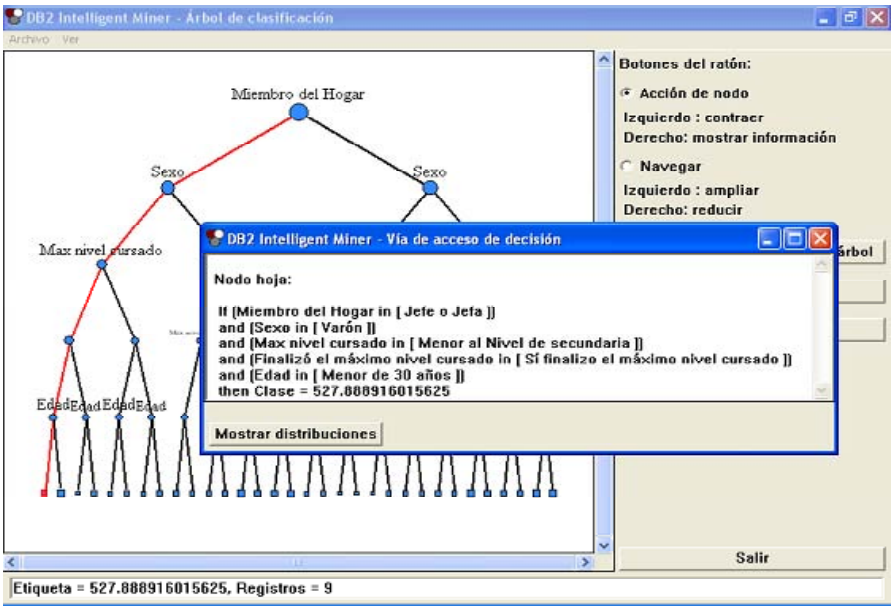

Figura 6.204: Visualizácion de la regla N^{o}32 del árbol de decisión con su respectiva rama involucrada.

Y para finalizar, se tiene a la fig. 6.204 de la pág. 281 que es la última regla extraída de árbol de decisión.

Si el individuo de estudio es de sexo masculino, es el jefe o jefa del hogar donde habita, posee menos de 30 años edad, su máximo nivel cursado es inferior al secundario y al mismo lo a finalizado, entonces el ingreso es de *527.88*.

Se puede también visualizar en la fig.6.200 de la pág. 277 el ingreso total individual, teniendo un valor de *527.88* con total de *9* de registros involucrados.

Capítulo 7

Extracción de Conocimiento con Pentaho Business Intelligence

7.1 Concepto de Inteligencia de Negocios Business Intelligence

La *Inteligencia de Negocios o Busines Intelligence* (*BI*) hace referencia a un conjunto de productos y servicios para acceder a los datos, analizarlos y convertirlos en información (ver fig. 7.1 de la pág. 284).

E-Business es la compleja fusión de los procesos de negocios, aplicaciones empresariales y estructura organizacional necesaria para crear un modelo de negocios altamente competitivo (Kalakota y Robinson).

La inteligencia en el negocio electrónico (*e-business)*, incluye actividades como el procesamiento analítico en línea (*OLAP*) y aprovechamiento de datos, también llamada extracción de datos o *Minería de Datos.*

Para obtener más información acerca de *Inteligencia de Negocios o Busines Intelligence* ver el Capítulo Nº1 *(Introducción a la Minería de Datos).*

Figura 7.1: En la BI se requiere aplicar herramientas de software a los datos que se encuentran en enormes almacenes para descubrir patrones significativos y tomar decisiones de negocios adecuadas.

7.2 Pentaho Business Intelligence (BI)

Pentaho Business Intelligence (BI) es una iniciativa en curso por la comunidad de *Open Source* que provee organizaciones con mejores soluciones para las necesidades de *Business Intelligence (BI)* a las empresa (ver fig. 7.2 de la pág. 284).

Figura 7.2: La corporacion *Pentaho* es el patrocinador primario y propietario del projecto *Pentaho Business Intelligence BI.*

La plataforma *Open Source Pentaho Business Intelligence* cubre amplias necesidades de análisis de los datos y de los informes empresariales.

Las soluciones de *Pentaho* están desarrolladas en *Java* y tienen un ambiente de implementación también basado en *Java.* Eso hace que *Pentaho* es una solución muy flexible para cubrir una amplia gama de necesidades empresariales tanto las típicas como las sofisticadas y específicas del negocio (ver fig. 7.3 de la pág. 285).

Las soluciones que *Pentaho* pretende ofrecer se componen fundamental-

mente de una infraestructura de herramientas de análisis e informes integrados con un motor de *workflow* de procesos de negocio.

La plataforma será capaz de ejecutar las reglas de negocio necesarias, expresadas en forma de procesos y actividades y de presentar y entregar la información adecuada en el momento adecuado, mediante analisis *OLAP, Cuadros de Mando, etc.*

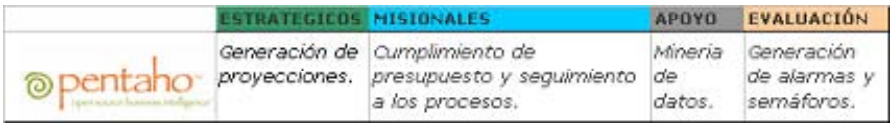

Figura 7.3: El siguiente esquema nos permite visualizar las distintas herramientas que la plataforma *Pentaho* utiliza para cada técnica de Business Intelligence.

7.2.1 Arquitectura de Pentaho

La solución *Business Intelligence OpenSource Pentaho* pretende ser una alternativa a las soluciones propietarias tradicionales más completas: *Business Objects, Cognos, Microstrategy, Microsoft, IBM, etc.*, por lo que incluye todos aquellos componentes que se pueden encontrar en las soluciones *Business Intelligence (BI)* propietarias más avanzadas:

- *Reporting.*
- *Análisis.*
- *Dashboards.*
- *Workflow.*
- *Data Mining.*
- *ETL.*
- *Single Sign-On. Ldap.*
- *Auditoría de uso y rendimiento.*
- *Planificador.*

- *Notificador.*
- *Seguridad. Perfiles.*

La fig. 7.4 de la pág. 286 permite visualizar la arquitectura estructurada de las diferentes componentes que forman parte de *Pentaho*.

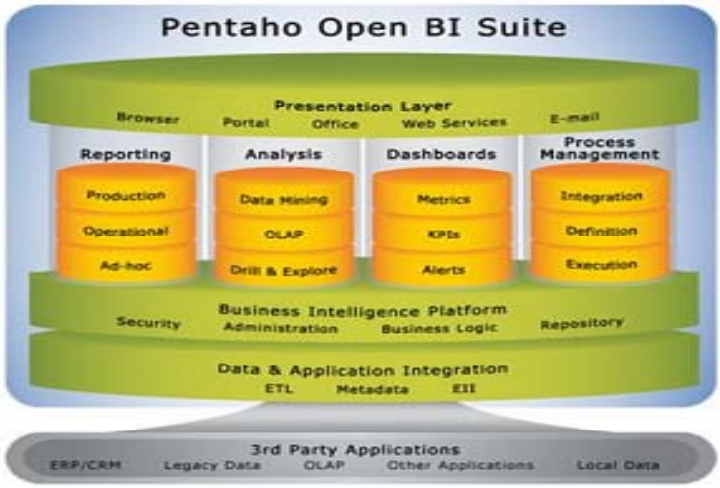

Figura 7.4: Visualización de la arquitectura de *Business Intelligence Open-Source Pentaho*.

7.2.2 Componentes del Pentaho

Business Intelligence Pentaho es una solución realizada en *Java* de código abierto flexible y muy potente que cubre prácticamente todas las necesidadesde una empresa.

Como la misma fue creada con el 100% *J2EE*, asegurando de esta forma la escalabilidad, integración y portabilidad.

Componentes Soportados

Servidor: *Pentaho* puede correr en servidores compatibles con *J2EE* como *JBOSS AS*, *IBM WebSphere, Tomcat, WebLogic y Oracle AS.*

Base de datos: Vía JDBC, *IBM DB2, Microsft SQL Server, MySQL, Oracle, PostgreSQL, NCR Teradata, Firebird.*

Sistema operativo: No existe dependencia; lenguaje interpretado.

Lenguaje de programación: *Java, Javascript, JSP, XSL (XSLT / XPath / XSL-FO).*

Interfaz de desarrollo: *Java SWT, Eclipse, Web-based.*

Todos los componentes están expuestos vía *Web Services* para facilitar la integración con *Arquitecturas Orientadas a Servicios (SOA)* .

También todos los repositorios de datos del *Business Intelligence Pentaho* están basados en *XML.*

La fig. 7.5 de la pág. 287 visualiza la interacción entre los diferentes componentes de *Pentaho.*

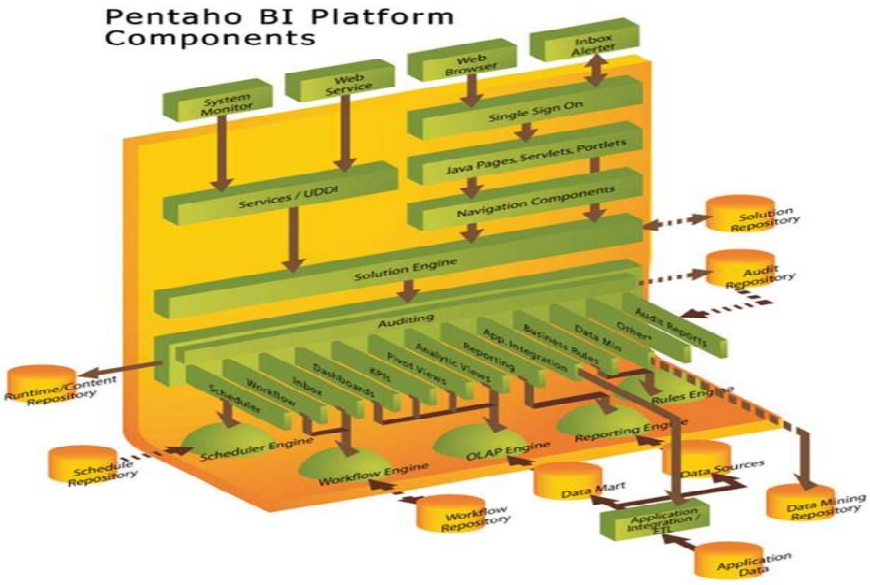

Figura 7.5: Visualización de los diferentes componentes soportados por *Pentaho Business Intelligence.*

Figura 7.6: *Pentaho Reporting* es un potente generador de informes: Permite la distribución de los resultados del análisis en múltiples formatos.

7.2.3 Características de Pentaho

Pentaho Business Intelligence abarca las siguientes aéreas de reporte:

Pentaho Reporting

La solución proporcionada por la plataforma *Business Intelligence OpenSource Pentaho* e integrada en su suite para el desarrollo de informes se llama *Pentaho Reporting* (ver fig. 7.6 de la pág. 288).

Existen tres productos con diferentes enfoques y dirigidos a diferentes tipos de usuarios:

- *Pentaho Report Designer*

Es un editor basado en *Eclipse* con prestaciones profesionales con capacidad de personalización de informes a las necesidades de los negocios destinado a desarrolladores.

Esta herramienta está estructurada de forma que los desarrolladores puedan acceder a sus prestaciones de forma rápida.

Incluye un editor de consultas para facilitar la confección de los datos que serán utilizados en un informe.

- *Pentaho Report Design Wizard*

Es una herramienta de diseño de informes, que facilita el trabajo y permite a los usuarios obtener resultados de forma inmediata. Está destinada a

usuarios con menos conocimientos técnicos.

- *Web ad-hoc reporting*

Es el similar a la herramienta *Pentaho Report Design Wizard*, pero via web.

Esta herramienta extiende la capacidad de los usuarios finales para la creación de informes a partir de plantillas preconfiguradas y siguiendo un asistente de creación.

La fig. 7.7 de la pág. 289 permite visualizar los distintos tipos de reportes desarrollados con cualquiera de las harramientas de *Pentaho Reporting.*

Figura 7.7: Visualización de los distintos reportes generados por *Pentaho Reporting.*

Pentaho Análisis

Ayuda a operar con máxima efectividad para ganar perspicacia y entender lo necesario para tomar optimas decisiones.

Las características generales son:

- Vista dimensional de datos (por ventas, por período, por empleados, etc.).
- Navegar y explorar (*Análisis Ad Hoc, Drill-down, etc.*).
- Interactuar con alto rendimiento mediante tecnologías optimizadas para la rápida respuesta interactiva.

La fig. 7.8 de la pág. 290 y la fig. 7.9 de la pág. 291 nos permite visualizar las distintas formas de análisis e interpretación de los datos que posee el *Pentaho Análisis*.

		Region			
Positions	Measures	Central	Eastern	Southern	Western
CEO	Actual	$549,625.00	$500,000.00	$500,000.00	$500,000.00
	Budget	$522,250.00	$488,750.00	$498,750.00	$478,750.00
	Variance Percent	-5.24%	-2.30%	-.25%	-4.44%
HR Generalists	Actual	$856,190.00	$961,000.00	$961,000.00	$961,000.00
	Budget	$771,225.00	$940,158.00	$940,158.00	$938,158.00
	Variance Percent	-11.02%	-2.22%	-2.22%	-2.43%
HR Training	Actual	$397,473.00	$271,200.00	$271,200.00	$271,200.00
	Budget	$443,570.00	$279,674.00	$279,674.00	$277,674.00
	Variance Percent	10.39%	3.03%	3.03%	2.33%

Figura 7.8: Visualización de los diferentes paneles de analisis con el *Pentaho Análisis.*

Pentaho Dashboards

Esta solución provee inmediata perspicacia en un rendimiento individual, departamental o empresarial. Pentaho Dashboards facilita a los usuarios de los

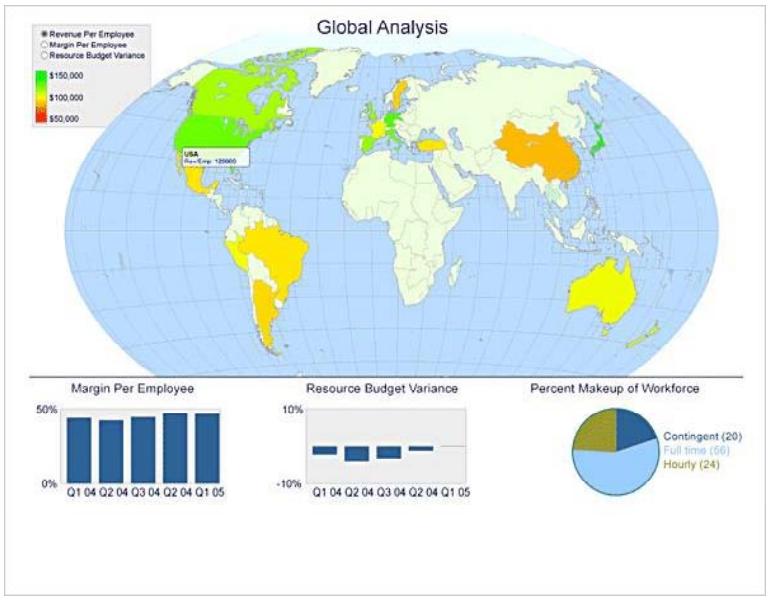

Figura 7.9: *Pentaho Análisis* permitira a el ususario final realizar diferentes analisis de las variables o de los campos de la bases de datos de estudio.

negocios información crítica que necesitan para entender y mejorar el rendimiento organizacional.

El *Pentaho Dashboards* es una potente herramienta que cuenta con las siguientes características:

- Identificación de métricas clave (*KPIs, Key Performance Indicators*), mediante la generación de Monitoreo/Métricas.
- Realización de investigaciones de detalles subyacentes, con reportes de soportes.
- Ejecución de seguimientos de excepciones, permitiendo pre-establecer alertas basadas en reglas del negocio.

Como se puede apreciar en la fig. 7.10 de la pág. 292, se ovservan todas las características antes mencionadas.

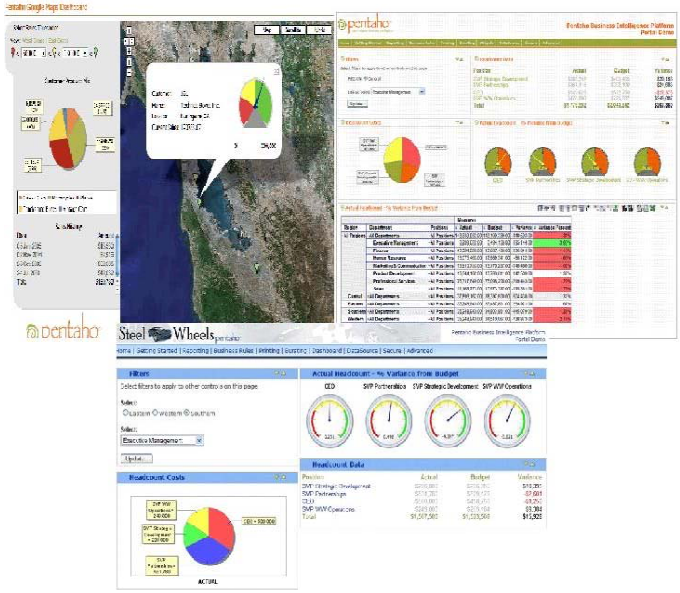

Figura 7.10: El *Pentaho Dashboards* es una potente herramienta que permite la incorporación de múltiples tipos de gráficos, tablas y velocímetros a un determinado proyecto de Business Intelligence.

Pentaho Data Integration

Los datos que alimentan a un sistema *data warehouse (DW)* proviene de diferentes fuentes, estas fuentes son los distintos sistemas operacionales que la empresa posee, generalmente ni son homogéneos entre sí ni concuerdan exactamen con lo que se necesita, por lo que será necesario realizar todas las adaptaciones pertinentes.

También muchas organizaciones tienen información disponible en aplicaciones y base de datos separadas.

Pentaho Data Integration abre, limpia e integra esta valiosa información y la pone en manos del usuario. Provee una consistencia, una sola versión de todos los recursos de información, que es uno de los más grandes desafíos para las organizaciones TI hoy en día.

Pentaho Data Integration permite una poderosa *ETL (Extract, Transform, Load) Extracción, Transformación y Carga*.

El uso de la solución *Kettle* permite evitar grandes cargas de trabajo manual frecuentemente difícil de mantener y de desplegar.

La arquitectura de *Pentaho Data Integration* viene representada por el esquema de la fig. 7.11 de la pág. 294.

Data Mining

La plataforma *Business Intelligence OpenSource Pentaho* ofrece diferentes soluciones para el desarrollo de un proyecto de *Business Intelligence*.

En este caso se hará referencia a la solución integrada al paquete *Business Intelligence Pentaho* para el desarrollo de proyectos de *Data Mining*.

El *Weka* (*Waikato Enviroment for Knowledge Analysis*) es un conjunto de librerías *JAVA* para la extracción de conocimientos desde bases de datos (ver fig. 7.12 de la pág. 295).

Es un software que ha sido desarrollado bajo licencia *GPL* lo cual ha impulsado que sea una de las suites más utilizadas en el área en los últimos años.

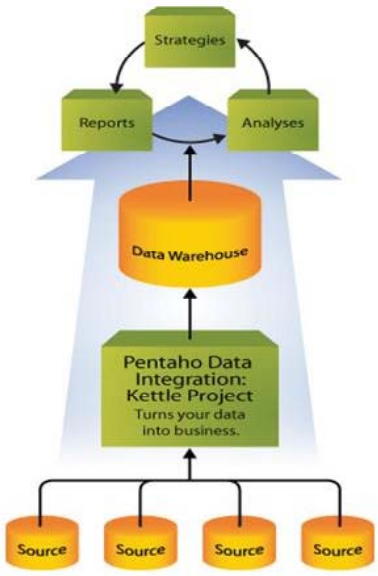

Figura 7.11: Visualización del esquema de *Pentaho Data Integration.*

Características Generales del Weka Esta herramienta *Open Source* incluye las siguientes características:

- Diversas fuentes de datos (ASCII, JDBC).
- Interfaz visual basada en procesos / flujos de datos (rutas).
- Distintas herramientas de minería de datos:
  - Reglas de asociación (a priori, Tertius, etc.).
  - Agrupación / segmentación / conglomerado (cobweb, EM y k-medias).
  - Clasificación (redes neuronales, reglas y árboles de decisión, aprendizaje bayesiano).
  - Regresión (regresión lineal, SVM, etc.).
  - Manipulación de datos (pick & mix, muestreo, combinación, separación, etc.).
  - Combinación de modelos (bagging, boosting, etc.).
  - Entorno de experimentos, con la posibilidad de realizar pruebas estadísticas (T-test).

Figura 7.12: *Weka (Waikato Enviroment for Knowledge Analysis)* http://www. cs. waikato. ac. nz.

Entorno de Trabajo del Weka En la fig. 7.13 de la pág. 296 se visualizará el ambiente de trabajo del weka y posteriormente se podrá analizar en detalle cada entornos de trabajo que esta potente herramienta onpen source posee.

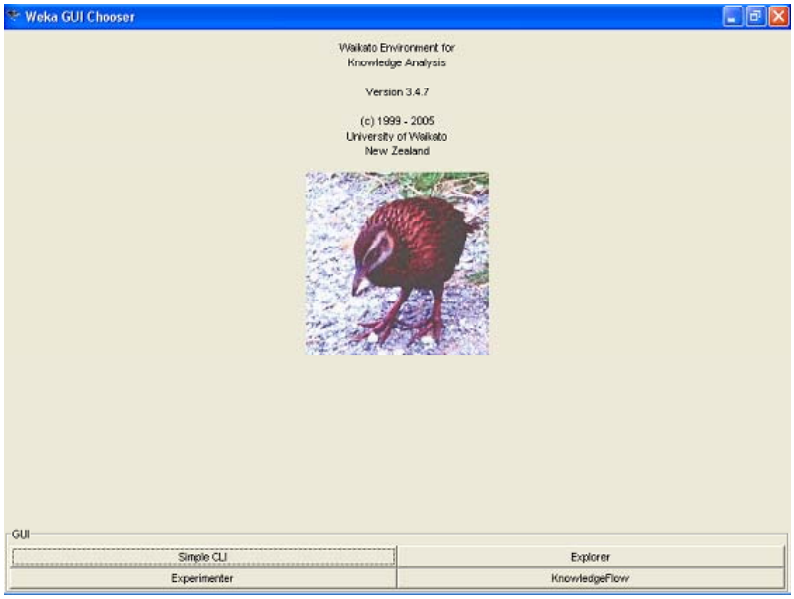

Figura 7.13: Visualización de la ventana principal del *Weka*.

Como se puede ver en la parte inferior de la fig. 7.13 de la pág. 296, *Weka* define cuatro entornos de trabajo diferentes.

Estos entornos son los siguientes:

- *Simple CLI*: Es un entorno consola que permite la invocación directa mediante *Java* a todos los paquetes de *weka.*
- *Explorer*: Es un entorno visual que ofrece una interfaz gráfica para el uso de los paquetes de *weka.*
- *Experimenter*: Entorno centrado en la automatización de tareas de manera que se facilite la

realización de experimentos a gran escala.

- *KnowledgeFlow*: Permite generar proyectos de *minería de datos* mediante la generación de flujos de información o *workflow*.

En este apartado se tratará únicamente el entorno *Explorer*, ya que permite el acceso a la mayoría de las funcionalidades integradas en *Weka* de una manera más sencilla.

La siguiente imagen permiten visualizar el entorno de trabajo que posee *Explorer* (ver fig. 7.14 de la pág. 297).

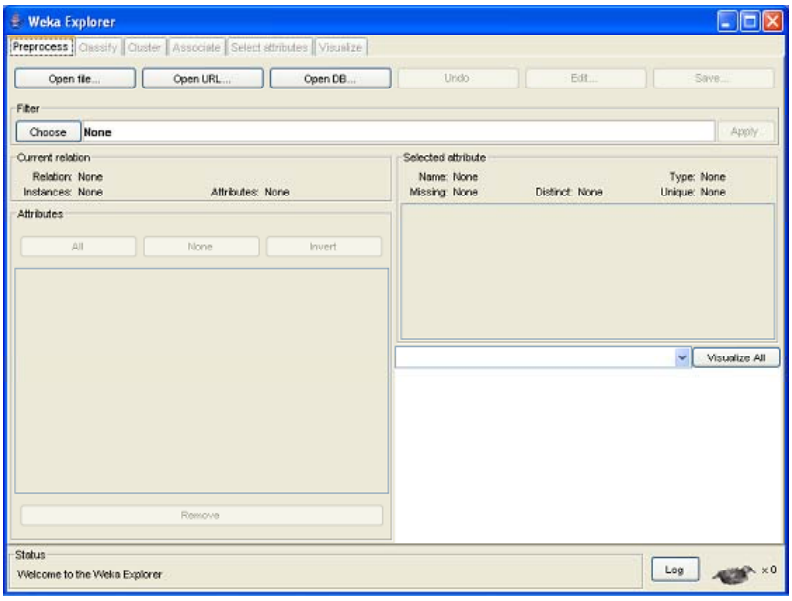

Figura 7.14: Visualización de la ventana del Explorador.

Como se puede observar en la fig. 7.14 de la pág. 297, existen seis subentornos de ejecución:

- *Preprocess:* Incluye las herramientas y filtros para cargar y manipular los datos.
- *Classification:* Acceso a las técnicas de clasificación y regresión.
- *Cluster:* Integra varios métodos de agrupamiento.
- *Associate:* Incluye una pocas técnicas de reglas de asociación.

- *Select Attributes:* Permite aplicar diversas técnicas para la reducción del número de atributos.

- *Visualize:* En este apartado podemos estudiar el comportamiento de los datos mediante técnicas de visualización.

7.3 Proceso de Minería de Datos Aplicando Business Intelligence OpenSource Pentaho

Como se había mencionado anteriormente, el *Proceso de Minería*, está compuesto por los siguientes pasos:

- *Definir el problema.*

- *Preparar los datos.*

- *Explorar los datos.*

- *Generar modelos.*

- *Explorar y validar los modelos.*

- *Implementar y actualizar los modelos.*

En el diagrama de la fig. 7.15 de la pág. 299 se describen las relaciones existentes entre cada paso de un proceso de generación de un modelo de minería de datos.

Aunque el proceso que se ilustra en la fig. 7.15 de la pág. 299 es circular, esto no significa que cada paso conduzca directamente al siguiente.

La creación de un modelo de minería de datos es un proceso dinámico e iterativo.

El objetivo de este apartado no es más que de utilizar las mismas problemáticas volcadas en el *capítulo Nº 6*, en la sección *"Proceso de Minería Aplicada a la EPH"*, donde la principal diferencia se basará en que en este caso se manejarán herramientas del ámbito *Open Source*.

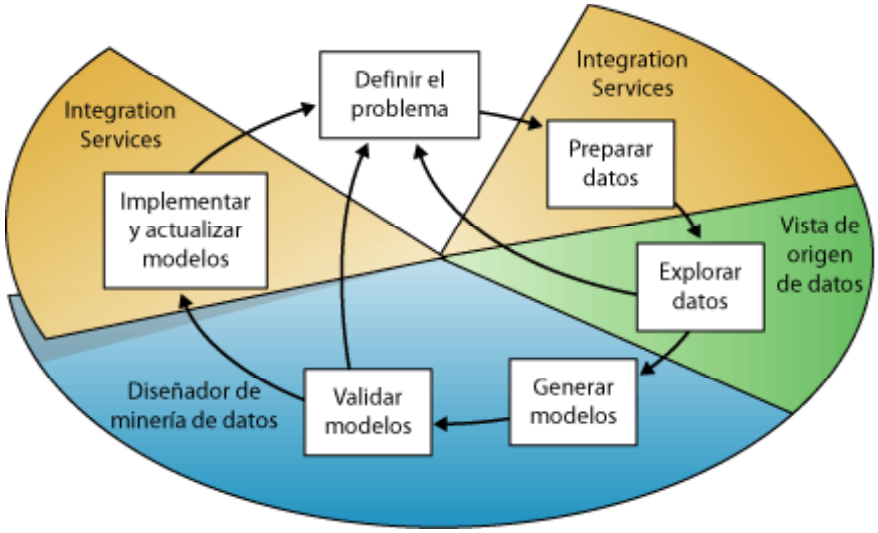

Figura 7.15: Visualización del proceso de generación de un modelo de *minería de datos*.

7.3.1 Definición de los Problemas

Como se hacía referencia anteriormente, no se añadirá ninguna problemática a las que se establecieron en el anterior capítulo.

En esta sección se tratarán los mismos objetivos de estudio ya fijados con anterioridad.

Objetivos Específicos:

- *Describir la composición del empleo en Corrientes.*
- *Conocer los perfiles socio demográficos de los individuos de la población de Corrientes.*

7.3.2 Preparación de los Datos

Nativamente *Weka* trabaja con un formato denominado *arff*, acrónimo de *Attribute-Relation File Format.*

Este formato está compuesto por una estructura claramente diferenciada en tres partes:

- **Cabecera**: Se define el nombre de la relación. Su formato es el siguiente:

  @relation <nombre-de-la-relación>

- **Declaraciones de atributos:** En esta sección se declaran los atributos que compondrán el archivo junto a su tipo. La sintaxis es la siguiente:

  @attribute <nombre-del-atributo> <tipo>

  Donde:

  <nombre-del-atributo> es de tipo string.

  <tipo> acepta diversos tipos, estos son:

  - *NUMERIC* Expresa números reales.
  - *INTEGER* Expresa números enteros.
  - *DATE* Expresa fechas.
  - *STRING* Expresa cadenas de texto.

- **Sección de datos**: Se declaran los datos que componen la relación separando con comas los atributos y con saltos de línea las relaciones. La sintaxis es la siguiente:

  @data

  4,3.2

Una vez conocido el formato de los datos soportado por el *Weka*, se pasará al confeccionado del archivo con extensión *arff*.

Los mismo se pueden convertir ficheros de texto conteniendo un registro por línea con los atributos separados por comas (formato *csv*) a ficheror *arff* mediante el uso de un filtro convertidor.

Con la información recolectada a través de la *EPH (Encuesta Permanente de Hogares)* se han generado una base de datos *Microsoft Access*.

La información será recabada en una planilla de hoja de cálculos *Microsoft Excel*, luego se la convertirá a un documento de texto plano (.txt, .doc, etc.) para su posterior transformación a un archivo de formato específico de datos legible por el *Weka*, el formato *.arff* (ver fig. 7.16 de la pág. 301).

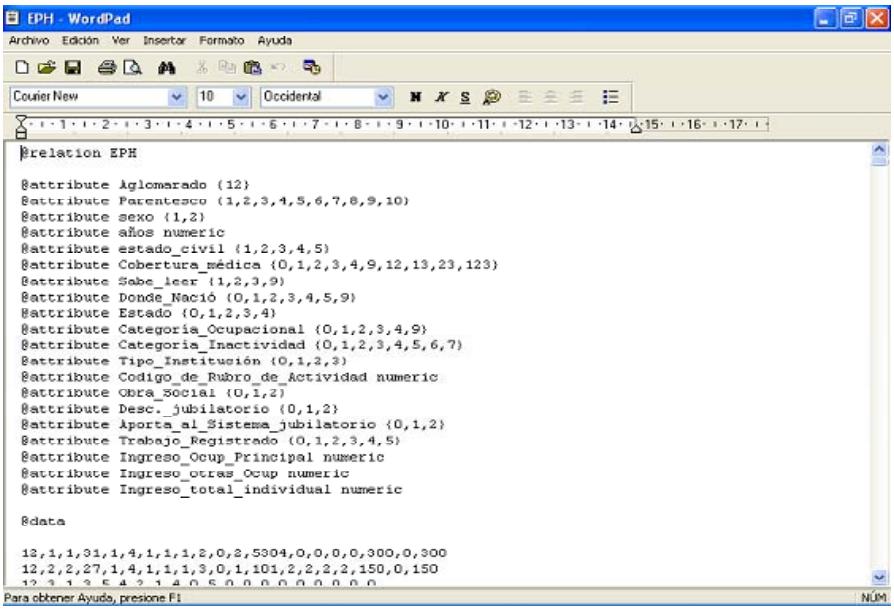

Figura 7.16: Conversión de un archivo plano *(.txt, .doc, etc.)* para la creación de un archivo .arff

7.3.3 Exportación de los Datos

Una vez culminada la etapa de preparación, se pasa a la etapa de exploración de datos.

En este período se comenzará a interactuar con la herramienta.

A continuación se visualizará el archivo confeccionado en el paso anterior, donde este archivo será ejecutado (ver fig. 7.17 de la pág. 302).

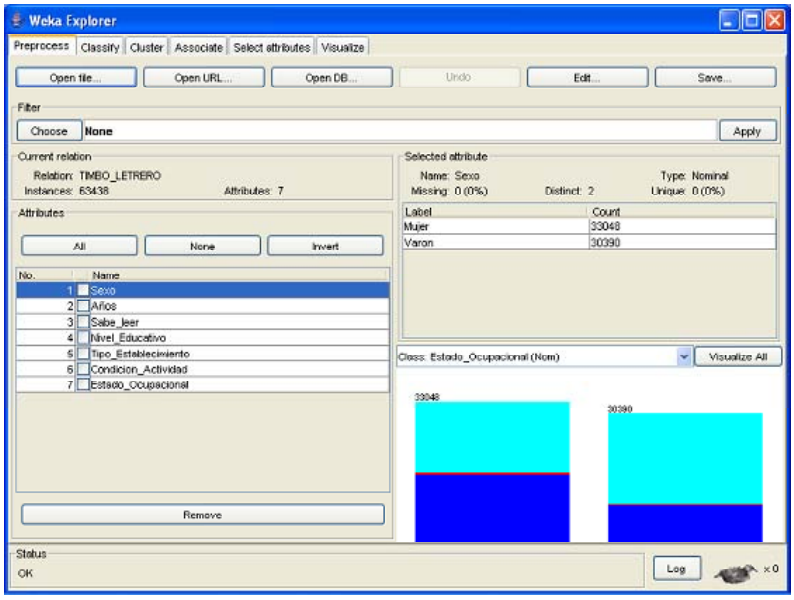

Figura 7.17: Visualización del archivo .arff ejecutado por el *Weka*.

7.3.4 Generación Modelos

En este apartado se analizarán los modelos impuestos en el apartado *"Definición de los Problemas"*.

Describir la composición del empleo en Corrientes.

Como se hacía referencia anteriormente, con el *Weka* no solamente se podrá aplicar técnicas de minería de datos.

En el transcurso del estudio relacionado con este objetivo se utilizará únicamente análisis de las variables.

A continuación se visualizan distintos análisis de las variables referentes a los estados de actividad de los individuos de la provincia de Corrientes.

En el gráfico de la fig. 7.18 de la pág. 303 se puede visualizar la frecuencia absoluta (número de casos) de la variable de estudio que en este caso es *estado (condición de actividad)*.

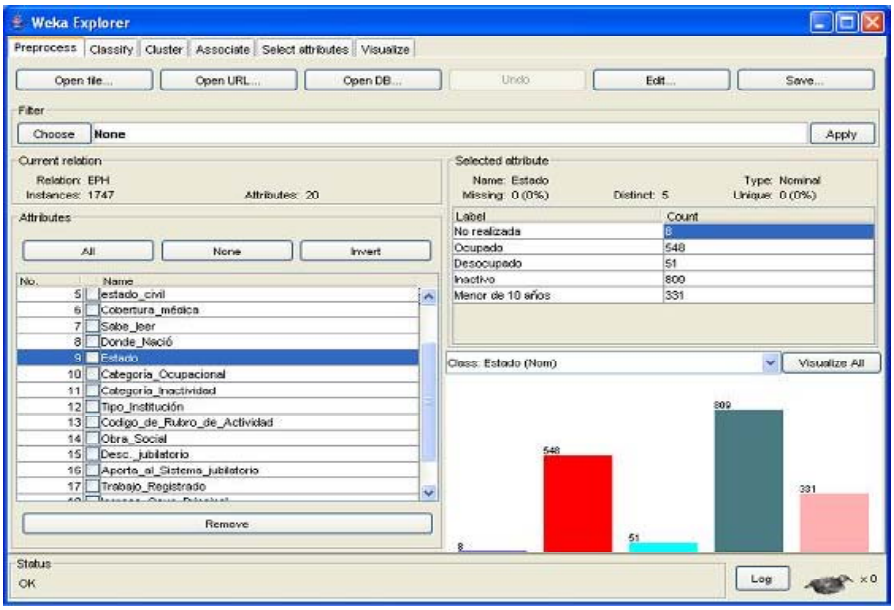

Figura 7.18: Muestreo de la población total con respecto a la condición de actividad (variable estado).

También se pude visualizar cómo se manifiesta la variable *estado (condición de actividad)* con las demás variables de la muestra (ver fig. 7.19 de la pág. 304).

Figura 7.19: Visualizacion de todas las varibles con respecto al condicion de actividad.

Como resultado de esta clasificación, se visualizó el número de *ocupado, desocupados, inactivos, etc.*

Similar al anterior procedimiento, se puede realizar con las variables *cat_ ocup (categoría ocupacional)* o incluso con *cat_ inac (categoría de inactividad)* (ver fig. 7.20 de la pág. 305 y fig. 7.21 de la pág. 306).

Conocer los perfiles socio demográficos de los individuos de la población de Corrientes.

A diferencia del anterior apartado, en este se utilizarán técnicas de *Minería de Datos.*

Lo que interesa en este caso es descubrir los diferentes perfiles de los individuos que poseen planes asistenciales en la provincia de Corrientes.

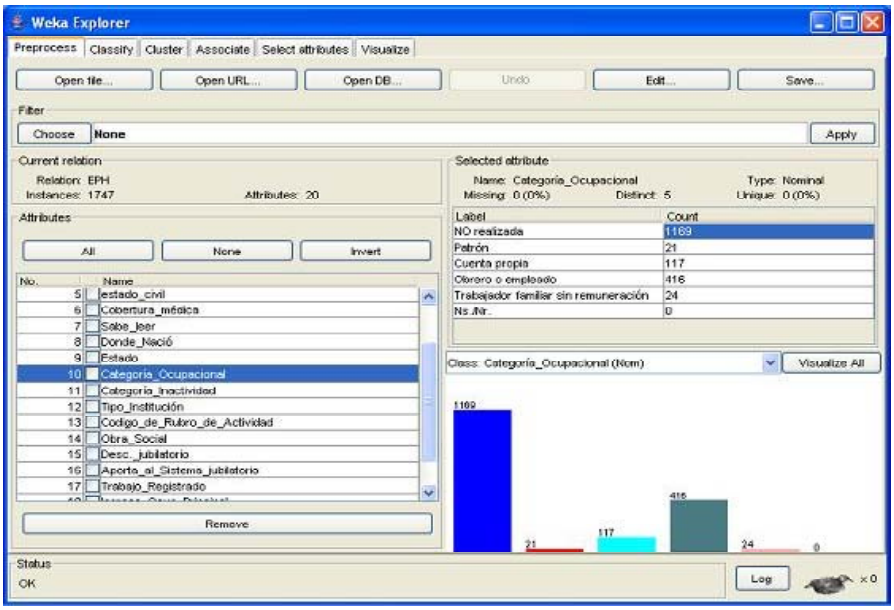

Figura 7.20: Muestreo de la población total con respecto a la categoría de ocupacional (variable CAT_OCUP).

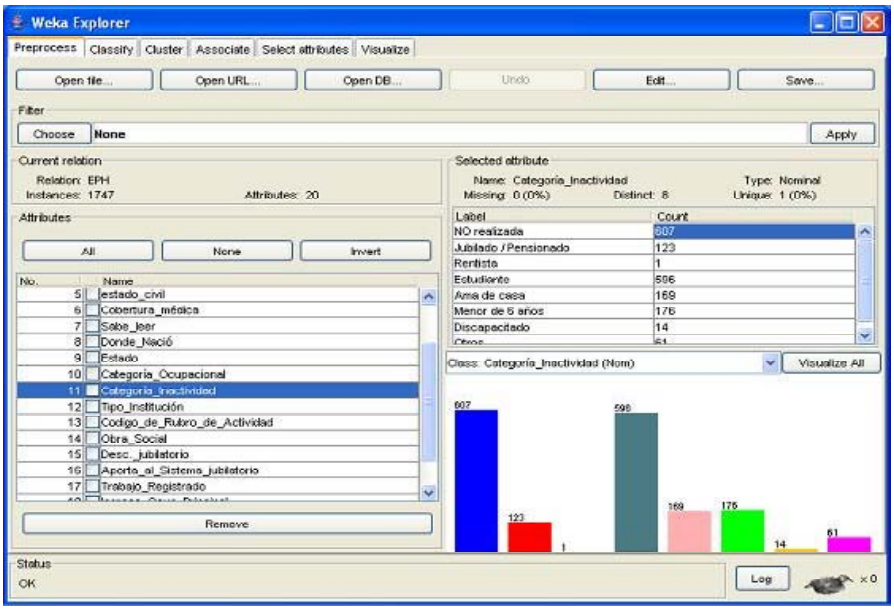

Figura 7.21: Muestreo de la población total con respecto a la categoría de incatividad (variable CAT_INAC).

Para ello se empleará la técnica de *Clustering* con un algoritmo *SimpleKMeans* , utilizando el atributo *sexo* para la distribución de los grupos.

Se obtendrá un modelo de minería de datos donde se dividirán todos los individuos de la población de Corrientes en los grupos correspondientes a la variable sexo.

Una vez culminado el proceso de *Clustering* la herramienta nos permite observar los resultados de modo textual (ver fig. 7.22 de la pág. 307) o también de manera grafica (ver fig. 7.24 de la pág. 309).

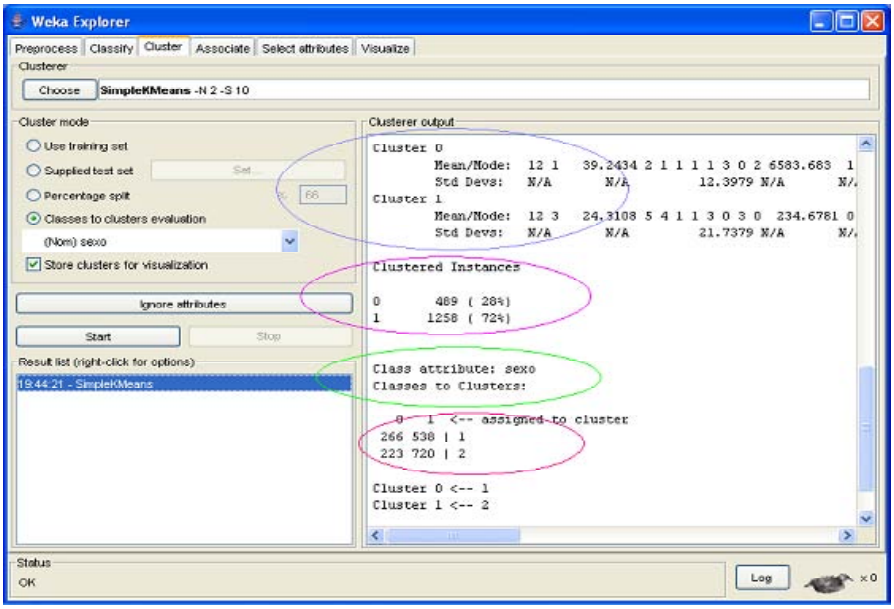

Figura 7.22: Información detallada de los cluster como ser (centros, instancias, asignación).

En la fig. 7.22 de la pág. 307 se puede apreciar la información sobre el número de clusters involucrados, las instancias de estos, como así también las clases y los atributos que participan en este análisis.

También se visualiza a los dos grupos, donde:

Cluster 0 <– 1 (Varón)

Cluster 1 <– 2 (Mujer)

Además se pude apreciar en la 7.22 de la pág. 307 a cuatro círculos de diferentes colores; cada uno de estos destacan la siguiente información:

- El primero círculo de color violeta destaca la distribución de los atributos en cada cluster.
- El segundo, de color rosado muestra en porcentaje y en frecuencia el número de instancias por cluster.
- El tercer círculo visualiza el atributo en este caso es la variable sexo con el cual se realizó el análisis.
- El último muestra la asignación de cada cluster por cada valor de la variable sexo, con su respectivo número de casos.

Si se presiona con el botón derecho del ratón sobre la lista de resultado (ver fig. 7.23 de la pág. 308) se puden observar los correspondientes resultados extraídos de la técnica de *Clustering* en forma gráfica (ver fig. 7.24 de la pág. 309).

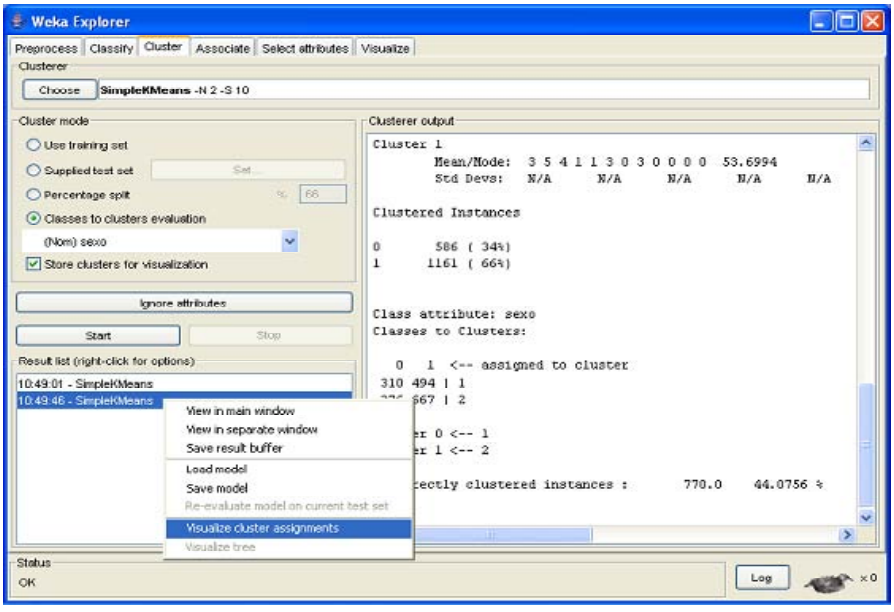

Figura 7.23: Accediendo a la visualización de los cluster de manera gráfica.

A continuación se visualizarán los resultados resultados extraídos de la técnica de *Clustering*.

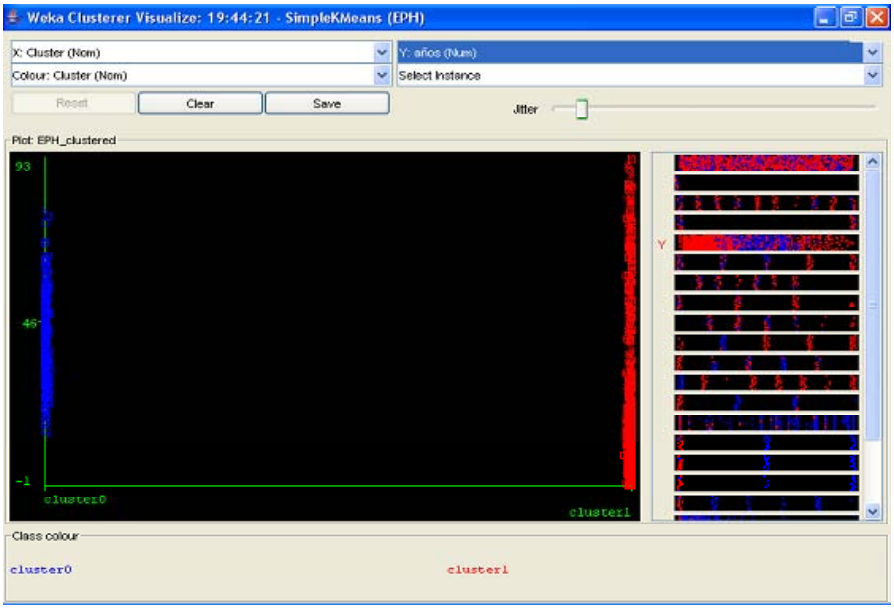

Figura 7.24: Visualización de la distribución de los cluster con respecto de la variable años.

En la fig. 7.24 de la pág. 309 se muestra la dispersión de la variable *años* en cada cluster.

Donde:

- *Cluster 0* de colo azul.
- *Cluster 1* de color rojo.

En la fig. 7.25 de la pág. 310 se pueden observar los valores que toma cada cluster de la variable que indica el analfabetismo.

En el gráfico de la fig. 7.25 de la pág. 310 permite extraer la siguiente infomación:

- El *Cluster 0* asume el valor 1 (1= Sí sabe leer y escribir; 2= No sabe leer y escribir).

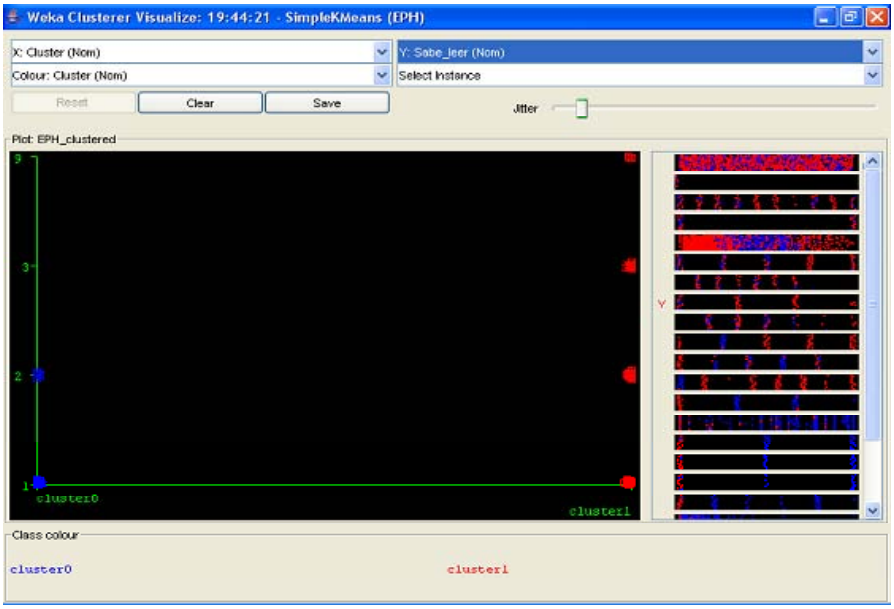

Figura 7.25: Visualización de la distribución de los cluster con respecto de la variable *analfabetismo*.

- El *Cluster 1* asume todos los valores restantes.

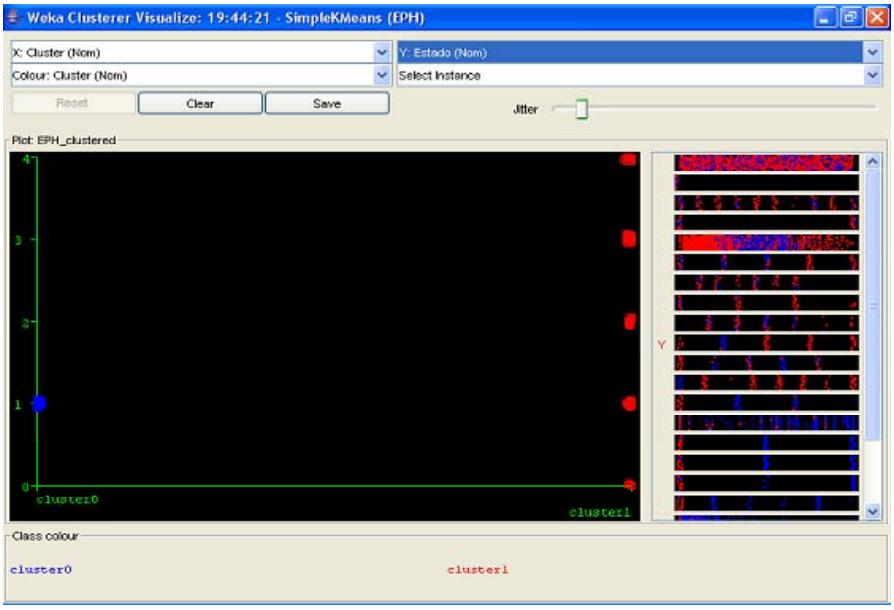

Figura 7.26: Visualización de la distribución de los cluster con respecto de la variable *estado (condición de actividad).*

Como se puede comprobar en la fig. 7.26 de la pág. 311, el *Cluster 0* asume únicamente el valor 1 (*Ocupado*), en cambio el *Cluster 1* el resto de los valores.

Cuando se contrasta la variable *cat_ ocup (categoria ocupacional)* con respecto a los cluster se puede comprobar lo siguiente:

- *Cluster 0* asume todos los valores exepto el 0 (cero), con mayor presencia de instancia en el valor 3 y con un importante número menor en la opción 4.
- *Cluster 1* asume todos los valores inclusive el 0 (cero).

Donde:

  - 0 = Entrevista individual no realizada.

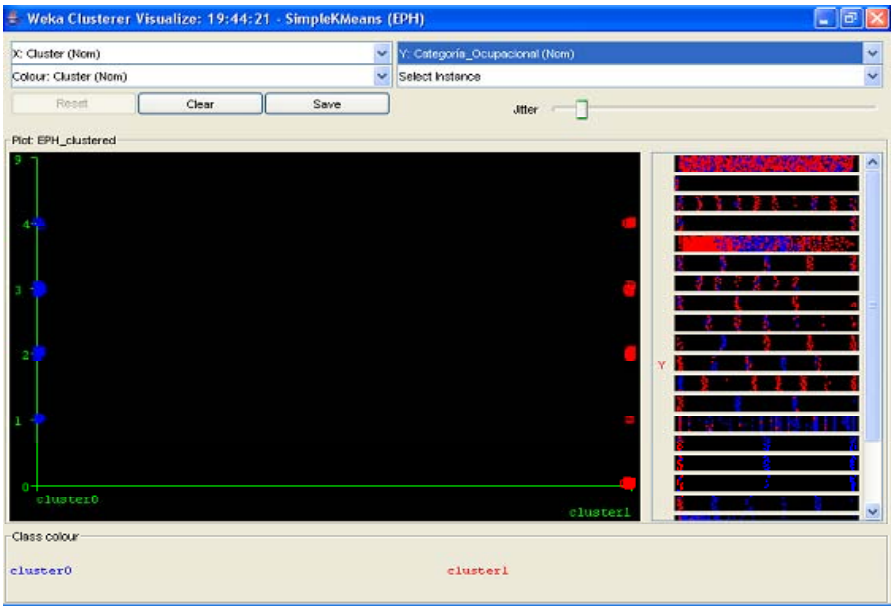

Figura 7.27: Visualización de la distribución de los cluster con respecto de la variable *cat_ ocup (categoría ocupacional)*.

- 1 = Patrón.
- 2 = Cuenta propia.
- 3 = Obrero o empleado.
- 4 = Trabajador familiar sin remuneración.
- 9 = Ns./Nr.

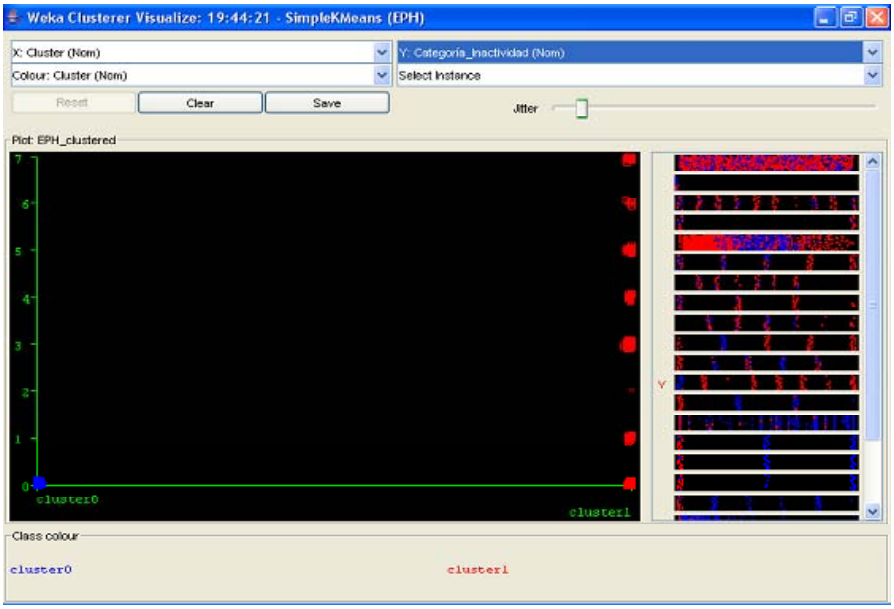

Figura 7.28: Visualización de la distribución de los cluster con respecto de la variable *cat_ inac (categoría de inactividad).*

En la fig. 7.28 de la pág. 313 permite observar los valores que poseen los diferentes cluster.

Como por ejemplo:

- *Cluster 0* asume el único valor 0 (cero), 0 = Entrevista individual no realizada.
- *Cluster 1* toma todos los valores que asume la variable, es decir:
  - 1 = Jubilado / Pensionado.

- 2 = Rentista.
- 3 = Estudiante.
- 4 = Ama de casa.
- 5 = Menor de 6 años.
- 6 = Discapacitado.
- 7 = Otros.

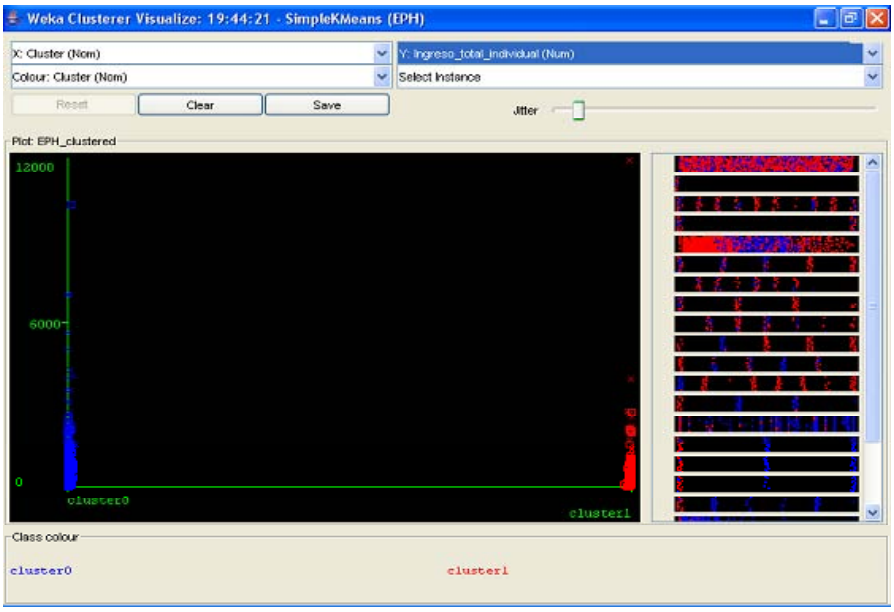

Figura 7.29: Visualización de la distribución de los cluster con respecto de la *ingreso total individual.*

En la fig. 7.29 de la pág. 314 se puede visualizar que la distribución de los ingresos de los individuos en el *Cluster 0* es superior que el *Cluster 1.*

El gráfico 7.29 de la pág. 314 permite comprobar que el *Cluster 0* supera los 6000 pesos y el *Cluster 1* no solamente no supera esta cifra si no que también posee menor número de casos.

Lo que se realizó hasta aquí es una descripción de *perfiles de los individuos* por la variable *sexo.*

Si se quisiera conocer la representación de los perfiles pero en este caso utilizando la variable *estado*, se procederá como se detalla a continuación.

Como se puede visualizar en la fig. 7.30 de la pág. 315 la variable *estado* es la *variable activa* y parentesco, sexo, estado civil, cobertura médica, sabe leer, donde nació, categoría ocupacional, categoría inactividad, tipo de institución, obra social, desc. jubilatorio, aporta al sistema jubilatorio, trabajo registrado y ingreso total individual son las *variables complementarias* de dicho proceso.

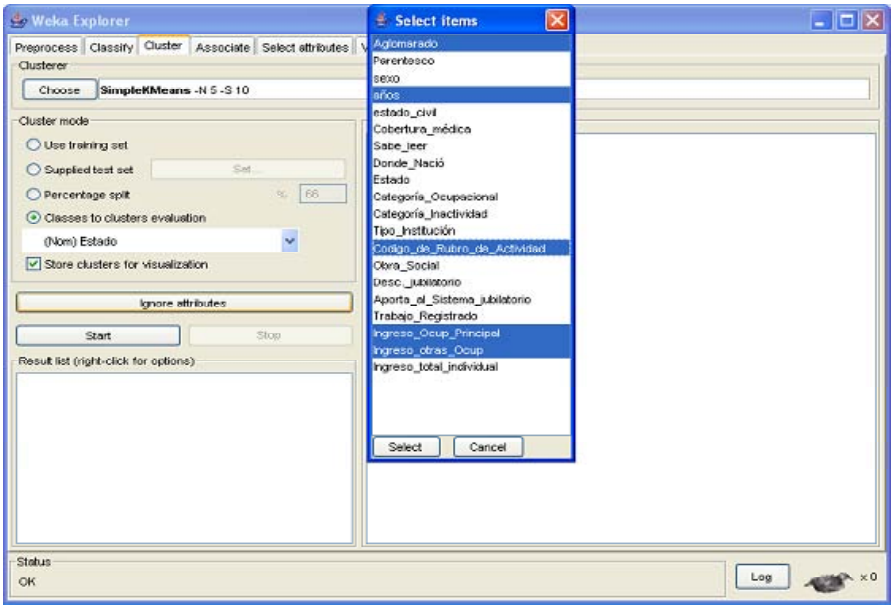

Figura 7.30: Selección de las variables activas y complementarias de este proceso de minería de datos.

Una vez ejecutado dicho proceso se obtienen los siguientes resultados, ya sea en formato textual como gráfico (ver fig. 7.31 de la pág. 316 y fig. 7.32 de la pág. 317).

En la fig. 7.31 de la pág. 316 se puede observar la siguiente información: *el número de cluster involucrados, las instancias de estos, como así también las clases y los atributos que participan en este análisis.*

En la fig. 7.32 de la pág. 317 se puede observar la formación de los diferentes clústers, donde los mismos representan distintos *estados.*

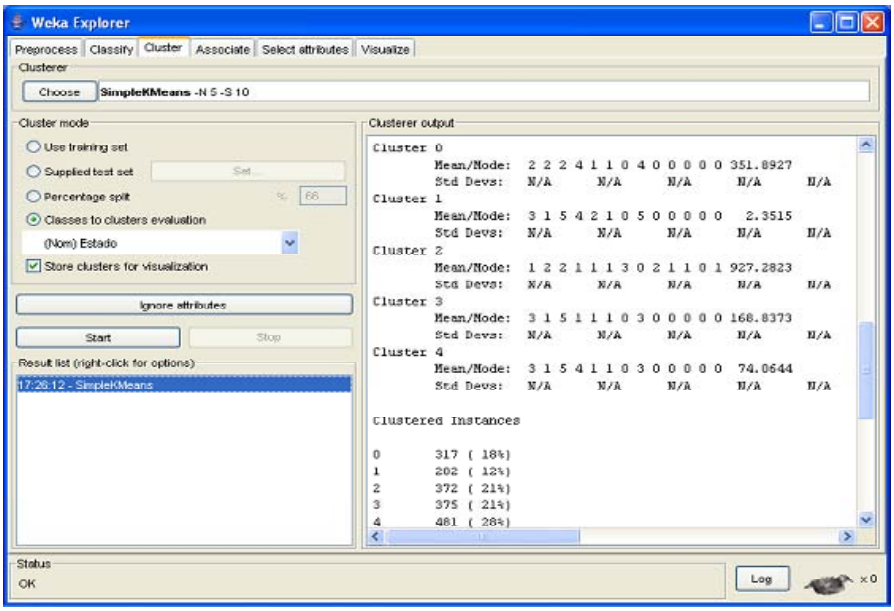

Figura 7.31: Visualización de los resultados de manera textual, donde se pueden observar entre otras cosas *número de cluster involucrados, los atributos que participan en este análisis, etc.*

Figura 7.32: Visualización de los distintos grupos conformados por la herramienta.

Por ejemplo:

- El *clúster Nº 0* asume los siguientes valores (ocupado, desocupado e inactivo).
- El *clúster Nº 1* está compuesto por mayor presencia de la población menor de 10 años y con una inferior representación en las poblaciones inactivas y ocupadas.
- El *clúster Nº 2* posee únicamente a los individuos que se encuentren ocupados.
- Los *clúster Nº 3* y *4* poseen casi la misma distribución, con la diferencia que el *clúster Nº 4* asume el valor 0 (cero) que es la no respuesta al cuestionario individual.

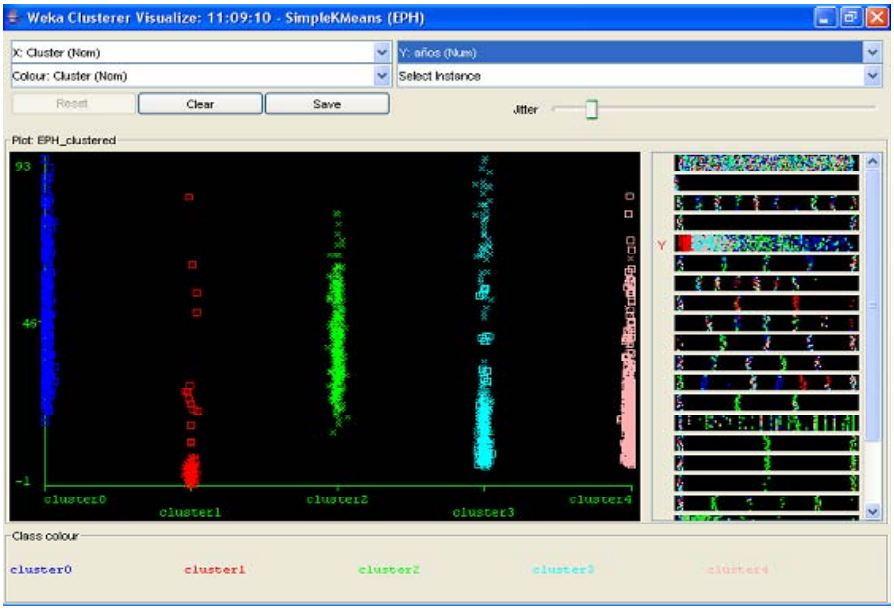

Figura 7.33: Visualización de la confección de los grupos con respecto a las edades.

Como se puede visualizar en el gráfico 7.31 de la pág. 316, las edades están estrechamente relacionadas a los estados (condición de actividad) de los individuos:

- El *clúster Nº 0* corresponde a los individuos que no respondieron al cuestionario individual.
- El *clúster Nº 1* (población menor de 10 años de edad) contiene a los menores de 10 años en el gráfico.
- El *clúster Nº 2* (población ocupada) siendo estas las edades más productivas de la población.
- El *clúster Nº 3* (población inactiva) posee una distribución de edades diferente a la anterior distribución ya que en este grupo se encuentran estudiantes, amas de casa, etc., entre otros.
- El *clúster Nº 4* (población desocupada) corresponde a los desocupados, con su respectiva distribución de edades.

Capítulo 8

Aplicación Web Multiplataforma

8.1 Descripción

Este trabajo se basa en el estudio del software que permite el desarrollo de aplicaciones *Web multiplataforma* con acceso a base de datos distribuidas y en el desarrollo de una aplicación *Web* que brinda información sobre resultados de procesos de *minería de datos*.

El objetivo es realizar una aplicación *Web multiplataforma* desarrollada en *Java* , mediante la cual el usuario pueda contar con un medio de visualización de resultados de procesos de minería.

El sistema funciona en distintas plataformas mediante el uso de *software multiplataforma*.

Considerando que la información es poder, es muy importante el lugar en donde se almacena, su organización y la forma en que ésta puede brindarse a los distintos usuarios.

Además, el desarrollo del presente trabajo se ve motivado por la posibilidad de obtener experiencias y conocimientos vinculados con entornos de trabajo propios del mercado comercial en gran escala, de la seguridad que precisa en el manejo de la información y de la importancia actual de la interrelación de las actividades de las organizaciones desde el punto de vista de su gestión interna

y su interrelación con el medio mediante la *Web*, todo ello en el contexto mayor de la sociedad de la información y el conocimiento.

En la fig. 8.1 de la pág. 322 se puede visualizar a la página principal del sistema. Mediante cualquier navegador de *Internet* se puede ingresar y navegar por el sitio *Web* de la aplicación.

Figura 8.1: Página Principal de la Aplicación Web.

Como se hacía referencia anteriormente el usuario que ingrese a la página deberá ingresar su usuario y su contraseña, una vez registrado podrá acceder a toda la información disponible en el *sitio Web*.

En la página *resul.html* (ver la fig. 8.2 de la pág. 323) se puede elegir qué resultados se desea visualiza.

Estos resultados son los siguientes:

- *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.*

Figura 8.2: Visualización de la Página *resul.html*.

Los resultados fueron extraídos de la bases de datos de la *Encuesta Permanente de Hogares (EPH)*. Para obtener más información acerca de Extracción de Conocimiento con *IBM DB2 Intelligent Miner for Data* ver el Capítulo Nº6 *(Extracción de Conocimiento con IBM DB2 Intelligent Miner for Data)*.

Según la opción elegida en los links de la página *resul.html* (ver la fig. 8.2 de la pág. 323), se pueden visualizar los resultados de minería obtenidos con los datos de la *Encuesta Permanente de Hogares (EPH)*.

Si la opción elegida es *Perfiles Socio-Demográficos*, se podrán visualizar los perfiles demográficos de los individuos que posean planes asistenciales del aglomerado de Corrientes. Esta información estará disponible en la página *demografico.html* (ver la fig. 8.3 de la pág. 324).

En la fig. 8.3 de la pág. 324 se pueden observar distintos clúster con sus respectivos porcentajes como resultado general.

También se puede visualizar en detalle la composición de cada uno de estos

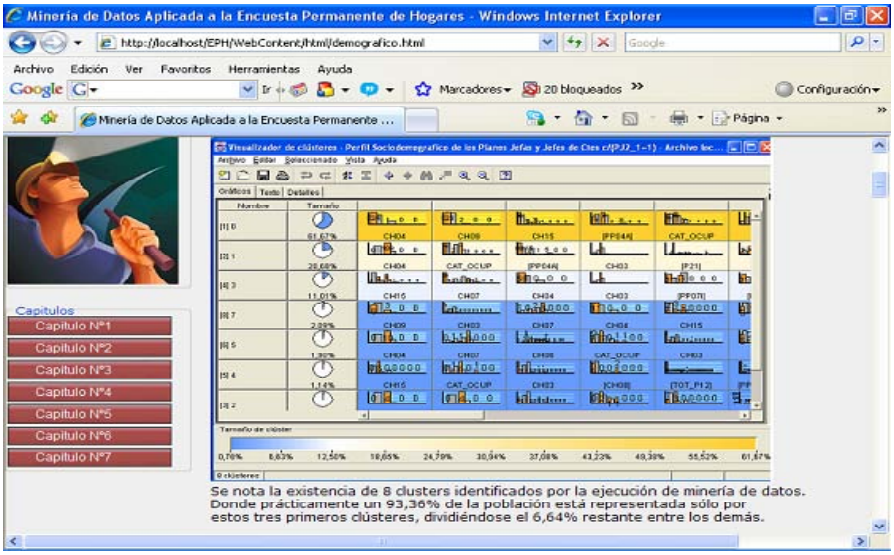

Figura 8.3: Visualización de la página Web (*demografico.html*).

clústeres, como se puede comprobar en la fig. 8.4 de la pág. 325, la fig. 8.5 de la pág. 325 y la fig. 8.6 de la pág. 326.

También se puede obtener información a nivel general respecto a los resultados seleccionados en la página (*resul.html*). Por ejemplo si se seleciona "*Clasificación de las Características Educacionales de Cada Individuo, en Base a sus Ingreso*", el tipo de información disponible este caso será la representación grafica de los resultados extraídos de un modelo de *Árboles de Decisión* (ver la fig. 8.7 de la pág. 326).

A continuación se pueden observar cada una de las reglas extraídas del *Árbol de Decisión* (ver la fig. 8.8 de la pág. 327, la fig. 8.9 de la pág. 327 y la fig. 8.10 de la pág. 328).

En esta aplicación no solamente se puede visualizar resultados extraídos con el *DB2 Intelligent Miner for Data* o *Weka* (ver la fig. 8.2 de la pág. 323), si no también se puede recurrir a información bibliográfica, conclusiones e incluso observar todos los capítulos del libro en formato digital (ver la fig. 8.11 de la pág. 329, la fig. 8.12 de la pág. 330 y la fig. 8.13 de la pág. 331).

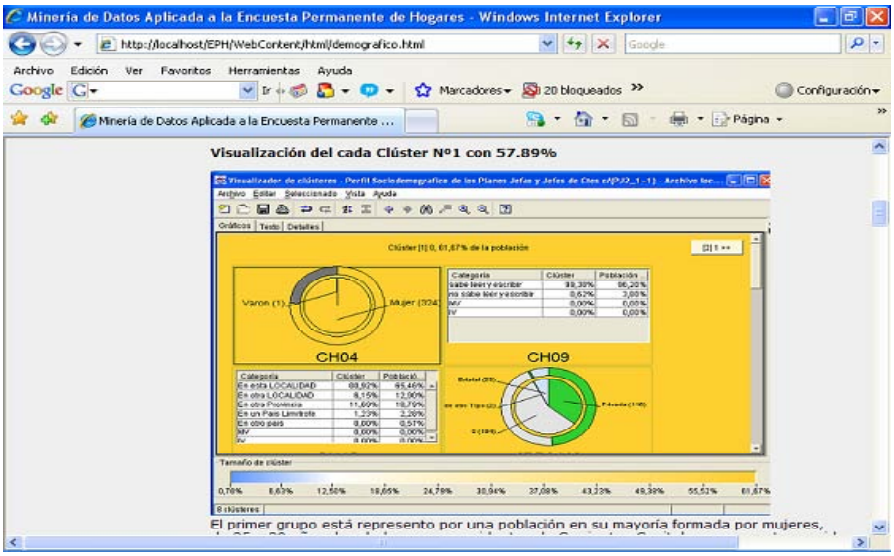

Figura 8.4: Visualización de la página Web (*demografico.html*), resultados del Clúster N^o 1 57.89 de la población total. %.

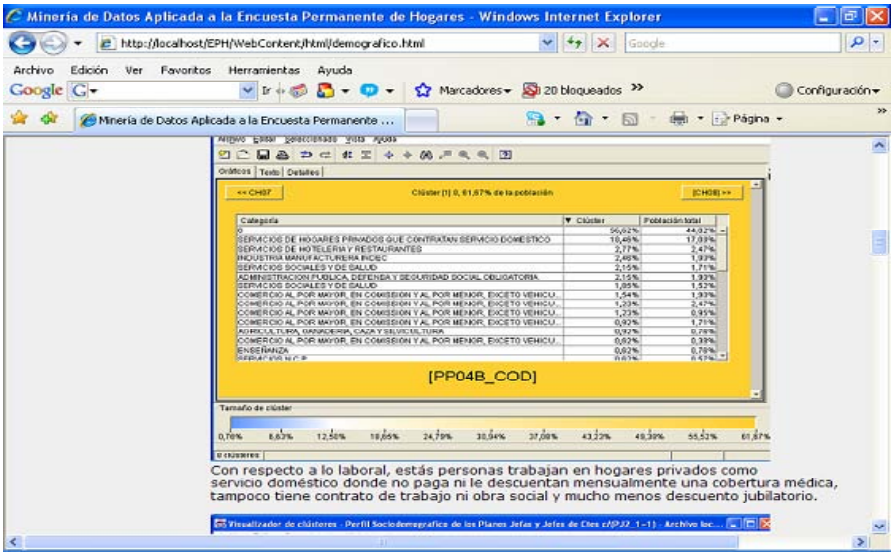

Figura 8.5: Visualización de la variable PP04_COD, del clúster Nº1.

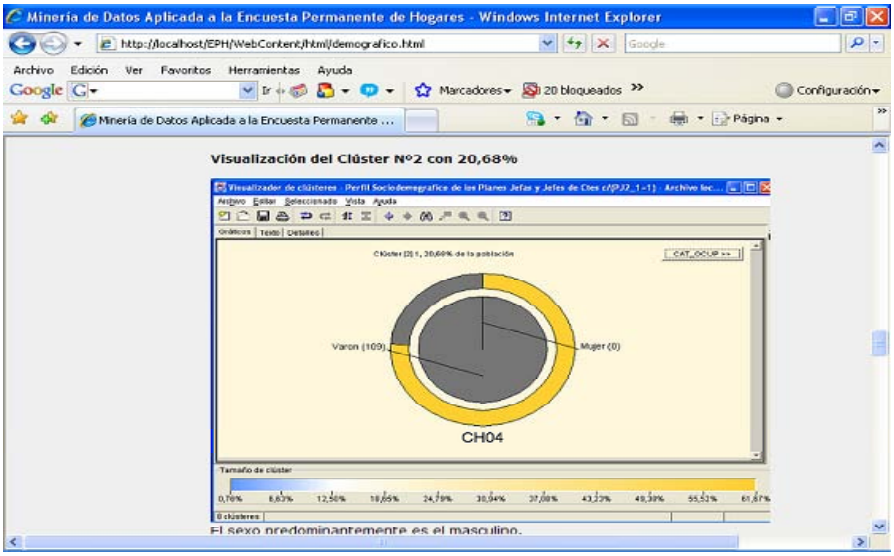

Figura 8.6: Visualización de la variable CH04 (sexo), del clúster Nº2 con 20,68% de la población total.

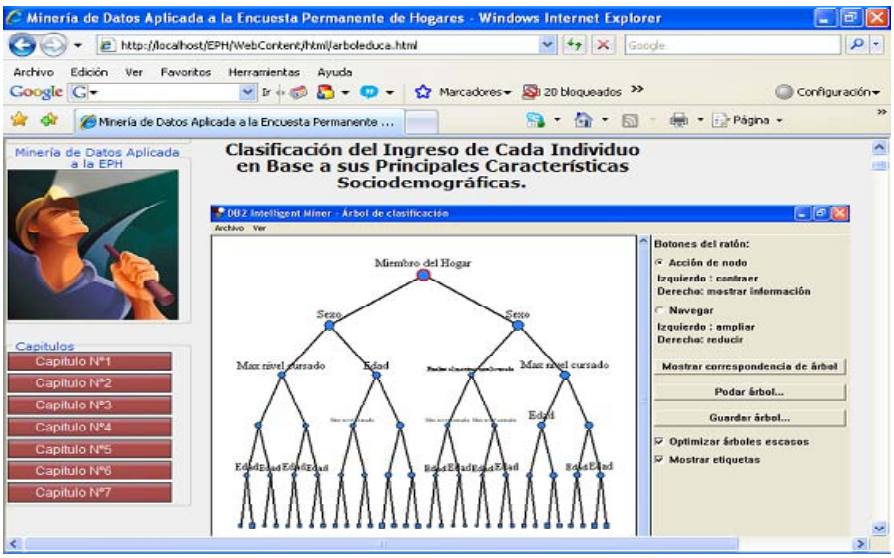

Figura 8.7: Visualización de la página Web (*arboleduca.html*).

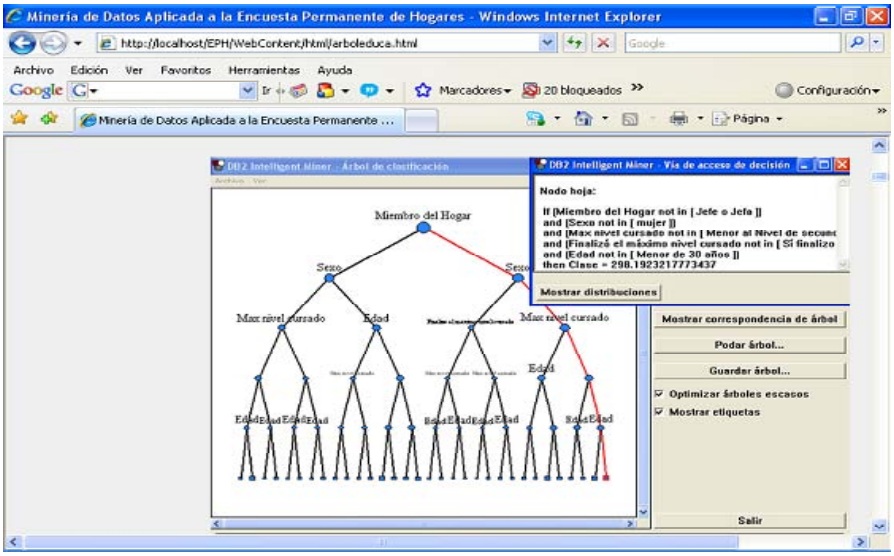

Figura 8.8: Visualización de las regla N^o 1 del Árbol de Decisión en la página web (*arboleduca.html*).

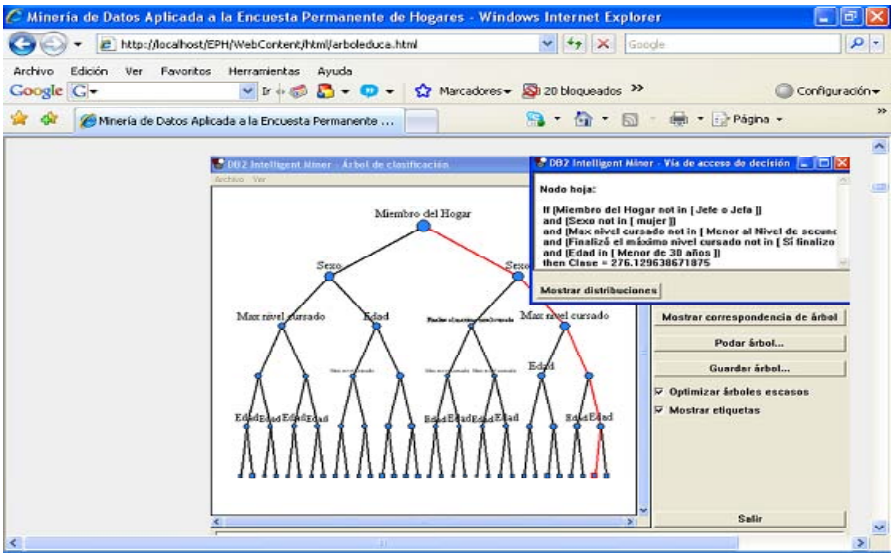

Figura 8.9: Visualización de las regla N^o 2 del Árbol de Decisión en la página web (*arboleduca.html*).

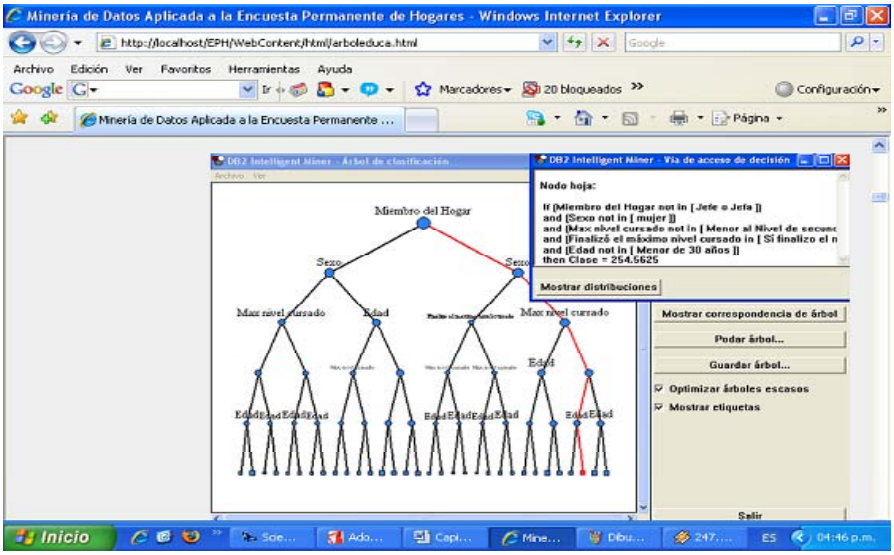

Figura 8.10: Visualización de las regla Nº 3 del Árbol de Decisión en la página web (*arboleduca.html*).

8.2 Ejemplos de Servlet y Páginas en HTML

A continuación se transcribe el servlet que integra la aplicación.

Login_Controller.java

```
import java.io.IOException;

import javax.servlet.ServletException;

import javax.servlet.http.HttpServlet;

import javax.servlet.http.HttpServletRequest;

import javax.servlet.http.HttpServletResponse;

import javax.servlet.ServletConfig;

import java.sql.*;

import javax.servlet.http.HttpSession;
```

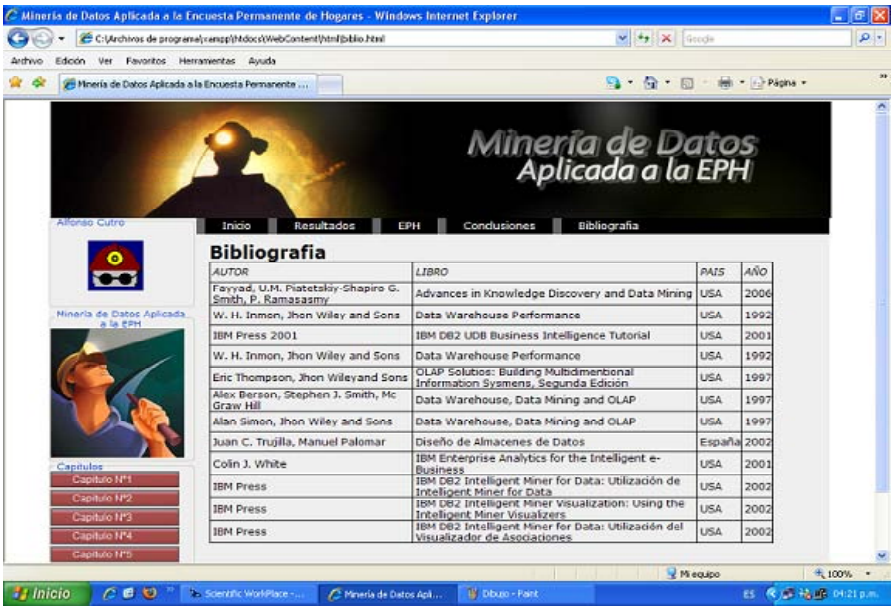

Figura 8.11: Visualización de la pagína que contiene informacíon biografíca (*biblio.html*).

Figura 8.12: Visualización de la pagína que posee las conclusiones (*conclu.html*).

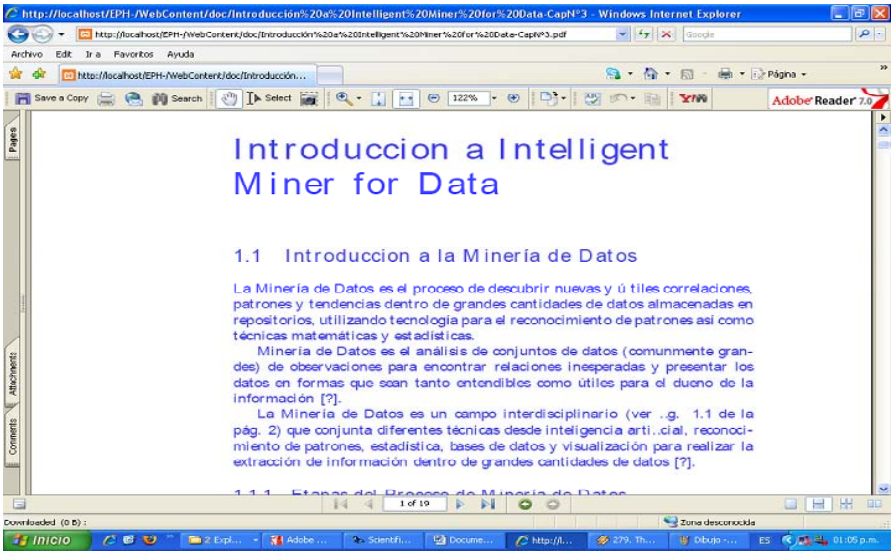

Figura 8.13: En este portal Web se puede visualizar todos los capítulos de libro.

```
/**
 * @version 1.0
 * @author
 */
public class Login_Controller extends HttpServlet {
 /**
 * @see javax.servlet.http.HttpServlet#void (javax.servlet.http.HttpServletRequest,
      javax.servlet.http . HttpServletResponse)
 */
 private Connection conn = null;
 public void doGet(HttpServletRequest req, HttpServletResponse resp)
 throws ServletException, IOException {
 String usuarioID = req.getParameter("usuario");
```

```
String clave = req.getParameter("clave");

Statement stmt = null;

ResultSet rs = null;

String select = "select * from sergio.usuario where usuario="'+ usuarioID +"' and
    "+ "clave="'+ clave +""';

System.out.println(select);

try

{

stmt = conn.createStatement();

rs = stmt.executeQuery(select);

if (rs.next()!=false)

{

HttpSession session = req.getSession();

session.setAttribute("usuarioNombre", rs.getString("nombre"));

session.setAttribute("usuarioApellido", rs.getString("apellido"));

getServletContext().getRequestDispatcher("logueado.jsp").forward(req, resp);

}else

{ System.out.println("error login");

String error = new String("si");

req.setAttribute("error", error);

getServletContext().getRequestDispatcher("index.jsp").forward(req, resp);

}

}catch(Exception e){}

}

/**
```

```
* @see javax.servlet.http.HttpServlet#void (javax.servlet.http.HttpServletRequest,
    javax.servlet.http . HttpServletResponse)

*/

public void doPost(HttpServletRequest req, HttpServletResponse resp)

throws ServletException, IOException {

String nombre = req.getParameter("nombre");

String apellido = req.getParameter("apellido");

String usuario = req.getParameter("usuario");

String clave = req.getParameter("pass");

Statement stmt = null;

String select;

try

{

stmt = conn.createStatement();

select = "INSERT INTO SERGIO.USUARIO (ID,NOMBRE, APELLIDO, USUA-
    RIO, CLAVE) VALUES (DEFAULT," + ""'+ nombre + "','"+ apellido +
    "','"+      usuario + "','" + clave + "')";

System.out.println(select);

int nfilas = stmt.executeUpdate(select);

//System.out.println(nfilas); imprime la cantidad de filas involucradas en la consulta

}catch(Exception e){ System.out.println("error al ejecutar sentencia sql");}

finally {

try {

if (stmt != null) stmt.close();

} catch (SQLException e) {}
```

```
}
getServletContext().getRequestDispatcher("registro.html").forward(req, resp);
}
/**
* @see javax.servlet.GenericServlet#void ()
*/
public void init(ServletConfig config) throws ServletException {
super.init(config);
try {
Class.forName("COM.ibm.db2.jdbc.app.DB2Driver");
conn = DriverManager.getConnection("jdbc:db2:EHP");
} catch(Exception e) {
System.out.println("Error al cargar el driver");
System.out.println(e.getMessage());
}
}
}
```

Seguidamente se transcribe una de las páginas que integran la aplicación, por considerársela representativa de la mayoría de las páginas utilizadas.

biblio.jsp

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN"" http://www.w3.org
    / TR/xhtml11/DTD/xhtml11.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
```

```
<head>
<%@ page
language="java"
contentType="text/html; charset=ISO-8859-1"
pageEncoding="ISO-8859-1"
session="true"
%>
<meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1" />
<meta name="GENERATOR" content="IBM WebSphere Studio" />
<meta http-equiv="Content-Style-Type" content="text/css" />
<link href="theme/eph.css" rel="stylesheet" type="text/css"/>
<title>biblio.jsp</title>
</head>
<body>
 <div id="all">
 <div class="header">
 <table width="900" border="0">
 <tr>
 <td><img src="img/datawarehouse.jpg" width="110" height="140" ></td>
 <td><h1>Minería de Datos Aplicada a la EPH</h1></td>
 </tr>
 </table>
 </div><!–end header–>
 <div class="content">
```

```
<div class="column left">
<fieldset class="buscador">
<% String usuario = (String)session.getAttribute("usuarioNombre");
String apellido = (String)session.getAttribute("usuarioApellido");
%>
<legend><%=usuario%> <%=apellido%></legend>
<table width="150" >
<tr>
<td> </td>
</tr>
<tr>
<td align="center"><img src="img/idmsyst2.gif" /></td>
</tr>
<tr>
<td> </td>
</tr>
</tr>
</table>
</fieldset><!– fin del buscador–>
<fieldset class="buscador">
<legend>Minería de Datos Aplicada
<p> a la EPH</legend>
<div id="navcontainer">
<img src="img/minero.JPG" >
```

```
</div>
</fieldset>
<fieldset class="buscador">
<legend>Capitulos</legend>
<div id="navcontainer">
<ul id="navlist">
<li>
<a href="doc/tfcutro.pdf">Capitulo Nº1</a>
</li>
<li>
<a href="doc/Introducción a el DB2-CapNº2.pdf " >Capitulo Nº2</a>
</li>
<li>
<a href="doc/Introducción a Intelligent Miner for Data - CapNº3.pdf " >
Capitulo Nº3 </a>
</li>
<li>
<a href="doc/Introducción al WebSphere Studio - Cap Nº4.pdf " >
Capitulo N º4</a>
</li>
<li>
<a href="doc/Creacion del Data Warehouse - CapNº5.pdf " >
Capitulo Nº5</a>
</li>
```

```
<li>
<a href="doc/Extraccion de Conocimientos con IBM Intelligent Miner - CapN º6 .
    pdf " > Capitulo N º6 </a>
</li>
<li><a href="doc/">Capitulo Nº7</a></li>
</ul>
</div>
</fieldset><!– fin del buscador–>
</div><!–end left–>
<div class="column main2col">
<div id="header-menu">
<ul id="navi">
<li><a href="index.jsp">Inicio</a></li>
<li><a href="html/resu.html">Resultados</a></li>
<li><a href="EPH.jsp">EPH</a></li>
<li><a href="conclu.jsp">Conclusiones</a></li>
<li><a href="biblio.jsp">Bibliografia</a></li>
</ul>
</div>
<!– ak ingresa toda la informacion que va variar (imagens, textos)–>
<div class="contenido">
<h2>Bibliografia</h2>
<table width="650" border="0" cellpadding="0" cellspacing="0" class="biblio">
<tr>
<td><i>AUTOR</i></td>
```

```
<td><i>LIBRO</i></td>
<td><i>PAIS</i></td>
<td class="derecho"><i>AÑO</i></td>
</tr>
<tr>
<td>Fayyad, U.M. Piatetskiy-Shapiro
G. Smith, P. Ramasasmy</td>
<td>Advances in Knowledge Discovery and Data Mining</td>
<td>USA</td>
<td class="derecho">2006</td>
</tr>
<tr>
<td>W. H. Inmon, Jhon Wiley and Sons</td>
<td>Data Warehouse Performance</td>
<td>USA</td>
<td class="derecho">1992</td>
</tr>
<tr>
<td>IBM Press 2001</td>
<td>IBM DB2 UDB Business Intelligence Tutorial</td>
<td>USA</td>
<td class="derecho">2001</td>
</tr>
<tr>
```

<td>W. H. Inmon, Jhon Wiley and Sons</td>

<td>Data Warehouse Performance</td>

<td>USA</td>

<td class="derecho">1992</td>

</tr>

<tr>

<td>Eric Thompson, Jhon Wileyand Sons</td>

<td>OLAP Solutios: Building Multidimentional Information Sysmens, Segunda Edición</td>

<td>USA</td>

<td class="derecho">1997</td>

</tr>

<tr>

<td>Alex Berson, Stephen J. Smith, Mc Graw Hill</td>

<td>Data Warehouse, Data Mining and OLAP</td>

<td>USA</td>

<td class="derecho">1997</td>

</tr>

<tr>

<td>Alan Simon, Jhon Wiley and Sons</td>

<td>Data Warehouse, Data Mining and OLAP</td>

<td>USA</td>

<td class="derecho">1997</td>

</tr>

<tr>

```
<td>Juan C. Trujilla, Manuel Palomar</td>
<td>Diseño de Almacenes de Datos</td>
<td>España</td>
<td class="derecho">2002</td>
</tr>
<tr>
<td>Colin J. White</td>
<td>IBM Enterprise Analytics for the Intelligent e-Business</td>
<td>USA</td>
<td class="derecho">2001</td>
</tr>
<tr>
<td>IBM Press</td>
<td>IBM DB2 Intelligent Miner for Data:
Utilización de Intelligent Miner for Data</td>
<td>USA</td>
<td class="derecho">2002</td>
</tr>
<tr>
<td>IBM Press</td>
<td>IBM DB2 Intelligent Miner Visualization:
Using the Intelligent Miner Visualizers</td>
<td>USA</td>
<td class="derecho">2002</td>
```

```
</tr>
<tr>
<td class="abajo">IBM Press</td>
<td class="abajo">IBM DB2 Intelligent Miner for Data :
Utilización de l Visualizador de Asociaciones  < / td>
<td class="abajo">USA</td>
<td class="derecho abajo">2002</td>
</tr>
</table>
</div>
</div><!–end middle–>
</div><!–end content–>
<div class="footer">
<table width="82%" height="35" border="0" align="center" cellpadding="2">
<tr>
<td width="20%"><div align="center" class="Estilo1"> Trabajo Final de Aplicaciones </A>
</div></td>
<td width="31%" class="Estilo1" align="center">Luis Alfonso Cutro</td>
<td width="29%" class="Estilo1" align="center">alfonsocutro@gmail.com</td>
</tr>
</table>
</div><!–end footer–>
</div><!–end all–>
</body>
```

```
</html>
```

conclu.jsp

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN" "http://www .
    w3 . org / TR / xhtml11/ DTD / xhtml11 .dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<%@ page
language="java"
contentType="text/html; charset=ISO-8859-1"
pageEncoding="ISO-8859-1"
session="true"
%>
<meta http-equiv="Content-Type"content="text/html;charset=ISO-8859-1"/ >
<meta name="GENERATOR" content="IBM WebSphere Studio" />
<meta http-equiv="Content-Style-Type" content="text/css" />
<link href="theme/eph.css" rel="stylesheet" type="text/css" />
<title>conclu.jsp</title>
</head>
<body>
<div id="all">
 <div class="header">
 <table width="900" border="0">
 <tr>
```

```
<td><img src="img/datawarehouse.JPG" width="110" height="140" ></td>
<td><h1>Minería de Datos Aplicada a la EPH</h1></td>
</tr>
</table>
</div><!–end header–>
<div class="content">
<div class="column left">
<fieldset class="buscador">
<% String usuario = (String)session.getAttribute("usuarioNombre");
String apellido = (String)session.getAttribute("usuarioApellido");
%>
<legend><%=usuario%> <%=apellido%></legend>
<table width="150">
<tr>
<td> </td>
</tr>
<tr>
<td align="center"><img src="img/idmsyst2.gif" /></td>
</tr>
<tr>
<td> </td>
</tr>
</tr>
</table>
```

```
</fieldset><!– fin del buscador–>
<fieldset class="buscador">
<legend>Minería de Datos Aplicada
<p> a la EPH</legend>
<div id="navcontainer">
<img src="img/minero.JPG" >
</div>
</fieldset>
<fieldset class="buscador">
<legend>Capitulos</legend>
<div id="navcontainer">
<ul id="navlist">
<li><a href="doc/tfcutro.pdf">Capitulo Nº1</a></li>
<li><a href="doc/Introducción a el DB2-CapNº2.pdf">
Capitulo Nº2</a> </li>
<li><a href="doc/Introducción a Intelligent Miner for Data-CapNº3.pdf">
Capitulo Nº3 </a> </li>
<li><a href="doc/Introducción al WebSphere Studio-CapNº4.pdf">
Capitulo Nº4 </a> </li>
<li><a href="doc/Creacion del Data Warehouse-CapNº5.pdf">
Capitulo Nº5 </a> </li>
<li><a href="doc/Extraccion de Conocimientos con IBM Intelligent Miner -
CapN º6 . pdf"> Capitulo Nº6 </a> </li>
<li><a href="doc/">Capitulo Nº7</a></li>
```

```
</ul>
</div>
</fieldset><!– fin del buscador–>
</div><!–end left–>
<div class=”column main2col”>
<div id=”header-menu”>
<ul id=”navi”>
<li><a href=”index.jsp”>Inicio</a></li>
<li><a href=”D:\IBM\wsad\wokspace\EPH\WebContent\html\resu.html”>
Resultados</a></li>
<li><a href=”EPH.jsp”>EPH</a></li>
<li><a href=”conclu.jsp”>Conclusiones</a></li>
<li><a href=”biblio.jsp”>Bibliografia</a></li>
</ul>
</div>
<!– ak ingresa toda la informacion que va variar (imagens, textos)–>
<div class=”contenido”>
<h2>Conclusión</h2>
<p class=”conclusion”>
Partiendo de los datos suministrados por el <span class= ” resalto”>
Instituto Nacional de Estadística y Censos (http://www.indec.mecon.ar / )</span>,
    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.
<br/>
```

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.

</p>

<!–end middle–>

<!–end content–>

<table width=”82%” height=”35” border=”0” align=”center” cellpadding=”2”>

<tr>

<td width=”20%”>

Trabajo Final de Aplicaciones</A>

</td>

<td width=”31%” class=”Estilo1” align=”center”>

Luis Alfonso Cutro</td>

<td width=”29%” class=”Estilo1” align=”center”>

alfonsocutro@gmail.com</td>

</tr>

</table>

<!–end footer–>

<!–end all–>

</body>

```
</html>
```

EPH.jsp

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN" "http://www.w3.org
    / TR / xhtml11 / DTD / xhtml11.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<%@ page
language="java"
contentType="text/html; charset=ISO-8859-1"
pageEncoding="ISO-8859-1"
session="true"
%>
<meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1" />
<meta name="GENERATOR" content="IBM WebSphere Studio" />
<meta http-equiv="Content-Style-Type" content="text/css" />
<link href="theme/eph.css" rel="stylesheet" type="text/css"/>
<title>EPH.jsp</title>
</head>
<body>
 <div id="all">
 <div class="header">
 <table width="900" border="0">
 <tr>
```

```
<td><img src="img/datawarehouse.JPG" width="110" height="140" ></td>
<td><h1>Minería de Datos Aplicada a la EPH</h1></td>
</tr>
</table>
</div><!–end header–>
<div class="content">
<div class="column left">
<fieldset class="buscador">
<% String usuario = (String)session.getAttribute("usuarioNombre");
String apellido = (String)session.getAttribute("usuarioApellido");
%>
<legend><%=usuario%> <%=apellido%></legend>
<table width="150">
<tr>
<td> </td>
</tr>
<tr>
<td align="center"><img src="img/idmsyst2.gif" /></td>
</tr>
<tr>
<td> </td>
</tr>
</tr>
</table>
```

```
</fieldset><!– fin del buscador–>

<fieldset class="buscador">

<legend>Minería de Datos Aplicada

<p> a la EPH</legend>

<div id="navcontainer">

<img src="img/minero.JPG" >

</div>

</fieldset>

<fieldset class="buscador">

<legend>Capitulos</legend>

<div id="navcontainer">

<ul id="navlist">

<li><a href="doc/tfcutro.pdf">Capitulo Nº1</a></li>

<li><a href="doc/Introducción a el DB2-CapNº2. pdf"> Capitulo Nº2</a>
</li>

<li><a href="doc/Introducción a Intelligent Miner for Data-CapNº3. pdf"> Capitulo Nº3 </a> </li>

<li><a href="doc/Introducción al WebSphere Studio-CapNº4. pdf"> Capitulo Nº4 </a> </li>

<li><a href="doc/Creacion del Data Warehouse-CapNº5. pdf"> Capitulo Nº5 </a> </li>

<li><a href="doc/Extraccion de Conocimientos con IBM Intelligent Miner - CapNº6 . pdf " > Capitulo Nº6 </a> </li>

<li><a href="doc/">Capitulo Nº7</a></li>

</ul>

</div>
```

```
</fieldset><!– fin del buscador–>
</div><!–end left–>
<div class=”column main2col”>
<div id=”header-menu”>
<ul id=”navi”>
<li><a href=”index.jsp”>Inicio</a></li>
<li><a href=”html/resu.html”>Resultados</a></li>
<li><a href=”EPH.jsp”>EPH</a></li>
<li><a href=”html/conclu.html”>Conclusiones</a></li>
<li><a href=”biblio.jsp”>Bibliografia</a></li>
</ul>
</div>
<!– ak ingresa toda la informacion que va variar (imagens, textos)–>
<br/>
<div class=”contenido”>
<h2>EPH</h2>
<p class=”conclusion”>
La misma contiene información de la nueva EPH , cuya muestra incluye 25.000
    familias de las 28 aglomerados urbanos y rurales
de la República Argentina con una frecuencia de cada tres meses.</p>
</div>
</div><!–end middle–>
</div><!–end content–>
<div class=”footer”>
<table width=”82%” height=”35” border=”0” align=”center” cellpadding=”2”>
```

```
<tr>
<td width="20%"><div align="center" class="Estilo1">
Trabajo Final de Aplicaciones</A>
</div></td>
<td width="31%" class="Estilo1" align="center">
Luis Alfonso Cutro</td>
<td width="29%" class="Estilo1" align="center">
alfonsocutro@gmail.com</td>
</tr>
</table>
</div><!–end footer–>
</div><!–end all–>
</body>
</html>
```

index.jsp

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN" "http://www.w3.org
    / TR / xhtml11 / DTD/xhtml11 . dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<%@ page
language="java"
contentType="text/html; charset=ISO-8859-1"
pageEncoding="ISO-8859-1"
```

```
%>
<meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1" />
<meta name="GENERATOR" content="IBM WebSphere Studio" />
<meta http-equiv="Content-Style-Type" content="text/css" />
<link href="theme/eph.css" rel="stylesheet" type="text/css" />
<title>index.jsp</title>
</head>
<div id="all">
 <div class="header">
 <table width="900" border="0">
 <tr>
 <td><img src="img/datawarehouse.jpg" width="110" height="140" ></td>
 <td><h1>Minería de Datos Aplicada a la EPH</h1></td>
 </tr>
 </table>
 </div><!–end header–>
 <div class="content">
 <div class="column left">
 <fieldset class="buscador">
 <form action="Login_Controller" method="get">
 <legend>Login</legend>
 <table>
 <tr>
 <td>Usuario</td>
```

```
</tr>
<tr>
<td><input type="text" name="usuario"/></td>
</tr>
<tr>
<td>
Clave
</td>
<tr>
<td><input name="clave" type="password" /></td>
</tr>
<tr>
<td><input type="submit" value="ingresar" align="right"/></td>
</tr>
<% String error = (String)request.getAttribute("error");
if((error !=null)&&(error.compareTo("si")==0))
{%>
<tr>
<td align="center"><span style="color:red; font-size: 10px;">
Error en el login</span></td>
</tr>
<%}%>
</table></form>
</fieldset><!– fin del buscador–>
```

```
<fieldset class="buscador">
<legend>Minería de Datos Aplicada<p> a la EPH</legend>
<div id="navcontainer">
<img src="img/minero.JPG" >
</div>
</fieldset><!– fin del buscador–>
<fieldset class="buscador">
<legend>Registrarse</legend>
<form method="post" action="Login_Controller">
<table>
<tr>
<td>Nombre</td>
</tr>
<tr>
<td><input type="text" name="nombre"/></td>
</tr>
<tr>
<td>
Apellido
</td>
<tr>
<td><input name="apellido" type="text" /></td>
</tr>
<tr>
```

```
<td>
Usuario
</td>
<tr>
<td><input name="usuario" type="text" /></td>
</tr>
<tr>
<td>
clave
</td>
<tr>
<td><input name="pass" type="password" /></td>
</tr>
<tr>
<td><input type="submit" value="Enviar" align="right"/></td>
</tr>
</table></form>
</fieldset><!– fin del buscador–>
</div><!–end left–>
<div class="column main2col">
<div id="header-menu">
<ul id="navi">
<li><a href="#" title = " Acceso invalido , debe registrarse antes de acceder
a alguna opción"> Inicio </a> </li>
```

```
<li><a href="#" title= " Acceso invalido , debe registrarse antes de acceder
a alguna opción">Resultados</a></li>
<li><a href="#" title= " Acceso invalido , debe registrarse antes de acceder
a alguna opción"> EPH </a> </li>
<li><a href="#" title= " Acceso invalido , debe registrarse antes de acceder
a alguna opción">Conclusiones</a></li>
<li><a href="#" title= " Acceso invalido , debe registrarse antes de acceder
a alguna opción">Bibliografia</a></li>
</ul>
</div>
<!– ak ingresa toda la informacion que va variar (imagens, textos)–>
<div class="contenido">
<center><img src="img/datamine5.jpg" /></center>
</div>
</div><!–end middle–>
</div><!–end content–>
<div class="footer">
<table width="82%" height="35" border="0" align="center" cellpadding="2">
<tr>
<td width="20%"><div align="center" class="Estilo1">
Trabajo Final de Aplicaciones</A>
</div></td>
<td width="31%" class="Estilo1" align="center">
Luis Alfonso Cutro</td>
```

```
 <td width="29%" class="Estilo1" align="center">
alfonsocutro@gmail.com</td>
 </tr>
</table>
 </div><!–end footer–>
 </div><!–end all–>
</body>
</html>
```

logueado.jsp

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN" "http://www.w3.org
    / TR / xhtml11 / DTD / xhtml11 . dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<%@ page
language="java"
contentType="text/html; charset=ISO-8859-1"
pageEncoding="ISO-8859-1"
session="true"
%>
<meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1" />
<meta name="GENERATOR" content="IBM WebSphere Studio" />
<meta http-equiv="Content-Style-Type" content="text/css" />
<link href="theme/eph.css" rel="stylesheet" type="text/css"/>
```

```
<title>logueado.jsp</title>
</head>
<body>
<div id="all">
<div class="header">
<table width="900" border="0">
<tr>
<td><img src="img/datawarehouse.jpg" width="110" height="140" ></td>
<td><h1>Minería de Datos Aplicada a la EPH</h1></td>
</tr>
</table>
</div><!–end header–>
<div class="content">
<div class="column left">
<fieldset class="buscador">
<% String usuario = (String)session.getAttribute("usuarioNombre");
String apellido = (String)session.getAttribute("usuarioApellido");
%>
<legend><%=usuario%> <%=apellido%></legend>
<table width="150">
<tr>
<td> </td>
</tr>
<tr>
```

```
<td align="center"><img src="img/idmsyst2.gif" /></td>
</tr>
<tr>
<td> </td>
</tr>
</tr>
</table>
</fieldset><!– fin del buscador–>
<fieldset class="buscador">
<legend>Minería de Datos Aplicada<p> a la EPH</legend>
<div id="navcontainer">
<img src="img/minero.JPG" >
</div>
</fieldset><!– fin del buscador–>
<fieldset class="buscador">
<legend>Capitulos</legend>
<div id="navcontainer">
<ul id="navlist">
<li><a href="doc/tfcutro.pdf"> Capitulo Nº1</a></li>
<li><a href="doc/Introducción a el DB2-CapNº2.pdf"> Capitulo Nº2</a></li>
<li><a href="doc/Introducción a Intelligent Miner for Data-CapNº3.pdf"> Capitulo Nº </a> </li>
<li><a href="doc/Introducción al WebSphere Studio-CapNº4.pdf"> Capitulo Nº4 </a> </li>
<li><a href="doc/Creacion del Data Warehouse-CapNº5.pdf"> Capitulo Nº5 </a> </li>
```

 Capitulo N^{o}6

Capitulo N^{o}7

</fieldset><!– fin del buscador–>

<!–end left–>

<div class=”column main2col”>

<ul id=”navi”>

Inicio

Resultados

EPH

Conclusiones

Bibliografia

<!– ak ingresa toda la informacion que va variar (imagens, textos)–>

<!–end middle–>

<!–end content–>

<table width=”82%” height=”35” border=”0” center”cellpadding=”2”>

```
<tr>
<td width="20%"><div align="center" class="Estilo1">
Trabajo Final de Aplicaciones</A>
</div></td>
<td width="31%" class="Estilo1" align="center">
Luis Alfonso Cutro</td>
<td widht="29%" class="Estilo1" align="center">
alfonsocutro@gmail.com</td>
</tr>
</table>
</div><!–end footer–>
</div><!–end all–>
</body>
</html>
```

registro.html

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http:
    / /www.w3.org /TR / xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1" />
<title>Minería de Datos Aplicada a la Encuesta Permanente de Hogares</title>
<link href="theme/eph.css" rel="stylesheet" type="text/css"/>
</head>
<body>
```

<table width="900" border="0">

<tr>

<td><img src="img/datawarehouse.jpg" width="110" height="140" ></td>

<td><h1>Minería de Datos Aplicada a la EPH</h1></td>

</tr>

</table>

<!–end header–>

<fieldset class="buscador">

<form action="Login_Controller" method="get">

<legend>Login</legend>

<table>

<tr>

<td>Usuario</td>

</tr>

<tr>

<td><input type="text" name="usuario"/></td>

</tr>

<tr>

<td>

Clave

```
</td>
<tr>
<td><input name="clave" type="password" /></td>
</tr>
<tr>
<td><input type="submit" value="Ingresar" align="right"/></td>
</tr>
</tr>
</table>
</form>
</fieldset><!– fin del buscador–>
<fieldset class="buscador">
<legend>Minería de Datos Aplicada<p> a la EPH</legend>
<div id="navcontainer">
<img src="img/minero.JPG" >
</div>
</fieldset><!– fin del buscador–>
<fieldset class="buscador">
<legend></legend>
<table>
<tr>
<td> </td>
</tr>
<tr>
```

<td>Se ha registrado correctamente. Ingrese con sus datos</td>

</tr>

<tr>

<td>

</td>

<tr>

<td> </td>

</tr>

<tr>

<td>

</td>

<tr>

<td> </td>

</tr>

<tr>

<td> </td>

</tr>

</tr>

</table>

</fieldset><!– fin del buscador–>

<!–end left–>

<div class=”column main2col”>

```
<div id="header-menu">
<ul id="navi">
<li><a href="#">Inicio</a></li>
<li><a href="#">Resultados</a></li>
<li><a href="#">EPH</a></li>
<li><a href="#">Conclusiones</a></li>
<li><a href="#">Bibliografia</a></li>
</ul>
</div>
<!– ak ingresa toda la informacion que va variar (imagens, textos)–>
<div class="contenido">
<center><img src="img/datamine5.jpg" /></center>
</div>
</div><!–end middle–>
</div><!–end content–>
<div class="footer">
<table width="82%" height="35" border="0" align="center" cellpadding="2">
<tr>
<td width="20%"><div align="center" class="Estilo1"> Trabajo Final de Aplicaciones </A>
</div></td>
<td width="31%" class="Estilo1" align="center">Luis Alfonso Cutro</td>
<td width="29%" class="Estilo1" align="center">alfonsocutro@gmail.com</td>
</tr>
</table>
```

```
</div><!–end footer–>
</div><!–end all–>
</body>
</html>
```

La totalidad del código desarrollado se encuentra en el DVD adjunto.

Capítulo 9

Conclusiones

Conclusiones Acerca de las Tecnologías y Software Utilizados

Se ha podido comprobar las grandes ventajas de la utilización de tecnologías y software de última generación, tanto de base de datos como de desarrollo de aplicaciones, que soportan sistemas distribuidos multiplataforma.

Esto ha resultado de gran utilidad al momento de desarrollar una aplicación con *WebSphere Application Developer v.5.0* y *DB2 UDB WorkGroup Server Edition v. 8.1*, *DB2 Intelligent Miner for Data v7.1*, bajo el sistema operativo *Windows XP*, utilizando *Java ESE 6.7.*

Se ha comprobado la facilidad del uso de los aplicativos mencionados, lo cual permitió actualizar los conocimientos en cuanto a las tecnologías que demanda el mercado actual.

Conclusiones Acerca de los Objetivos Propuestos

Respecto de los resultados obtenidos mediante la realización del presente trabajo, cabe mencionar que el proceso de extracción de conocimientos realizado sobre los datos provenientes del *"Instituto Nacional de Estadísticas y Censo (INDEC) - Encuesta Permanente de Hogares"*, revela una gran cantidad de información, la cual permite conocer a la población de la cuidad de Corrientes en un elevado nivel de detalle socio-demográfico y educacional.

Conclusiones Respecto del Proceso de Extracción del Conocimiento

El desarrollo de un *Almacén de Datos (Data Warehouse)* con su correspon-

diente esquema en estrella, permitió adquirir conocimientos adicionales sobre el diseño y utilización de esta tecnología.

Respecto de las fuentes de datos utilizadas, se puede destacar la excelente calidad y consistencia de los mismos, lo que agilizó notablemente su estudio ya que prácticamente no fue necesaria una etapa de depuración de datos.

Cabe destacar la eficiencia de los siguientes algoritmos aplicados:

- *"Clustering"*: permitió obtener un modelo con los datos socio demográficos y de educación de los individuos de la población estudiada.
- *"Árboles de decisión y clasificación"*: permitió la generación de reglas que ilustran las relaciones existentes entre los ingresos y el nivel socio demográfico, como también entre los ingresos y la educación de cada individuo.

Líneas Futuras de Acción

- Avanzar en la investigación mediante la aplicación de otras técnicas de minería de datos tales como *Redes Neuronales, Redes Bayesianas, etc.*
- Investigar acerca de la aparición de nuevas herramientas de *Inteligencia de Negocios (Business Intelligent)* y aplicarlas con el fin de obtener nuevos resultados y poder realizar comparaciones.
- Mejorar la aplicación generada agregando conceptos de *RIA (Rich Internet Applications).*

Bibliografía

[1] Jhon Wiley Alan Simon and Sons. *Data Warehouse, Data Mining and OLAP*. USA, 1997.

[2] Mc Graw Hill Alex Berson, Stephen J. Smith. *Data Warehouse, Data Mining and OLAP*. USA, 1997.

[3] Bart Jacob Carla Sadtler, John Ganci. *WebSphere Product Family Overview and Architecture*. IBM Press, USA, 2004.

[4] Heikki Mannila David Hand and Padhraic Smyth. *Principles of Data Mining*. MIT Press, Cambridge, MA, 2001.

[5] E. Marcos E.Bertino. *Concepts and Architectures*. Addison-Wesley, USA, 1993.

[6] E.Marcos E.Bertino. *Advanced Databases*. Artech House, USA, 2000.

[7] G.; Smith P.; Ramasasmy U. Fayyad, U.M.; Piatetskiy-Shapiro. *Advances in Knowledge Discovery and Data Mining*. AAAI Press / MIT Press, 2006.

[8] IBM Software Group. *Enterprise Data Warehousing whit DB2: The 10 Terabyte TPC-H Benchmark*. IBM Press, USA, 2003.

[9] José M. Guitiérrez. *Data Mining Extracción de Conocimiento en Grandes Bases de Datos*. España, 2001.

[10] Manuel Palomar Juan C. Trujilla. *Diseño de Almacenes de Datos*. España, 2002.

[11] Rolf Stadler JaapVerhees Peter Cabena, Pablo Hadjinian and Alessandro Zanasi. *Discovering Data Mining: From Concept to Implementation*. Prentice Hall, Upper Saddle River, NJ„ 1998.

[12] IBM Press. *IBM DB2 Intelligent Miner for Data: Utilización del Visualizador de Asociaciones*. IBM Press, USA, 1999.

[13] IBM Press. *IBM DB2 Warehouse Manager Guía de Instalación Version 7*. IBM Press, USA, 2001.

[14] IBM Press. *IBM DB2 Intelligent Miner for Data: Utilización de Intelligent Miner for Data*. IBM Press, USA, 2002.

[15] IBM Press. *IBM DB2 Intelligent Miner Visualization: Using the Intelligent Miner Visualizers*. IBM Press, USA, 2002.

[16] Ana M. Pérez Rubio. *Empleo, desempleo e informalidad: la composición del mercado región NEA. Una caracterización con los datos de la EPH, en Laboratorio - Estudios sobre Cambio Estructural y Desigualdad Social*. Facultad de Ciencias Sociales, UBA, Uriburu 950 6ž piso oficina 21, Cdad. de Buenos Aires, 2007.

[17] Rudyanto Linngar Saida Davies, Surech Amujuri. *WebSphere Business Integration Pub/Sub Solutions*. IBM Press, USA, 2004.

[18] Henry F.; Susarshan S. Silberschatz, Abraham; Korth. *Aprenda Servlets de Java como si estuviera en Segundo*. Editorial McGraw-Hill, USA, 1993.

[19] VV.AA. *Introducción a las Bases de Datos*. THOMSON PARANINFO, S.A., USA, 2005.

[20] Platesky Shapiro C. Matheus W. Frawley, G. *Knowledge Discovery in Database An Overview*. Al Magazine, 1992.

[21] Jhon Wiley W. H. Inmon and Sons. *Data Warehouse Performance*. John Wiley, USA, 1992.

[22] Jhon Wiley W. H. Inmon and Sons. *Building the Data Warehouse*. John Wiley, USA, 1996.

[23] Colin J. White. *IBM Enterprise Analytics for the Intelligent e-Business*. IBM Press, USA, 2001.

Índice de Materias