UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Facultad de Estudios Superiores Acatlán

TÉSIS

Que para obtener el Título de licenciado en Matemáticas Aplicadas y Computación

Presenta:

Hugo Tovar Núñez

"EL PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO EN BASE DE DATOS (KNOWLEDGE DISCOVERY IN DATABASES) UTILIZANDO SOFTWARE LIBRE PARA LA TOMA DE DECISIONES"

Asesor
Lic. Sergio Alejandro Matías Hernández

Abril 2010

Agradezco

A mis padres a quienes debo todo cuanto soy.

Muchas gracias mamá, papá siempre voy a estar agradecido por todo el amor que me han dado, el gran apoyo que siempre he tenido y los sacrificios que han hecho por mi y mis hermanos.

A mis hermanos Alicia, Pablo, Aristeo, Juan David, Roselia y Víctor quienes me regalaron una niñez inolvidable; han sido un ejemplo y se que siempre tendré a quien acudir en ustedes.

A mi asesor Alejandro Matías, así como a todos la plantilla de profesores de los cuales he tenido la oportunidad de aprender de su conocimiento y experiencia.

Además quiero expresar mi agradecimiento, pero sobretodo mi orgullo por ser parte de la UNAM, he tenido la fortuna de estar en áreas de gran competitividad y siempre he tratado de representar a la UNAM con dignidad y personalidad.

Muchas gracias...

Índice

Capítulo 1

Capítulo 2

Capítulo 3

Capítulo 4

Capítulo 5

Capítulo 6

Capítulo 7

Introducción

En los 90’s las aplicaciones y software desarrollados para la administración y explotación de los datos eran opción sólo para grandes organizaciones con los recursos para pagar por una licencia ó un sistema administrador de base de datos; actualmente con la apertura, el desarrollo tecnológico y la creciente comunidad de programadores, se tienen aplicaciones de uso libre con gran potencial que representan una alternativa para desarrollar sistemas similares o incluso mejores que los generados por una aplicación comercial, multiplataformas, integrados, de respuesta en tiempo real y capaces de satisfacer las necesidades de una organización a cualquier nivel.

El continuo desarrollo tecnológico y la apertura de múltiples canales de comunicación generan un ambiente con alcance global, por lo cual resulta una necesidad la implementación de sistemas de almacenamiento de datos suficientes, robustos y flexibles que generen datos e información diversa con procesos semiautomáticos y que exploten su potencial obteniendo conocimiento y con ello adoptar procesos de mejora continua.

La necesidad de información es requerida por igual tanto en organizaciones públicas como en privadas. El fundamento de este trabajo considera que la información es uno de los activos de mayor importancia para cualquier organización, y se establece a los datos como origen de información y conocimiento; aunque gran parte de las organizaciones presentan problemas por que tienen gran cantidad de datos históricos y generados en su operación día a día que simplemente se almacenan sin ninguna estructura, lo cual se traduce en datos redundantes, inconsistentes y sobre los cuales se toman decisiones.

El presente trabajo es una propuesta a organizaciones públicas ó personas que trabajan en estas y en su entorno, y trata de explicar de una manera clara, sencilla y objetiva la necesidad de planear, modelar, implementar y explotar un sistema de almacenamiento de datos masivo, con objetivos definidos, integral, estructurado, no volátil que permita extraer conocimiento útil y que soporte las decisiones en cualquier área de una organización.

En el Capítulo 1 se define el concepto de dato, información y la diferencia entre estos dos términos; el desarrollo de los diferentes sistemas de almacenamiento de datos que aparecen como concepto desde 1960 hasta hoy en que los principales receptáculos de datos son bases de datos objeto relacionales teniendo como principales características la administración lógica y física independiente así como la eliminación de redundancia. Además se presenta lo que es el lenguaje de consultas estructurado (SQL), sus operaciones así como principales características que lo hace hoy en día el lenguaje estándar para trabajar con bases de datos.

En el Capítulo 2 se trabajó con la definición de un sistema administrador de base de datos (DBMS), se mencionan los principales DBMS que existen en el mercado con distribución comercial y de uso libre, finalmente se hace énfasis de sus componentes, características principales y requerimientos para su instalación. En este Capítulo se define el sistema manejador de base de datos de uso libre PostgreSQL con el cual se va a diseñar e implementar una base de datos.

El Capítulo 3 aborda el concepto de Data Warehouse (DW), su arquitectura genérica que va desde los sistemas operacionales, transformación y carga de los datos a la aplicación, el acceso a ellos por parte de diferentes personas en una organización; sus características y entorno de negocio en una organización. Se plantean las ventajas y desventajas de su desarrollo principalmente desde el aspecto operacional.

En el Capítulo 4 se define el término de descubrimiento de conocimiento en bases de datos, por su acrónimo en inglés (KDD), (*Knowledge Discovery in Databases*), el cual basa su teoría en la extracción no trivial de conocimiento previamente desconocido y potencialmente útil, su proceso que considera la recopilación e integración de diferentes recursos de datos, su selección para modelarlos y una de sus fases de gran apertura y explotación comercial en estos días que es la minería de datos (DM) la cual, mediante técnicas estadísticas principalmente y con procesos semiautomáticos obtiene conocimiento que reside implícitamente en cantidades masivas de datos.

Otro concepto que tiene una relación directa con los datos, su almacenamiento y explotación es el CRM que se establece en el Capítulo 5. La gestión de la relación con el cliente por su acrónimo en inglés (CRM) (*Customer Relationship Managment*), es una estrategia o un conjunto de estrategias

que hoy en día resultan de gran rentabilidad para las organizaciones; se plantea su definición, características principales y su implementación teniendo como soporte el CRM analítico y procesos de descubrimiento de conocimiento, que basa su implementación en los diferentes canales de comunicación que integran el CRM operativo. Finalmente la integración de la interacción entre el cliente o usuario con la organización que es el objetivo del CRM colaborativo.

En el Capítulo 6 se presenta el concepto de Business Intelligence (BI), representando la etapa de integración de herramientas, procesos y tecnología para transformar datos en información, información en conocimiento y conocimiento como soporte de decisiones oportunas y consistentes, apoyado principalmente por herramientas de visualización y de un entorno grafico; las herramientas y procesos de business intelligence son hoy en día parte fundamental de sistemas administradores de bases de datos comerciales y un reto para los distintos sistemas de uso libre poder integrar herramientas de visualización y algoritmos estadísticos para su consulta en tiempo real.

Finalmente en el Capítulo 7 se plantea un caso de aplicación mediante el cual se desarrolla la teoría del trabajo. A través del DBMS PostgreSQL se diseña e implementa una base de datos que posteriormente es analizada con el objetivo de obtener información y conocimiento mediante técnicas de minería de datos con el software de uso libre WEKA. Estableciendo los resultados, de manera puntual se plantean las ventajas y desventajas de contar con información y conocimiento histórico para poder explicar lo que sucedió, entender lo que está sucediendo y anticipar lo que va a suceder.

Concluyendo con la necesidad de las diferentes áreas de aplicación y ámbitos sociales por establecer una cultura de mejora continua sustentada en el aprovechamiento optimo de sus recursos y en la explotación de uno de sus mayores activos que es la información

Capítulo 1

Resumen: En este como en cada Capítulo se presentan definiciones fundamentales para el desarrollo del trabajo como son: dato, información, características de la información, base de datos su evolución y características; modelo de datos y los diferentes modelos de datos, las reglas de Cood quien fundamentó el uso del modelo relacional en base al álgebra relacional; el uso de SQL como lenguaje estándar certificado para la operación de sistemas manejadores de base de datos.

Palabras clave: Dato, información, base de datos, SQL, modelo relacional

1 Datos, información su administración y almacenamiento

Actualmente ante el vertiginoso y continuo desarrollo de la tecnología y la informática se escucha el término “dato”, “información”, “almacenamiento”; como términos cotidianos de aplicación que usamos desde siempre y que en ocasiones usamos incorrectamente.

Si consideramos que se cuentan con registros de 3,500 A.C. donde los babilonios registraban intercambios de mercancía y los registraban en tablas de arcilla se podrían establecer gran número de definiciones de lo que es un dato, información y su almacenamiento.

A continuación se plantean diferentes definiciones bajo las cuales se trabajaran los conceptos a lo largo de todo el trabajo.

1.1 ¿Qué es un dato?

El dato (*del latín datum*), es una representación simbólica (numérica, alfabética, etc.), atributo o característica de una entidad. El dato no tiene valor semántico o sentido en sí mismo,

Pero convenientemente tratado o procesado se puede utilizar en la realización de cálculos o toma de decisiones.

"Los datos son hechos aislados y en bruto, los cuales situados en un contexto significativo mediante una o varias operaciones de procesamiento, permiten obtener deducciones relacionadas con la evolución e identificación de personas, eventos y objetos.", según: [Arnold; (1975)].

Considerando estas definiciones se trabajara el concepto de "dato", que como unidad simple e indivisible carece de sentido pero que relacionado con otros datos, se convierte en información. Entonces un dato puede ser una transacción bancaria, una compra, una serie de condiciones, una fecha, un objeto etc.; y que puede ser o no trivial para un individuo dependiendo de su contexto.

1.2 ¿Qué es información?

La información consiste en datos seleccionados y organizados respecto al usuario, problema, tiempo, lugar y función.

"La información es una acontecimiento o una serie de acontecimientos que llevan un mensaje, que al ser percibida por el receptor mediante alguno de sus sentidos amplia su conocimiento. Sólo el destinatario puede evaluar la significación y la utilidad de la información recibida.", según: [Arnold; (1975)].

Entonces se considerará el término información como un conjunto de datos organizados procesados que dan un conocimiento y/o descripción de un evento, fenómeno u objeto.

1.2.1 Diferencia entre dato e información

Los datos son la materia prima de la cual se deriva la información y la principal diferencia radica en su aplicación ya que la información brinda conocimiento de algún acontecimiento, fenómeno, objeto o razón mientras que los datos no son más que una representación simbólica de hechos aislados de poco valor por sí mismos.

Considerando la utilidad como principal diferencia se establece otro término de vital importancia entre estos conceptos que es el procesamiento de datos.

El procesamiento de datos es un término relativamente nuevo pero que se viene aplicando desde la segunda guerra mundial, y resulta ser un proceso de vital importancia ya que en el radica la información obtenida a partir de los datos.

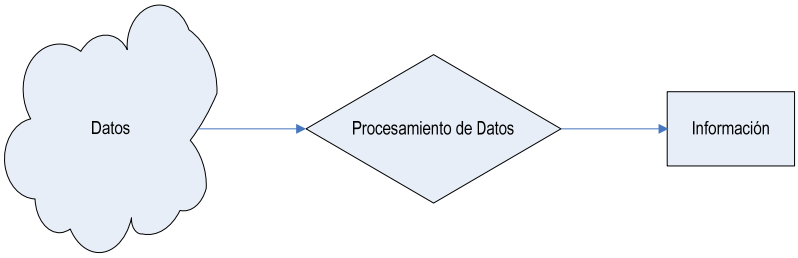


Fig.1 Obtención de información a partir del procesamiento de datos

1.2.2 Características del valor de la información

Considerando la información como un recurso imprescindible, resulta imperativo el considerar ciertos mecanismos y/o procesos que permitan establecer un ciclo para su procesamiento y así poder cumplir con todas y cada una de las características deseables.

Características del valor de la información:

- Accesible: Es la facilidad y rapidez con la que se obtiene la información requerida.
- Clara: Se refiere a la integridad y su entorno.
- Precisa: Que sea más exacta.
- Propia: Debe de haber relación entre el resultado y lo solicitado por el usuario.
- Oportuna: Menor duración del ciclo de procesamiento (Entrada- procesamiento-entrega usuario).
- Flexible: Adaptabilidad a todos los niveles de la organización (usuarios finales).
- Verificable: Que se pueda examinar y replicar la información.
- Imparcial: Que no sea volátil o vulnerable de acuerdo a los usuarios.
- Cuantificable: Todo dato procesado produce información

Considerando la información como uno de los recursos de mayor importancia en una organización; hoy en día es de vital importancia presupuestar su valor para llegar a obtener información con todas sus características deseables que sostengan cualquier decisión y que proyecten el comportamiento de una organización en un ambiente competitivo y abierto a cualquier tipo de cambio.

En décadas pasadas el costo de almacenamiento de la información resultaba costoso en todos los aspectos y se tenía que desechar ciclos, procesos o datos históricos que en su momento no resultaban importantes pero que hoy en día resultarían invaluables para hacer un análisis y tener argumentos para entender principalmente los cambios que se dieron a partir del desarrollo tecnológico y el crecimiento comercial.

1.3 Almacenamiento de datos

Como antecedentes principales antes de la era de las computadoras electrónicas como las que se conocen hoy en día se encuentra la era mecánica y la electromecánica; la construcción antes del siglo XX de la máquina aritmética de Pascal que realizaba las cuatro operaciones aritméticas y los

diseños de las maquinas de Babbage, que indujeron el planteamiento y desarrollo de los lenguajes de alto nivel que son el punto de partida para diseñar, programar, y administrar sistemas de información.

La evolución y desarrollo del almacenamiento de datos se tiene que plantear de manera formal desde que aparece la primera generación de computadoras (1946-1955), que es después de la Segunda Guerra Mundial.

El término base de datos comenzó a mencionarse en 1960 cuando en esa época sólo se manejaban archivos y conjuntos de datos carentes de cualquier relación y organizados estos de manera secuencial, se creaban archivos de copia del mismo archivo por que se guardaban generaciones anteriores, lo que traía consigo redundancia y una dependencia de datos total. Obviamente existían muchos problemas con esta manera de almacenar la información, el costo de almacenaje de un solo *bit* era muy caro y tener esa información replicada en varios archivos era un gran costo tanto operativo como funcional ya que si se quería hacer una reorganización de los datos se tenía que cambiar toda la programación, dicho de otra manera si se quería agregar una entidad o alguna característica a la aplicación que accedía al sistema de almacenamiento de datos se tenia que rehacer por completo el programa no había una independencia lógica de la estructura respecto a los datos.

Hacia 1970 se identifico la naturaleza de cambio de los diferentes sistemas y el objetivo era proteger el trabajo intelectual de las estructuras de los sistemas de almacenamiento por lo que el principal objetivo era separar la estructura lógica de la física es decir, se pretendía diseñar sistemas de almacenamiento de datos en los que se pudiera cambiar los datos de manera física sin que se alterara su estructura lógica.

En esta etapa ya se contaba con un acceso a archivos no únicamente secuenciales sino también directos o indexados y el gran avance que se genero fue que ya se podían cambiar las unidades de almacenamiento sin tener que modificar o rehacer los programas de aplicación.

De 1970 a 1980 ya no únicamente era una necesidad el acceso a archivos; el crecimiento exponencial en el volumen de los datos así como los grandes cambios comerciales trajeron consigo una reestructuración en el entorno de un sistema de información, los datos ya almacenados necesitaban nuevos campos y nuevas relaciones. Se logro tener acceso a datos de múltiples maneras lógicas y eliminar gradualmente la redundancia, se utilizaban formas de organización de datos muy complejas ajenas en todo momento a los programas de aplicación.

1.4 Definición de base de datos

"Una base de datos se define como una colección de datos interrelacionados almacenados en conjunto sin redundancias perjudiciales e innecesarias; su finalidad es la de servir a una aplicación o más, de la mejor manera posible; los datos se almacenan de modo que resulten independientes de los programas que los usan; se emplean métodos bien determinados para incluir datos nuevos y para modificar o extraer los datos almacenados." según: [Martín; (1975)]

Una definición generalizada por la comunidad de programadores es:

"Colección de datos integrados, con redundancia controlada y con una estructura que refleje las interrelaciones y restricciones existentes en el mundo real; los datos que han de ser compartidos por diferentes usuarios y aplicaciones, deben mantenerse independientes de éstas, y su definición y descripción, únicas para cada tipo de datos, han de estar almacenadas junto con los mismos.

Los procedimientos de actualización y recuperación, comunes, y bien determinados, habrán de ser capaces de conservar la integridad, seguridad y confidencialidad del conjunto de datos."

1.4.1 Evolución de base de datos

A continuación se presenta una breve descripción del desarrollo cronológico de los sistemas manejadores de bases de datos y sus principales características.

1.4.1.1 Sistemas manejadores de archivos 50´s

El acceso a los datos era de modo secuencial simple; se guardaban copias de generaciones anteriores lo que generaba alto nivel de redundancia entre estos, la estructura física era esencialmente igual a la lógica lo cual ocasionaba grandes costos ya que si se cambiaba la estructura física de los datos se tenia que volver a escribir los programas de la aplicación, compilarse y probarse lo cual era excesivamente caro además de que no mantenía un crecimiento intelectual de las aplicaciones realizadas.

1.4.1.2 Surgimiento de bases de datos jerárquicas 60´s

El acceso a los datos es secuencial y el gran avance se da en el acceso a registros al azar, se trabaja con hacer independiente la estructura lógica de la física, también se logra tener algunos recursos de seguridad aunque no del todo fiables, y la principal característica de esta generación fue la estructura jerárquica basada en uno nodo padre que tiene una relación con un hijo; aunque todavía existía mucha redundancia en los datos.

1.4.1.3 Surgimiento de bases de datos de red 70´s

Finalmente se logra la independencia de la organización física de los datos de los programas que lo usan lo cual deriva en diferentes maneras de acceso a datos, el cambio en unidades de almacenamiento de datos sin tener que modificar los programas y sobre todo mantener un control sobre la redundancia lo cual también ayuda a la conservación de los datos y su integridad; esta independencia de los datos de su estructura física y lógica es a partir de la integración en el esquema de una base de datos de un módulo de administración de datos (software), que permitía el acceso a los datos de diferentes maneras según los requisitos de la aplicación.

1.4.1.4 Surgimiento y auge de bases de datos relacionales 80´s 90´s

Se cuenta con un panorama global de la estructura lógica; también se crean medios para que un administrador de base de datos tenga control, seguridad y asegure que la organización de estos sea siempre la mejor para los usuarios en general. En esta etapa del desarrollo en base deDatos ya es posible una migración de datos sin que se pierdan históricos o se vean afectados en cuanto su integridad y seguridad.

1.4.2 Características de una base de datos

Respecto a su definición y tratando de preservar las propiedades de los datos y la información; una base de datos cuenta con cuatro características principales:

- Redundancia
- Consistencia
- Integridad
- Seguridad

1.4.2.1 Redundancia

La redundancia de datos se refiere, a la existencia de información repetida o duplicada en diferentes tablas dentro de una base de datos, y esto se origina principalmente en los procesos de carga ya que se pueden almacenar datos con diferentes aplicaciones y si no se controla detalladamente este o estos procesos se puede estar almacenando información repetida la cual aumenta los costos de almacenamiento y acceso a los datos, además de que puede originar la inconsistencia de los datos, es decir diversas copias de un dato que no concuerdan entre sí.

Dentro de una base de datos relacional la redundancia debe ser mínima y controlada. En ocasiones existirán motivos válidos de negocios o técnicos para mantener varias copias de los mismos datos almacenados en diferentes tablas.

1.4.2.2 Consistencia

Generalmente los problemas de consistencia son ocasionados por la redundancia y un mal diseño a nivel lógico; y es que al almacenar información en diversas tablas de una base de datos de manera innecesaria resulta probable que a la hora de efectuar modificaciones en ellas no se afecten todas las instancias y se generen incongruencias entre los datos obteniendo datos inconsistentes.

1.4.2.3 Integridad

La integridad en una base de datos se refiere principalmente a los valores que poseen los datos y su relación con otras tablas y su entorno. Es recomendable para lograr la integridad:

- Establecer de manera global reglas de negocio que consideren el contexto y la estructura lógica de la base de datos.
- El mantenimiento de una redundancia mínima y siempre controlada.
- La creación de reglas de validación durante la inserción y edición de datos.

1.4.2.4 Seguridad

La seguridad de una base de datos se refiere principalmente al control de acceso, modificación y definición, tanto de los datos como de la estructura de la base de datos por parte de los diferentes usuarios a la misma.

Se debe de contar con normas y políticas de acceso y uso en los diferentes niveles que garanticen integridad, seguridad y confidencialidad del conjunto de datos.

Respecto a la seguridad de puede establecer:

- Seguridad de Objetos: Se refiere a los permisos de los diferentes usuarios para poder hacer uso de tablas, procedimientos almacenados*, triggers*, etc.
- Seguridad de operaciones: Aquí se manejan permisos para poder modificar (insertar, borrar, actualizar) la base de datos.

1.4.3 Modelo definición

Es una representación de la realidad que contiene las características generales y el sentido abstracto de algo que se pretende representar mediante datos y sus relaciones. En base de datos, esta representación se elabora de forma gráfica.

1.4.3.1 ¿Qué es un modelo de datos?

Es una colección de herramientas conceptuales para describir los datos, las relaciones que existen entre ellos, semántica asociada a los datos y restricciones de consistencia.

Los modelos de datos se dividen en tres grupos:

- Modelos lógicos basados en registros.
- Modelos lógicos basados en objetos.
- Modelos físicos de datos.

Los modelos de datos se usan para describir datos en los niveles conceptual y físico, es decir, con un modelo representamos los datos de tal forma como nosotros los captamos en el mundo real, tienen una capacidad de estructura flexible y permiten especificar restricciones de datos explícitamente. Existen diferentes modelos de este tipo, pero el más utilizado por su sencillez y eficiencia es el modelo entidad-relación.

Denominado por sus siglas como: *E-R*; este modelo representa a la realidad del entorno con el que se trabaja a través de entidades, que son objetos que se distinguen de otros por sus características pero que se relacionan por reglas, asociaciones o condiciones.

1.4.3.2 Modelo jerárquico

En este tipo de modelos la organización se establece en forma de árbol, donde la raíz es un nodo ficticio y la relación entre datos es por medio de registros y sus ligas. La diferencia con otros modelos radica en que están organizados por conjuntos de árboles en lugar de gráficas.

1.4.3.3 Modelo de red

Una base de datos de red como su nombre lo índica, esta formado por una colección de registros, los cuales están conectados entre sí por medio de enlaces; el enlace es la asociación entre dos registros exclusivamente, así que podemos verla como una relación estrictamente binaria.

1.4.3.4 Modelo relacional

En este modelo se representan los datos y las relaciones entre estos, a través de una colección de tablas, en las cuales los renglones (*tuplas*) equivalen a cada uno de los registros que contendrá la base de datos y las columnas corresponden a las características (atributos) de cada registro localizado en la *tupla*.

En el modelo relacional el único elemento de representación es la tabla.

1.4.4 Evolución del modelo relacional

PERÍODO	EVENTOS
1968 - 1970	Surge el modelo
1970 …	Desarrollos teóricos
1970 - 1978	Prototipos (Ingres, Sistema R de IBM, etc.)
1978	QBE (Query By Example), de IBM
1979	ORACLE (1er SGDB)
1980	Ingres
1981	SQL
1982	DB2
1986	SQL/ANSI
1987	SQL/ISO
1989	SQL Addendum
1989	Manifiesto de los SGBO
1990	Modelo relacional versión 2
1990	Manifiesto de los SGBO -3G
1992	SQL 92
1995	3er manifiesto
1999	SQL 1999
2003	SQL 2003

1.4.5 Objetivos del modelo relacional

A finales de los años 60´s el Dr. Edgar Frank Codd introdujo la teoría matemática de las relaciones en el campo de las bases de datos.

El modelo relacional fue propuesto por Codd en su artículo titulado "*A relational model of data for large shared of data banks*" (Codd, 1970).

Y sus objetivos primordiales eran:

- Independencia física: La manera en que se almacenan los datos no influye en su manipulación lógica y por tanto, los usuarios que acceden a esos datos no tienen que modificar sus programas por cambios en el almacenamiento físico.
- Independencia lógica: El añadir, eliminar o modificar objetos de la base de datos no repercute en los programas y/o usuarios que están accediendo a subconjuntos parciales de los mismos (vistas).
- Flexibilidad: En el sentido de poder presentar a cada usuario los datos de la forma en que éste prefiera.
- Uniformidad: Las estructuras lógicas de los datos presentan un aspecto uniforme, lo que facilita la concepción y manipulación de la base de datos por parte de los usuarios.
- Sencillez: Las características anteriores, así como unos lenguajes de usuario muy sencillos, producen como resultado que el modelo de datos relacional sea fácil de comprender y de utilizar por parte del usuario final.

1.4.6 Reglas de Codd

En 1985 el Dr. Edgar Frank Codd publicó 12 reglas para evaluar si un DBMS (*Data Base Management System*) puede considerarse un RDBMS (*Relational DataBase Management System*),

O dicho más concisamente, si un sistema de bases de datos puede considerarse o no relacional.

0.- El sistema debe ser relacional

Base de datos y administrador de sistema. Ese sistema debe utilizar sus facilidades relacionales (exclusivamente) para manejar la base de datos.

1.- Representación de la información

Toda información almacenada en una base de datos relacional debe representarse explícitamente a nivel lógico, y de manera única, por medio de valores en tablas. Siendo éste el principio básico del modelo relacional.

2.- Acceso garantizado

Se garantiza que todos y cada uno de los datos (valor atómico) en una base de datos relacional pueden ser leídos recurriendo a una combinación del nombre de la tabla, valor de la llave primaria y nombre de la columna.

3.- El manejo sistemático de los valores nulos

En un RDBMS totalmente relacional se soportan los valores nulos (que son distintos de una cadena de caracteres vacía o de una cadena con caracteres en blanco o de cero o cualquier otro número), para representar información faltante o no aplicable de una forma consistente, independientemente del tipo de dato.

4.- Catálogo dinámico en línea basado en un modelo relacional

La descripción de la base de datos se representa en el nivel lógico de la misma forma que los datos ordinarios, de tal manera que los usuarios autorizados puedan aplicar el mismo lenguaje relacional para consultarla, que aquél que emplean para con sus datos habituales.

5.- Regla del sub-lenguaje de dato completo

Se debe contar con un sub-lenguaje que contemple la definición de datos, la definición de vistas, la manipulación de datos, las restricciones de integridad, la autorización, el inicio y fin de una transacción.

6.- Regla de actualización de vistas

Todas las vistas que teóricamente sean actualizables deberán ser actualizadas por medio del sistema.

7.- Inserción, actualización y eliminación de alto nivel

La posibilidad de manejar una relación base o una relación derivada como un sólo operador se aplica a la lectura, inserción, modificación y eliminación de datos.

8.- Independencia física de los datos

Los programas de aplicación y la actividad en terminales no deberán ser afectados por cambios en el almacenamiento físico de los datos o en el método de acceso.

9.- Independencia lógica de los datos

Los programas de aplicación y la actividad en terminales no deberán ser afectados por cambios de cualquier tipo que preserven la información y que teóricamente permitan la no afectación en las tablas base.

10.- Independencia de la integridad

Las restricciones de integridad de una base de datos deberán poder definirse en el mismo sublenguaje de datos relacional y deberán almacenarse en el catálogo, no en los programas de aplicación.

11.- Independencia de la distribución

Un DBMS relacional tiene independencia de distribución.

12.- Regla de la no subversión

Si un sistema relacional tiene un lenguaje de bajo nivel (un solo registro cada vez), ese bajo nivel no puede ser utilizado para suprimir las reglas de integridad y las restricciones expresadas en el lenguaje relacional de nivel superior (múltiples registros a la vez).

1.5 Lenguaje Estructurado de Consultas (SQL)

Fue definido inicialmente en la década de los 70´s por Donald Chamberlain que trabajaba para IBM con el objetivo de encontrar una herramienta de consulta para explotar todas las ventajas del modelo relacional de las bases de datos. Tras varios prototipos el potencial se vio realizado tras su normalización por parte de ANSI (*American National Standard Institute*) e ISO (*International Organization for Standardization*) lo cual llevo a que cualquier DBMS estuviera orientado a SQL o lo tuviera inmerso.

En los últimos años numerosas firmas comerciales han generado productos orientados a SQL, lo que hace que cuente con un entorno de trabajo sencillo, dinámico y flexible por lo cual se considera un lenguaje de cuarta generación 4GL, donde el usuario define lo que se debe hacer más no como se debe hacer.

1.5.1 Álgebra relacional

El álgebra relacional consiste en una colección de operaciones sobre relaciones donde cada operación toma una o más relaciones como su operando y produce otra relación como su resultado.

Dado que el resultado de una operación del álgebra relacional es una relación, ésta a su vez puede ser sujeto de posteriores operaciones algebraicas. En base de datos el álgebra relacional se fundamenta principalmente en:

- La teoría de conjuntos, relaciones y en el álgebra de conjuntos.
- Adicionalmente al conjunto básico de operadores como: unión, diferencia, producto cartesiano e intersección; incorpora operadores específicos de base de datos tales como proyección, selección y *join*.

1.5.2 Unión, intersección, diferencia, producto cartesiano, proyección, selección, join

- Unión: La unión de la relación de dos conjuntos denotada como R ∪ S es la relación de *tuplas* que están en R ó en S ó ambas; sólo se puede aplicar el operador al conjunto de *tuplas* que cuenten con el mismo esquema (mismos campos), entonces todas las *tuplas* del resultado de la unión de la relación tienen los mismos atributos; así como su orden.
- Intersección: Es la relación que contiene el conjunto de *tuplas* que están en R y en S y se denota como R ∩ S. Esto es, construye una relación formada por aquellas *tuplas* que aparezcan en las dos relaciones especificadas.
- Diferencia: La diferencia algebraica de un par de relaciones R Y S se denota como R – S y es el conjunto de *tuplas* en R pero no en S y al igual que en la unión sólo se puede aplicar el operador diferencia si cuentan con el mismo esquema respecto a sus atributos.
- Producto Cartesiano: Obtiene todas las *tuplas* que se construyen concatenando cada tupla de R con otra de S y se denota R X S. En este caso los atributos de R y S no tienen que ser los mismos. A partir de dos relaciones especificadas, construye una relación que contiene todas las combinaciones posibles de *tuplas*, una de cada una de las dos, esto es, los pares ordenados.
- Proyección: La proyección selecciona y genera un subconjunto con los atributos indicados de una tabla. También es conocida como operación vertical.
- Selección: La selección toma y genera un subconjunto con los renglones indicados de una tabla. También es conocida como operación horizontal.
- *Join*: La operación *join* es en esencia un producto cartesiano, donde se seleccionan las columnas que satisfagan las condiciones indicadas. Es la operación más común en las bases de datos relacionales.

Ejemplos de operaciones algebraicas, sean las relaciones R y S

D	E	F
b	g	a
d	a	f

Relación S

A	B	C
a	b	c
d	a	f
c	b	d

Relación R

R U S

a	b	c
d	a	f
c	b	d
b	g	a

R - S

a	b	c
d	a	f
c	b	d
b	g	a

R ∩ S

d	a	f

R X S

A	B	C	D	E	F
a	b	c	b	g	a
a	b	c	d	a	f
d	a	f	b	g	a
d	a	f	d	a	f
c	b	d	b	g	a
c	b	d	d	a	f

Proyección
A, C (R)

A		C
a		c
d		f
c		d

Selección
B=b, (R)

A	B	C
a	b	c
c	b	d

1.5.3 Concepto de llave primaria y llave foránea

Una llave o clave en base de datos es un campo (o campos), o atributos, que tiene un valor único para cada *tupla* de la tabla; es decir el campo a partir del cual se pueden inferir los demás campos de la tabla; por lo que cada *tupla* debe estar asociada con una llave que permita su identificación. Hoy en día cualquier DBMS utiliza su definición para garantizar la integridad de los datos, pero su mal uso puede causar redundancia no controlada.

Existen dos tipos de llaves:

- Llave primaria: Es aquel atributo que identifica de manera única a una *tupla*; no puede contener valores nulos (NULL) esto quiere decir que no puede haber *tuplas* repetidas respecto a su llave primaria.
- Llave secundaria (foránea o externa): Es una llave primaria en otra relación, es decir es una llave que esta siendo compartida por dos tablas que mantienen una relación.

1.6 Normalización

El proceso de normalización de una base de datos consiste en una serie de restricciones que se rigen a partir de la estructura del modelo relacional de la base de datos con el objetivo de evitar redundancia no controlada y evitar anomalías al realizar una inserción, eliminación o actualización de los datos.

Las formas de normalización fueron propuestas por Codd entre 1971 y 1972 y con el tiempo han surgido varias formas de normalización que complementan a las establecidas por Codd; en este caso se señalan las tres primeras formas normales y sus principales características.

- Primera forma normal (1NF): Dependencia funcional de los atributos que no son clave respecto a la llave primaria. Una relación está en primera forma normal si, y sólo si, todos los dominios de la misma contienen valores atómicos, es decir, no hay grupos repetitivos. Si se ve la relación gráficamente como una tabla, estará en 1FN si tiene un solo valor en la intersección de cada fila con cada columna.
- Segunda Forma Normal (2NF): Dependencia funcional completa de los atributos que no son clave respecto a la llave primaria. Una relación está en segunda forma normal si, y sólo si, está en 1FN y, además, cada atributo que no está en la llave primaria es completamente dependiente de la llave primaria.
- Tercera forma normal (3NF): Ninguna dependencia transitiva de los atributos que no son clave.

Una relación está en tercera forma normal si, y sólo si, está en 2FN y, además, cada atributo que no está en la llave primaria no depende transitivamente de la llave primaria. La dependencia es transitiva si existen las dependencias siendo atributos o conjuntos de atributos de una misma relación. Las principales ventajas de la normalización son:

- Evita anomalías en inserciones, modificaciones y borrados.

Mejora la independencia de datos

Capítulo 2

Resumen: Se presenta la definición de sistema administrador de base de datos (DBMS), sus componentes y esquema de seguridad genérico; la trascendencia del lenguaje SQL y su evolución para considerarse como el lenguaje estándar de los DBMS. Finalmente se presentan los principales DBMS comerciales y libres con sus características principales; focalizando en el proyecto GNU que ha tenido trascendencia desde 1983 y que hoy en día cuenta con aplicaciones tan robustas y potentes como un software comercial.

2. Administradores y manejadores de bases de datos en el mercado

Desde finales de los años 60´s se presenta una competencia estrecha entre las distintas compañías por mantener la vanguardia en cuanto a sistemas capaces de administrar y procesar datos; como se presentó en el Capítulo 1 la evolución de las bases de datos presenta un antes y un después con la propuesta de Cood, de representar a través del modelo relacional los datos, su almacenamiento y procesamiento. IBM siempre buscó con diferentes prototipos pero fue ORACLE en 1979 que logra desarrollar el primer sistema administrador de base de datos (DBMS), que cumplía con todas las tareas que ha de ejecutar un sistema administrador de base de datos.

2.1 Descripción y definición de un administrador y manejador de bases de datos relacional.

Ya se definió el concepto de base de datos, la evolución, sus características que hicieron que cada sistema tuviera una forma diferente de almacenamiento, integridad, seguridad y capacidad para poder modificar extraer o eliminar algunos de esos datos ya almacenados.

Para poder lograr un control estructurado del almacenamiento físico de los datos, independiente de su estructura lógica, que permita modificar, extraer y garantizar su integridad son necesarios una serie de procedimientos que ejecuten está labor; a esta serie de procedimientos (software), se le conoce como sistema de gestión de base de datos (SGBD).

Entonces un SGBD es una colección de programas que permiten la creación, manipulación, modificación y control de acceso a una o varias bases de datos.

Principales funciones de un SGBD.

- Crear y organizar bases de datos
- Definición de los datos a los distintos niveles de abstracción (físico, lógico y externo).
- Manipulación de los datos; es decir la inserción, modificación, borrado y acceso vía consultas a los mismos.
- Mantener integridad de la base de datos, conservando la estructura lógica independientemente de los cambios físicos, sus valores y sus relaciones.
- Control de privacidad y seguridad de los datos en la base de datos.

2.1.1 SQL su desarrollo

La historia de SQL (*Structured Query Lenguaje*), data desde 1974 cuando Donald Chamberlain que trabajaba para IBM propuso un lenguaje que trabajaba con las características de las bases de datos apegadas al modelo relacional, este prototipo se llamaba SEQUEL (*Structured, English Query Lenguaje*), que por motivos legales quedo su nombre en SQL, entonces IBM desde 1981 comenzó a trabajar con bases de datos relacionales y en 1983 introdujo al mercado comercial DB2. En los 70´s compañías como ORACLE, SYBASE comenzaron a trabajar con productos relaciones todos ellos basados en SQL lo cual hace que SQL se convierta en el estándar respecto a base de datos relacionales se refiere.

A continuación se presenta una breve reseña cronológica de la evolución y transformación de SQL, considerando los cambios más importantes y sus características principales.

2.1.1.1 ANSI SQL 89

En 1986 ANSI *(American Nacional Standar Institute)*, adopto el lenguaje SQL como estándar para los lenguajes relacionales y en 1987 se convierte en estándar ISO (*International Organization for Standardization*), el cual se llamaba SQL/86; pero en 1989 ANSI define SQL 89 basado en SQL/86 con una serie de mejoras (definición de llaves primarias, integridad de los datos)

Las características más importantes de ANSI SQL 89 son:

- Integridad referencial.
- Se establece que el lenguaje SQL está compuesto por comandos cláusulas, operadores y funciones de agregado, los cuales se combinan para definir y manipular una Base de datos.
- Se establecen los componentes de un DBMS *Data Definition Lenguaje* (DDL), *Data Modification Lenguaje* (DML), *Data Control Lenguaje* (DCL); así como la sintaxis relacionada a cada uno de ellos.
- Establecimiento de las cláusulas del comando select que son: From, Where, Group By, Having, Order By.
- Definición de operadores de comparación.
- Definición de operadores lógicos AND, OR Y NOT.
- Definición de operadores de comparación.
- Se determinan las funciones de agregado: SUM, COUNT, AVG, MAX, MIN.

2.1.1.2 ANSI SQL 92

Toma todas las características de SQL 89 además de incorporar mejoras sustanciales respecto al tipo de datos y procedimientos para definir tablas, esquemas y vistas.

Las características más importantes de ANSI SQL 92 son:

- Definición de esquemas: Es una cláusula que permite al dueño del esquema dar permisos para la manipulación de los datos del mismo esquema.
- Nuevos tipos de datos: Se establecen nuevos tipos de datos como numéricos exactos (*Small*, decimal, *integer*, etc.), numéricos aproximados (real, doublé precisión, *float*, etc.), cadenas de bits (*bit(n), bit varying* (n).), fechas y horas (*date*(fecha), *time*(hora)), intervalos (*year-month* (año- mes), *day time* (día - hora)).
- Definición de dominios: Se refiere a la determinación de algún tipo de dato en específico por medio de cláusulas que nosotros mismos creamos; dicho de otra manera son datos que nosotros generamos sus propiedades y tipo de dato.
- Definición de tablas: Se refiere a que se puede crear una tabla considerando un nombre para tenerla en la base de datos, los tipos de datos y restricciones que deben de cumplir.
- Menciona las consideraciones pertinentes para realizar consultas y subconsultas.

2.1.1.3 ANSI SQL 99

Toma todas las características de SQL 89 Y 92. Se caracteriza como “SQL orientado a objetos” y es la base de sistemas de manejo de base de datos orientados a objetos (ORACLE, *Informix Universal Server* entre otros);

Contenidos de SQL 99

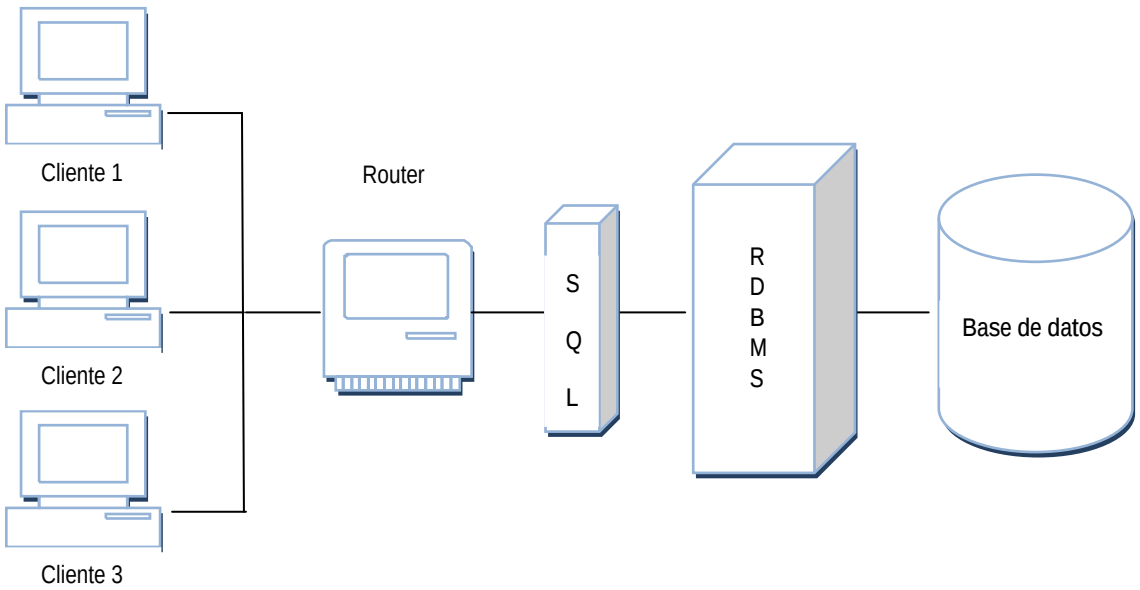
- Nuevos Tipos de Datos: CLOB (*Character Large Object*) funciona como una cadena de caracteres, pero tiene restricciones que impiden su uso como PRYMARY KEY, BLOB

(*Bynary Large Object*) cuenta con restricciones similares además de que no puede ser usada en cláusulas Group by y Order by.

- Otro tipo de dato nuevo que trajo esta versión es el bolean que permite registrar valores de verdad como "falso" y "desconocido"; con esto complejas combinaciones de enunciados pueden ser ahora expresadas de una manera más sencilla que antes.
- Nuevos predicados: LIKE, SIMILAR y DISTINCT
- Presenta dos nuevos operadores de totales EVER Y ANY
- Incorpora generadores de tipo de dato: ARRAY, REF Y ROW
- Incluye dos operadores EXISTS Y NOT EXISTS.

2.1.2 Arquitectura cliente servidor

Desde su explotación comercial, las bases de datos se han utilizado en sistemas que se ajustan a una arquitectura conocida como cliente-servidor. En ella los datos residen en un ordenador que actúa como servidor, ejecutando el software que denominamos servidor de datos. Los usuarios, desde ordenadores remotos, se sirven de un software cliente para comunicarse con el servidor de datos, ese software cliente es específico para cada servidor de datos existente y su respectivo DBMS.



2.1.3 Esquemas de seguridad en los manejadores de bases de datos relacionales

La seguridad de los datos se refiere a la protección de estos contra el acceso a personas o programas (software), no autorizados contra su indebida destrucción o algún tipo de alteración que afecte de cualquier manera sus características y entorno de almacenamiento.

Existen tres características respecto a seguridad se refiere que debe mantener una base de datos y son:

- Seguridad
- Confidencialidad
- Disponibilidad de los datos.

Por lo cual deben de existir políticas de acceso a los datos referentes a cada organización pero además en el entorno físico debe de haber una estructura u esquema bien definido que garantice en todo momento seguridad, confidencialidad y disponibilidad de los datos.

En un sistema manejador de base de datos (DBMS), se cuenta con un esquema "multicapas" el cual considera tres instancias de manera jerárquica que son: servidor, base de datos, objetos y datos; es decir el usuario final debe tener una cuenta o acceso valido en cuanto a la capa del servidor (seguridad a nivel servidor); el usuario final debe tener una cuenta de usuario en la capa de la base de datos (seguridad a nivel base de datos) y finalmente el usuario final deberá tener una cuenta para tener permisos sobre objetos y comandos lo cual se representa en la Fig. 2.1.

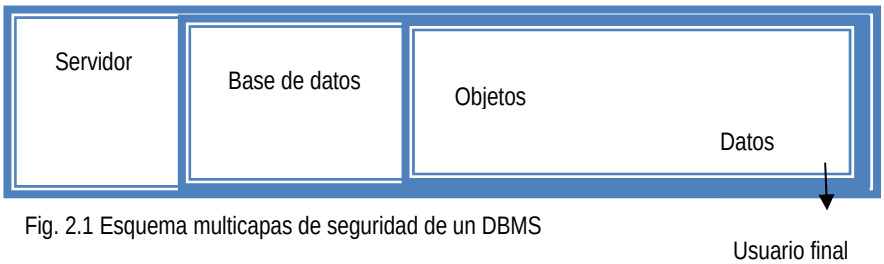


Fig. 2.1 Esquema multicapas de seguridad de un DBMS

2.2 Componentes de un sistema manejador de bases de datos relacional (RDBMS)

Los componentes de un RDBMS son aquellos procedimientos que hacen posible el correcto almacenamiento, su definición y control sobre las modificaciones a nivel físico y lógico, considerando garantizar sus características como base de datos diseñada a partir del modelo relacional.

2.2.1 Lenguaje de definición de datos (DDL) *(Data Definition Language*)

Es un lenguaje artificial basado en un determinado modelo de datos, en este caso el relacional, que permite la representación lógica de ellos la cual se compila y genera una representación orientada a la máquina que es la que utiliza por el RDBMS.

Se utiliza para crear, eliminar o modificar tablas, índices, vistas, *triggers*, procedimientos; es decir, nos permite definir la estructura de la base de datos mediante comandos.

- Create: Utilizado para crear nuevas bases de datos, tablas, campos, índices, vistas, *defaults*, reglas, procedimientos, *triggers*.
- Alter: Utilizado para modificar la estructura de una tabla para agregar campos o *constrainst*.
- Drop: Utilizado para eliminar bases de datos, tablas, campos, índices, vistas, *defaults*, reglas, procedimientos, *triggers*.

2.2.2 Lenguaje de modificación de datos (DML) *(Data Manipulation Language*)

Es un lenguaje artificial en el cual se realizan dos funciones específicas en la gestión de los datos:

- La definición del nivel externo o de usuario de los datos.
- La manipulación de los datos, es decir la inserción, borrado, modificación y recuperación de los datos almacenados en la base de datos.

Al igual que un DDL, un DML está basado en un modelo de datos y por tanto los DBMS basados en distintos modelos de datos tienen diferentes DML.

Dependiendo del modelo de datos y del DBMS existen dos tipos de DML:

- Procedimentales: En las sentencias del lenguaje se específica que datos se van a manipular, que datos se desean obtener y que acciones/operaciones deben ejecutarse para ello.
- No Procedimentales: Los cuales sólo se requiere que en la sentencia se especifiqué que datos se van a manipular, que se desea obtener siendo el propio lenguaje DML el que se encarga de hacer los procedimientos más efectivos para cumplir con lo requerido.

La diferencia está en que los procedimentales tratan a los datos de manera individual y los no procedimentales operan con un conjunto de datos.

Las instrucciones relacionadas con este componente son:

- Select: Permite realizar consultas a la base de datos.
- Insert: Empleado para agregar registros a una tabla.
- Update: Utilizado para modificar los valores de los campos de una tabla.

- Delete: Utilizado para modificar los valores de los campos de una tabla.

2.2.3 Lenguaje de control de datos (DCL) (*Data Control Language*)

Se utiliza para la definición de los privilegios de control de acceso y edición a los elementos que componen la base de datos (seguridad), es decir, permitir o revocar el acceso a objetos y/o datos.

Los permisos a nivel base de datos pueden otorgarse a usuarios para ejecutar ciertos comandos dentro de la base o para que puedan manipular objetos y los datos que puedan contener estos.

Las instrucciones relacionadas con este componente son:

- Grant. Permite otorgar permisos a los usuarios sobre los objetos definidos en la base de datos, así como las operaciones a utilizar sobre ellos.
- Revoke. Permite revocar permisos sobre los objetos definidos en la base de datos y las operaciones sobre los mismos.

2.2.4 Diccionario de datos (DD) (*Data Dictionary*)

El diccionario de datos es uno o un conjunto de archivos que contienen información acerca de los datos que son almacenados en la base de datos.

En el diccionario de datos se encuentra almacenado:

- El esquema lógico de la base de datos.
- El esquema físico de la base de datos.
- Los subesquemas de la base de datos.
- Restricciones de privacidad y acceso a los datos almacenados.
- Las reglas normas o restricciones referentes a la seguridad de los datos.

Sus principales funciones son las siguientes:

- Describe todos los elementos en el sistema.
- Los elementos se centran en los datos.
- Comunica los mismos significados para todos los elementos del sistema.
- Documenta las características del sistema.
- Facilita el análisis de los detalles para evaluar las características y determinar cómo deben realizarse los cambios.
- Localiza errores y omisiones del sistema.

2.3 Sistemas manejadores de bases de datos comerciales y libres, *General Public License* (GNU)

En la actualidad existen muchos sistemas administradores de bases de datos (DBMS), que son imprescindibles para almacenar y gestionar el gran volumen de datos de una organización pero que también se hacen presentes en los ordenadores de una persona que tal vez lo use en sus practicas escolares, como plataforma de un negocio familiar, o simplemente como una herramienta en su PC que utilice como cualquier otro programa para diversas tareas. Como se expuso en el Capítulo 1 en los 70's esto era prácticamente imposible ya que el almacenamiento era costoso, no existía la independencia física y lógica de los datos. Actualmente existen aplicaciones totalmente gratuitas y con código abierto para su desarrollo, instituciones de prestigio internacional como la Universidad de Berkeley en California tienen proyectos específicos para el desarrollo de código de uso libre, la Universidad de Waikato en Nueva Zelanda hoy famosa por su potente aplicación de minería de datos WEKA, sólo por mencionar algunas.

El principal antecedente del desarrollo de software libre se tiene en 1983 con Richard Stallman bajo el objetivo de crear un sistema operativo completamente libre que retomara el espíritu de cooperación que prevalecía en los inicios de la comunidad de usuarios de computadoras. Este proyecto adopto el nombre de GNU que significa "*General Public Licence*" mediante el cual se genera software de distribución libre y que se puede obtener de innumerable cantidad de sitios de Internet. Este proyecto ha generado una apertura a programadores, a generar aplicaciones de gran potencial; potencial que se puede incluso comparar con el desarrollado por grandes marcas trasnacionales.

En la actualidad los DBMS se clasifican por su estructura lógica que se basa en un modelo de datos, por su compatibilidad en cuanto a sistemas operativos, por su costo, por los usuarios, por su distribución incluso por su comercialización.

El modelo lógico de mayor uso en los diferentes DBMS es el relacional por su potencial con el lenguaje normalizado SQL.

A continuación se presentan los DBMS de mayor representación tanto comercial como de uso libre a nivel internacional, con sus características principales.

2.3.1 Microsoft SQL Server.

Distribuido bajo un entorno de Windows por Microsoft; SQL Server es un sistema administrador para bases de datos relacionales basadas en la arquitectura cliente / servidor

2.3.1.1 Características principales:

- Compatibilidad con estándares de W3C, incluyendo XML, Xpath, XSL, HTTP.
- Obtiene código XML de las consultas realizadas con SQL.
- Manipulación de documentos XML.
- Manejo de bases de datos distribuidas.
- Manejo de varias particiones físicas para almacenamientos de datos flexibles.

- Permite realizar algunas tareas de mantenimiento y administración de la base de datos sin tener que darla de baja.
- Permite realizar acciones OLAP (*Online Analitycal Processing*), herramienta que permiten analizar datos almacenados en una base da datos, por medio de cubos de información.
- Consulta y modificación de cubos virtuales de manera gráfica.
- Conectividad con clientes ODBC y JDBC.
- Data warehousing

2.3.2 SYBASE

Desde su fundación en Berkeley, California (EE.UU.), en 1984, Sybase se ha ganado la confianza de muchas de las compañías más importantes del mundo por su habilidad en la gestión de información.

Sybase es la compañía de software empresarial más grande enfocada exclusivamente a la gestión y movilización de información, desde el centro de datos, hasta el punto de acción.

Sus soluciones abiertas y multiplataforma entregan la información de manera segura, en cualquier momento y lugar, permitiendo que los clientes creen una "ventaja de información".

Sybase IQ es un motor de bases de datos altamente optimizado para inteligencia empresarial y una ventaja sobre otros DBMS es su rapidez en consultas hasta 100 veces más rápidas que un RDBMS tradicional.

2.3.2.1 Características principales:

- Diseñado para soportar aplicaciones OLTP (*OnLine Transaction Procesor*), ambiente diseñado para insertar, actualizar y borrar datos en una base datos.
- Permite integrar aplicaciones basadas en XML con la base de datos y crear reglas de negocio que ejecuten Java Beans.
- Conectividad con clientes ODBC y JDBC.

- Soporte para BLOB's (*Large Objects*).
- Permite realizar *querys* XQL, lo cual significa que utiliza un motor abierto para búsqueda dentro de contenidos XML almacenados en la base de datos, o en un URL.
- Tamaño expandido de filas y datos, es decir soporta filas más grandes, columnas más grandes. Se soportan ahora tamaños de páginas de 2k, 4k, 8k o 16k (entre más tamaño de página mayor rendimiento de operaciones SQL)
- Compresión de copias de respaldo.
- Bloqueo a nivel de fila, nivel de página de datos, nivel de página (datos e índices), nivel de tabla.

2.3.3 ORACLE

La arquitectura Oracle es una herramienta cliente - servidor para la gestión de base de datos creada por Oracle Corporation, es considerado el RDBMS más complejo y es vendido a nivel mundial aunque su robustez y su elevado precio lo limita a ser accesible sólo por empresas de gran jerarquía y trasnacionales.

2.3.3.1 Características principales:

- Ofrece varias plataformas de desarrollo para Internet y aplicaciones tradicionales, tales como: XML, Enterprise Java Engine, SQL y PL/SQL, C, C++, entre otras.
- Soporte Unicode.
- Extiende las habilidades de una base de datos para Internet.
- Amplía distintos mecanismos para protección de datos.
- Soporta OLTP y OLAP.
- Data Warehousing.
- Contiene mecanismos de gran funcionalidad y flexibilidad para compartir la información almacenada en la base de datos con otras bases de datos o aplicaciones.
- Conectividad con clientes ODBC y JDBC.

- Soporte para BLOB's.
- Ofrece escalabilidad y performance sin modificar las aplicaciones instaladas.
- Soporta columnas con cifrado de datos.
- Permite replicación de bases de datos (bases de datos distribuidas).
- Ofrece distintas herramientas para la administración de la base de datos.
- Redefinición de tablas en línea.
- Respaldo y recuperación en línea.

2.3.4 Informix y DB2

Es un gestor de base de datos creado por Informix software Inc. Incluye un RDBMS basado en SQL. En 2001 Informix tiene problemas financieros e IBM compra los derechos sobre el RDBMS

2.3.4.1 Características principales:

- Soporta bases de datos de más de 4 TB.
- Soporte para acceso a la base de datos vía Web.
- Provee acceso a cualquier tipo de cliente.
- Permite manejo de base de datos distribuidas.
- Capacidad de replicación de bases de datos.
- Permite realizar querys en paralelo.
- Contiene plataformas de desarrollo con SPL: (*Informix Stored Procedure Language*), C, Java, XML.
- Conectividad vía ODBC, JDBC, OLE/DB.
- Soporta aplicaciones para e-Commerce, e inteligencia de negocios.
- Soporta OLAP y OLTP.
- Data warehousing

Los RDBMS libres de mayor uso en el mercado son PostgresSQL y MySQL.

2.3.5 PostgreSQL

Este ha sido un proyecto que ha evolucionado con el modelo relacional para el manejo de base de datos, teniendo sus inicios a finales de los años 80´s, su característica principal es que tiene la licencia BSD (*Berkeley Software Distribution*) que permite el uso libre y la modificación del código lo cual ha permitido que se hayan liberado nuevas versiones con el paso del tiempo. En 1996 ya se habían lanzado cuatro versiones del RDBMS y es entonces cuando se decide cambiar de nombre de Postgres95 a PostgresSQL.

2.3.5.1 Características principales:

- Base de datos de distribución libre.
- Velocidad.
- Confiabilidad.
- Flexibilidad.
- Bajos costos de operación.
- Conformación a estándares ANSI.
- Estrategia de almacenamiento MVCC para grandes volúmenes.

2.3.6 MySQL

Es un sistema de gestión de base de datos relacional, multihilo y multiusuario con más de seis millones de instalaciones. Sun Microsystems desarrolla MySQL como software libre en un esquema de licenciamiento dual, por un lado se ofrece bajo la GNU para cualquier uso compatible con esta licencia, pero las empresas que quieran incorporarlo en productos privativos pueden comprar a la empresa una licencia específica que les permita este uso. Está desarrollado en su mayor parte en ANSI C.

2.3.6.1 Características principales:

- Soporta los estándares ANSI.
- Contiene esquemas de almacenamiento independiente que se pueden seleccionar de acuerdo a las necesidades.
- InnoDB para transacciones y bloqueo de registros.
- *MyISAM* sin transacciones
- Soporte para SSL.
- *Querys* con manejo de cache que puede incrementar el performance de la base de datos en un 200%.
- Permite manejo de replicación de bases de datos.
- Soporta indexado de texto.

Capítulo 3

Resumen: Se plantea de una manera sencilla y objetiva el concepto data warehouse, antecedentes, características, componentes, funciones y un esquema de funcionamiento genérico, con lo cual se puede integrar un criterio y considerar ventajas de desarrollar este concepto; también se consideran posibles desventajas. Actualmente un sistema de almacenamiento con estas características representa uno de los activos de mayor peso y con gran utilidad a corto y mediano plazo.

Palabras clave: Data warehouse, data mart, OLAP (Online Analytical Processing), sistemas operacionales.

3. ¿Qué es un data warehouse?

El concepto data warehouse (DW) se presenta a mitad de los años 80´s ante la necesidad de implementar aplicaciones con la capacidad de almacenar, administrar y explotar cantidades masivas de datos. La principal diferencia entre una aplicación data warehouse y un sistema de almacenamiento convencional era la de integrar datos de diferentes recursos los cuales podrían ser base de datos internas o externas, históricos de venta, datos de diferentes canales de distribución, la Web etc.

Data warehouse es un concepto, no es un producto que se puede comprar, considera componentes de hardware y software que se utilizan para analizar grandes cantidades de datos y que debe ser el soporte de las decisiones en cualquier ámbito de la organización. Los datos bajo los cuales funciona y opera una organización representan una abundancia de conocimiento, lo cual es un activo que quizás no se explota a su máxima capacidad.

Se considera a Bill Inmon el padre del concepto data warehouse y lo define de la siguiente manera: “Es una colección de datos orientados a temas, integrados, no-volátiles y variante en el tiempo, organizados para soportar necesidades empresariales” según: [Inmon; (1996)].

Un aspecto importante de la teoría de Bill Inmon y que causa discrepancias entre otros autores sobre el tema es que considera el diseño de un data warehouse bajo una metodología descendente (*top – down*), lo cual considera el diseño y la implementación de la aplicación para después determinar los diversos data *marts* que han de integrar la aplicación.

Por otra parte Ralph Kimball otro autor que ha trabajado constantemente en el concepto y su entorno define un data warehouse como: “Una copia de las transacciones de datos, específicamente estructurada para la consulta y el análisis”; y a diferencia de Inmon considera una metodología ascendente (*bottom- up*) en su desarrollo, ya que considera que un data warehouse es la unión de todos los data *marts* que convergen a una sola entidad.

3.1 Características de un data warehouse

Una aplicación DW tiene diversos procesos y herramientas que lo caracterizan respecto a su diseño, entorno, plataforma, usos y concurrencia, pero debe de englobar cuatro características fundamentales:

- Orientado al tema
- Integrado
- De tiempo variante
- No volátil

3.1.1 Orientado al tema

Los datos almacenados en un DW están orientados a temas, es decir son datos relacionados a sujetos o a entidades no a operaciones o funciones, una diferencia respecto de un sistema de almacenamiento convencional es que son datos a gran escala, resumidos, que brindan información respecto a los acumulados y que una vez cargados a la aplicación no se pueden actualizar; otra diferencia es la interrelación entre los datos mientras que en un DW hay datos con un horizonte de tiempo mucho mayor (históricos), y con relaciones no necesariamente vigentes, el de un sistema operacional contiene datos con una relación entre dos tablas basadas en un hecho vigente que almacena datos de una ventana de tiempo de menor amplitud y de acuerdo al entorno los datos se pueden actualizar y modificar.

3.1.2 Integrado

Hoy en día una organización opera bajo diferentes entornos, lo cual genera que los datos provengan de diferentes plataformas, aplicaciones, o de diferentes procesos; antes del proceso de carga de la fuente al DW se deben establecer criterios de codificación para que independientemente de las aplicaciones los datos que tengan diferentes representaciones se almacenen de una sola manera; el nombramiento de atributos, su tipo de dato, su longitud o definición que lo generan se debe homologar para almacenar un solo tipo de dato.

3.1.3 De tiempo variante

La información almacenada en un DW tiene un horizonte de tiempo de gran amplitud, es decir se encuentran históricos de años o de varios años dependiendo de las necesidades y de la granularidad de almacenamiento que más convenga a la organización; a diferencia en un sistema operacional se tiene datos actuales, de vigencia no mayor a 3 o 4 meses, los cuales se pueden actualizar, modificar, reemplazar mientras que en la aplicación DW una vez cargados los datos no se pueden actualizar, es por eso que en un DW se considera la información como una serie de (*snapshots*).

3.1.4 No volátil

En un DW sólo hay dos tipos de operaciones: carga de los datos y acceso a los mismos; a diferencia de un sistema operacional en donde los datos se actualizan a cada momento por las operaciones que se ejercen sobre ellos, se insertan nuevos datos, se remplazan, y se modifican. Las actualizaciones se hacen antes de cargarlos lo cual evita un sin numero de problemas y garantiza información consistente y no volátil.

3.2 Elementos que integran un data warehouse

Considerando que una aplicación DW es un concepto que integra componentes tanto de software y hardware podemos generalizar tres componentes principales independientemente del entorno de la organización que lo implemente.

3.2.1 Metadato

Es un componente fundamental ya que es donde se tiene la información del esquema global tanto lógico como físico de la aplicación, en el también se encuentra información respecto a los datos, sus características, reglas de negocio, procesos, y cualquier definición respecto a la organización.

En los Metadato se debe documentar la siguiente información:

- Jerarquías, dimensiones, y definiciones de los datos.
- Entidades y sus relaciones.
- Reglas de negocio.
- Condiciones operacionales.
- Acciones de contingencia.

3.2.2 Funciones de extracción, transformación y carga (ETL) (*Extract Transform Load*)

En un DW sólo existen dos operaciones la carga y la extracción de los datos aunque la transformación no se considera como tal una fase operativa dentro de la arquitectura de un DW, es fundamental y considera la consolidación, agregación y creación de las diferentes dimensiones que han de representar los datos.

A continuación se definen cada uno de estos procesos dentro de un DW:

- Extracción: Es el proceso para obtener la información deseada a partir de los datos almacenados en las diversas fuentes. Generalmente son procesos batch que apuntan a diferentes sistemas operacionales, recursos u almacenes de datos.
- Transformación: Cualquier operación ya sea aritmética, de selección o exclusión sobre los datos para que estos puedan ser cargados en el DW o se puedan migrar a otras bases de datos.
- Carga: Después de homologar e integrar los datos de las diferentes fuentes, estos se almacenan en el DW.

En una aplicación DW las funciones de extracción, transformación y carga se hacen con sistemas OLAP (*On-line Analitycal Processing*).

Un sistema OLAP se puede entender como la generalización de un generador de informes. Los sistemas OLAP evitan la necesidad de desarrollar interfaces de consulta, y ofrecen un entorno único valido para el análisis de cualquier información histórica, orientado a la toma de decisiones. A cambio, es necesario definir dimensiones, jerarquías y variables, organizando de esta forma los datos.

Existen dos tipos de sistemas OLAP: R-OLAP y M-OLAP.

- R-OLAP: Es la arquitectura en la que los datos se encuentran almacenados en una base de datos relacional se caracteriza por sólo almacenar información relativa a los datos en detalle, evitando acumulados (evitando redundancia innecesaria).
- M-OLAP: Los datos se encuentran almacenados en archivos con estructura multidimensional, los cuales reservan espacio para todas las combinaciones de todos los posibles valores de todas las dimensiones de cada una de las variables, incluyendo los valores de dimensión que representan acumulados.

3.2.3 Middleware

Middleware es un término genérico que se utiliza para referirse a todo tipo de software de conectividad que funciona sobre aplicaciones distribuidas en plataformas heterogéneas. Funciona como una capa de abstracción de software distribuida que sirve en diferentes entornos sin tener que preocuparse por los protocolos de red o el sistema operativo. En un DW el *Middleware* garantiza la conectividad con las diversas fuentes de datos.

3.3 Concepto de data *mart*

Un data *mart* es una versión especial de un DW, o también considerado como un DW de menor dimensión. Son subconjuntos de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones. Los datos existentes en este contexto pueden ser agrupados, explorados y propagados de múltiples formas para que diversos grupos de usuarios realicen la explotación de los mismos de la forma más conveniente según sus necesidades.

El data *mart* es un sistema orientado a la consulta, en el que se producen procesos *batch* de carga de datos (altas) con una frecuencia baja y conocida. Es consultado mediante herramientas OLAP que ofrecen una visión multidimensional de la información.

Sobre estas bases de datos se pueden construir EIS (*Executive Information Systems*) y DSS (*Decision Support Systems*). Se puede generalizar el concepto de data *mart* como subconjunto de un DW ya que debe mantener las cuatro características principales del mismo.

3.4 Diseño de un data warehouse

Como ya se cito un DW no es algo tangible que se pueda comprar; entonces considerar alguna metodología para su proceso de diseño sería incorrecto. Actualmente el desarrollo tecnológico y comercial generan muchas herramientas las cuales han diversificado los canales de distribución de productos y servicios, se han transformado los puntos de ventas; entonces depende exclusivamente del entorno o ámbito de la organización el diseño de la aplicación DW.

Dicho de otra manera el diseño de un DW no depende de un proceso de requerimientos como tal sino depende de los requerimientos de la organización en particular.

Se recomienda tratar de mantener un criterio exhaustivo y consistente respecto a las necesidades de información considerando objetivos y reglas de negocio.

Como se comento anteriormente no hay un método o proceso para un diseño de un DW pero resulta buena práctica el considerar una representación de todo lo que se debe considerar, documentar, delegar, relacionar, etc.; para generar una aplicación que satisfaga las necesidades a cualquier nivel de una organización.

Una práctica que ha venido siendo usada considerando todos los aspectos que involucra una organización es la propuesta por Jhon Zachman que es uno de los pioneros de la llamada "arquitectura empresarial", y su criterio para un buen diseño de un DW se basa en establecer que todos los aspectos de una organización deben ser considerados y los representa mediante una matriz.

Organización	Datos (Qué)	Procesos (Cómo)	Redes (Dónde)	Personas (Quién)	Tiempo (Cuándo)	Motivación (Por qué)
Ámbito						
Modelo						
Modelo						
Modelo de sistema						
Modelo tecnológico						
Componentes						
Sistema Funcional						

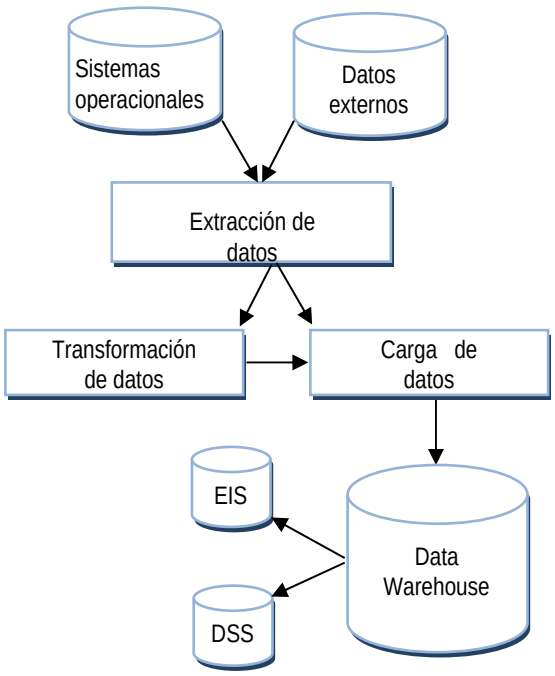
Fig 3.1 Estructura de Zachman considerando todas las necesidades de una organización

Entonces se recomienda comenzar con el modelo de datos, ya que el modelo empresarial de los datos sirve para generar el modelo operacional.

Uno de los aspectos de mayor trascendencia en el diseño es definir el volumen de datos que se pretende almacenar; contando con un panorama general de su uso se debe analizar un aspecto fundamental que es el de la granularidad o variable tiempo de almacenamiento de los datos; ya que se conozca sobre el uso y los usuarios de la aplicación se puede definir con mayor certeza la granularidad que facilite a los usuarios sus análisis, se contara con argumentos para generar consultas por unidad de tiempo que pueden ser por hora, por día, por semana, trimestrales, etc.

Puede resultar complejo apegarse a una estructura de este tipo; pero resulta buena práctica cuando se tratan todos los elementos; se tendrá certidumbre de que se está considerando toda la información que pueda ser parte del modelo de datos y no como suele suceder con los diseñadores de sistemas que solo consideran información operacional o vista desde el punto de vista de desarrollo.

3.5 Esquema de funcionamiento de un data warehouse genérico



El esquema de funcionamiento parte de los sistemas operacionales (canales de distribución, puntos de venta, cajeros, e-commerce y todo lo referente a transacciones que involucren la parte operativa y transaccional de una organización); los datos externos y terceros que los podemos considerar datos referentes a la organización como pueden ser resultados de una campaña publicitaria, monitoreo de competencia e investigación de mercado, estudios demográficos, estudios sociales, indicadores corporativos, generalmente producidos por terceros.

Considerando diferentes fuentes de datos, estos deben ser transformados obteniendo los niveles de agrupamiento y agregación que más convenga a la organización y garantizando la integridad de los datos. Un DW almacena historia por lo que hay que definir la composición o calculo de los diferentes atributos considerando siempre el respaldo y garantizando la continuidad del correcto almacenamiento de los datos; documentando y generando catálogos que definan el sentido abstracto de los atributos y de la integración de los nuevos atributos calculados.

Con los datos a un nivel final de homologación y agregación se procede a realizar la carga de los datos, aquí interviene otro punto trascendental y conveniente a cada organización; hay que recordar que una vez cargados los datos en el DW ya no se pueden actualizar, modificar, eliminar por lo cual hay que evaluar si las fechas de carga de los datos no involucran un cambio, actualización o cualquier tipo de alteración que pueda surgir en el procesamiento de los datos.

Finalmente el acceso a un DW es para gente que busca información para tomar decisiones, comúnmente ejecutivos, analistas, asesores de las diferentes áreas de la organización con pocos o casi nulos conocimientos de programación para poder manipular o hacer consultas complejas, por lo que se pueden crear aplicaciones a partir de las necesidades particulares de los usuarios. Las aplicaciones DSS (*Decision Suport System)* y EIS (*Executive Informant System*) tienen en común que son aplicaciones con características similares en cuanto a su operación, sencilla con una interfaz amigable e intuitiva que permiten su operación por personas que no tengan un perfil técnico para manipular un DBMS. Este tipo de sistemas pueden ser cubos que contengan diferentes vistas, dimensiones periodos de tiempo, gráficas, tableros de control (*Dashboards*), etc.

Un sistema de información ejecutivo (EIS) y un sistema de soporte de decisiones (DSS), son herramientas que están orientadas a la consulta por niveles gerenciales que permiten monitorear las actividades de la organización a partir de la información interna y externa; El DW es la fuente directa de este tipo de herramientas y sus principales ventajas son el entorno fácil y amigable, la capacidad de presentar datos agrupados bajo un periodo de tiempo de mucho mayor amplitud que cualquier sistema operacional.

Otro punto que caracteriza este tipo de aplicaciones es el esquema de seguridad, su grado de concurrencia y la conectividad, respecto a esta se plantea que se trabaje bajo una estructura cliente servidor lo cual establece un esquema de seguridad en capa como ya se mencionó en el Capítulo 2 donde primero se tiene el acceso a los datos, pudiendo accesar a tablas, objetos como usuario final y el nivel de concurrencia depende de la estructura lógica de los sistemas, así como de la capacidad de disco y del servidor.

3.6 ¿Por qué implementar un data warehouse?

Pueden existir muchas justificaciones para la implementación de aplicaciones bajo el concepto DW, pero es una realidad que hoy en día ante mercados globalizados se vuelve imprescindible la cultura de integración, homologación, trabajar bajo métricas equivalentes y tener la capacidad de volver a la historia para comparar históricos y proyectar desarrollo, tendencias y resultados.

Su costo es elevado, quizás una de las razones por las cuales no es una prioridad en cualquier organización pero más que un gasto es una inversión que generara utilidades a corto mediano y largo plazo.

En el ámbito comercial la adopción de este concepto o cultura es necesaria por la competencia a la que están sujetas hoy en día las empresas, las cuales buscan la lealtad de sus clientes y la captación de nuevos así como la maximización de sus ganancias al costo más bajo, entre sus prioridades para mantenerse en el mercado.

Un DW resulta una ventaja competitiva ya que con el conocimiento de sus clientes pueden anticiparse a su competencia para generar acciones tanto operativas como comerciales así como alianzas o promociones entre muchas estrategias más.

3.6.1 Ventajas de un data warehouse

Se consideran ventajas generalizando respecto a lo que genera un data warehouse genérico.

- Liderazgo: Contar con un DW representa hoy día una ventaja competitiva ya que es el motor de cualquier estrategia y cambio en ellas. La capacidad de innovación te lleva al liderazgo, y está se sustenta en la toma de decisiones con cierto nivel de certidumbre o bien tomando riesgos que sean costeables.
- Competitividad en el mercado: El contar con información y conocimiento permite ofrecer mejores productos y servicios al mejor costo lo cual mantiene en el ámbito competitivo a cualquier organización.
- Tendencias del mercado: El análisis de históricos, comparaciones, simulaciones, proyecciones, ayudan a tener un panorama con cierto grado de certidumbre lo cual genera un panorama o posibles escenarios de lo que va a suceder.
- Flexibilidad respecto al entorno tecnológico: El continuo desarrollo tecnológico debe ser un área de oportunidad para cualquier organización, la capacidad de conectividad, almacenamiento, manipulación de un DW permite gran flexibilidad respecto a nuevos sistemas o programas para su operación y explotación.
- Entorno de trabajo fácil, amigable e intuitivo para áreas de negocios: La interfaz gráfica, sistemas de consulta, reportadores, tableros de control son amigables e intuitivos orientados a la consulta, fáciles de interpretar y sin mucho o nada de código a ejecutar.
- Versión única de los datos: Quizás la característica y ventaja de mayor trascendencia de un DW, ya que representa integridad, seguridad y la certeza de que se trabaja con datos fidedignos para tomar cualquier decisión.

- Fácil accesos a gran variedad de datos: Independientemente del ámbito de la organización se puede obtener gran variedad de datos, bajo diferentes vistas, dimensiones, ventanas de tiempo, relaciones, restricciones, exclusiones etc.
- Tiempo de análisis e implementación de nuevos proyectos: El tener los datos integrados en una sola fuente (DW) ayuda mucho y realiza casi el 70-80% del trabajo de cualquier reporte, análisis e informe sobre un proyecto o análisis en particular.
- Conocimiento de los clientes: Ya sea una organización pública o privada se puede tener gran conocimiento, perfil demográfico, características principales, transacciones, comportamiento de sus clientes o usuarios lo cual es un insumo para tomar decisiones, abrir nuevas cedes, generar nuevos canales, cambiar estrategias, evaluar el funcionamiento y servicio todo orientado al cliente o usuario.
- Estados financieros: Es un hecho que un DW representa una inversión pero a través del almacenamiento de los datos se debe considerar indicadores, que permitan evaluar la inversión, y su desarrollo en el ámbito costo-beneficio y en su caso determinar el retorno de la inversión, incluso se puede determinar un pronostico sobre cuando se tendrá el retorno de la inversión (ROI), y bajo que condiciones o circunstancias.
- Capacidad de reacción para situaciones de contingencia: Se considera así a situaciones o casos atípicos o anormales en un ámbito en particular; pueden ser las ventas semanales, el lanzamiento de un producto, la cantidad de transacciones de una cuenta bancaria, una epidemia en un sector de la población entre muchas otras situaciones en particular lo importante es la capacidad de reacción para en base a los datos almacenados determinar posibles desviaciones, casos atípicos o anormales y situaciones criticas que requieran un análisis y seguimiento en particular.

3.6.2 Desventajas de un data warehouse

- Alta inversión: Puede resultar una inversión fuera de la capacidad adquisitiva de una organización y en relación con el uso que habrá de dársele.

Esto suele suceder en organizaciones públicas principalmente ya que compran software y usan plataformas sin la previa capacitación para su explotación.

- Problemas de implementación: Si no se tiene una estructura previa, o un sistema operacional como bases de datos, estructuras de datos, diagramas de procesos, se pueden presentar omisiones en el diseño, desarrollo y operación pueden resultar ineficientes.
- Un data warehouse puede tener un ciclo corto de vida: Es decir se puede hacer relativamente obsoleto dependiendo de su entorno y de las herramientas tecnológicas con las que se trabaje: Es fundamental evaluar la capacidad de almacenamiento, el software con el cual se opera y se operara en corto y mediano plazo así como la razón de implementar un DW.
- Falta de conocimientos por parte de áreas de negocios para su explotación y mantenimiento: No puedes tener un DW y no explotarlo y no puedes implementar un DW y no saber para que se desarrollo, un DW necesita mantenimiento además de su implementación debe estar apegado a un modelo ya sea relacional o multidimensional lo cual con el tiempo puede sufrir cambios; se pueden generar nuevas relaciones, pueden no estar vigentes muchas hay que evaluar y probar periódicamente la operación del DW. Y establecer indicadores o puntos de control.

Capítulo 4

4 Descubrimiento de conocimiento en base de datos

Resumen: Considera la definición, antecedentes, y proceso del descubrimiento de conocimiento en base de datos (KDD), el cual tiene como parte de su proceso la aplicación de técnicas de minería de datos (DM), con el objetivo de generar conocimiento previamente desconocido, apoyado principalmente de modelos descriptivos y modelos predictivos. Se describe el uso de WEKA, software de uso libre que además de aplicar técnicas de minería de datos cuenta con la conectividad a bases de datos y por ser desarrollado en Java es compatible casi con cualquier plataforma o sistema operativo.

Palabras clave: Descubrimiento de conocimiento en base de datos (KDD), minería de datos (DM), modelos descriptivos y modelos predictivos.

El descubrimiento de conocimiento en base de datos es un proceso que se ha venido desarrollando con mayor trascendencia en la última década; pero que tiene como precedente el análisis exploratorio de datos desarrollado por Jhon Tuckey en los 70's – 80´s.

En el trabajo se ha enfatizado sobre los datos, su almacenamiento y explotación que conforman una tecnología creciente en todos los aspectos que ha rebasado cualquier expectativa y con tendencias sobre su desarrollo muy alentadoras; su análisis es un proceso clave en un mercado competitivo y globalizado.

En la actualidad es prácticamente imposible trabajar con una base de datos o un sistema de información operacional sin el apoyo de aplicaciones o programas de análisis; una característica de estas es la del análisis de grandes cantidades de datos a través de procesos semiautomáticos,

Que generen resultados útiles, comprensibles, novedosos y con eficacia que se puedan replicar para construir algoritmos o para replicarlos de manera semiautomática.

Hay muchos métodos diferentes que son clasificados como técnicas de descubrimiento de conocimiento en base de datos. métodos cuantitativos, como los probabilísticas y los estadísticos, hay métodos que utilizan las técnicas de visualización, hay métodos como la clasificación de Bayes, lógica inductiva, análisis de decisión, redes neuronales y los métodos híbridos que combinan dos o más técnicas.

Si se pudiera caracterizar a un proceso de descubrimiento de conocimiento se tiene que considerar los siguientes aspectos:

- Interactivo
- Comparativo
- Iterativo

4.1 Definición

El descubrimiento de conocimiento que en inglés es “Knowledge Discovery in Databases” (KDD), se define cómo: “La extracción no trivial de información implícita, desconocida y potencialmente útil”, según: [Shapiro; (2000)]

Es un proceso que considera varias fases y que hoy en día tiene gran desarrollo por que en ocasiones se usa indistintamente con otro concepto muy comercial por su entorno “minería de datos” el cual forma parte del proceso KDD.

El proceso KDD tiene como característica principal la de extraer conocimiento a partir de grandes cantidades de datos, pero ese conocimiento busca cumplir con las siguientes propiedades generales:

- Valido

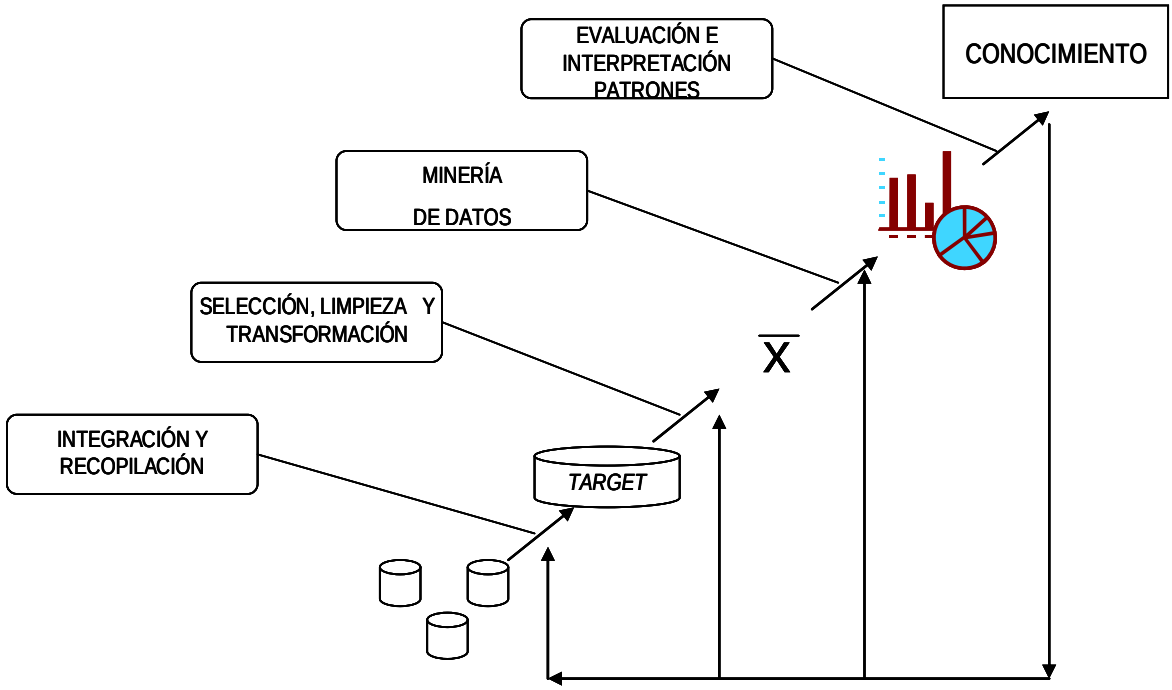
- Novedoso
- Potencialmente útil
- Comprensible

4.2 Proceso de descubrimiento de conocimiento en base de datos

El descubrimiento de conocimiento en base de datos es un proceso el cual está integrado por varias fases que de manera genérica se clasifican de la siguiente manera:

- Integración y recopilación
- Selección, limpieza y transformación
- Minería de datos
- Evaluación e Interpretación de patrones
- Difusión, uso y monitorización del conocimiento

PROCESO GENÉRICO DE DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS (*KNOWLEDGE DISCOVERY IN DATABASES (KDD)*)



4.2.1 Integración y recopilación

El desarrollo tecnológico así como los diferentes canales de distribución y los diferentes sistemas operacionales hacen necesaria una etapa de recopilación e integración de los datos. En esta fase del proceso se busca recopilar datos de las diferentes fuentes de información; estas pueden ser internas y externas y lo que se busca es tener datos consolidados e integrados y tener información homogénea y valida.

4.2.2 Selección, limpieza y transformación

Esta fase está compuesta por actividades fundamentales en el proceso de extracción de conocimiento:

- Selección: Hay que considerar además de las diferentes fuentes de información; que son cantidades masivas por lo cual la selección de los datos a ser analizados deben ser totalmente relevantes y no considerar datos irrelevantes y que resulten redundantes, triviales o innecesarios.
- Limpieza: Hay que prever que pueden existir datos faltantes, errores en los proceso de captura, formatos etc. lo cual genera datos poco precisos y que puede sesgar la información obtenida lo que hace necesaria la tarea de “limpiar”, considerando la posibilidad de exclusiones, filtros, cruces y la aplicación de reglas de negocio.
- Transformación: Las actividades cotidianas están sujetas a un proceso definido que se ajusta para dar resultados a una estructura operacional, pero quizás no sea factible para un análisis de históricos, o de mayor complejidad que un reporte; por lo cual se recomienda considerar la transformación o funciones de agregación de los datos para un análisis mas complejo como puede ser la minería de datos.

4.2.3 Minería de datos (*Data Minning* (DM))

Es la etapa de mayor trascendencia en el proceso KDD y por lo cual en muchas ocasiones se utiliza indistintamente el concepto en el mercado, el objetivo es similar o técnicamente el mismo que es obtener nuevo conocimiento no explicito en conjuntos de datos que pueda ser útil para el usuario.

En los últimos años con gran desarrollo y potencial de seguir creciendo por su entorno y relación con otras disciplinas, la minería de datos hoy es una herramienta establecida y que con disciplina, desarrollo y continuidad puede ser la piedra angular de cualquier organización.

4.2.3.1 Definición

"Es el proceso de extraer conocimiento útil y comprensible, previamente desconocido desde grandes cantidades de datos almacenados en distintos formatos", según [Witten; (2000)]

"Es una fase del Proceso KDD que basada en algunos métodos (algoritmos), produce una enumeración de patrones (modelos), sobre los datos"

Considerando diferentes definiciones se establece que la minería de datos (DM), es un proceso que inmerso en otro proceso KDD involucra una metodología para generar patrones los cuales representan un modelo, una secuencia, reglas, asociaciones y relaciones sobre grandes cantidades de datos.

4.2.3.2 Características y objetivos principales de la minería de datos

De manera genérica la minería de datos genera dos tipos de modelos:

- Modelos descriptivos: Identifican patrones, fluctuaciones y variaciones que explican o resumen los datos explorados; es decir analizan las propiedades de los datos; no predicen nuevos datos.
- Modelos predictivos: Pretenden estimar valores a futuro o desconocidos de una variable de interés llamada variable objetivo o dependientes y se basan en otros datos o atributos que se conocen como variables predictivas o independientes.

El ajuste y validación de estos modelos tienen como objetivo generar información que se puede establecer como:

- Asociaciones
- Secuencias
- Agrupaciones
- Clasificaciones
- Pronósticos

4.2.3.3 Algunas técnicas de minería de datos

Las técnicas de minería de datos son la base analítica para determinar un conocimiento sobre grandes cantidades de datos. Actualmente existen métodos definidos que ajustan a la realidad y entorno de una organización y por lo cual ya se tiene acceso a ellos en software o incluso en sistemas manejadores de bases de datos.

Las técnicas de minería de datos de mayor representación son:

- Clasificación: Una de las técnicas de mayor uso por parte de la minería de datos es la clasificación la cual es un modelo predictivo que asigna a cada registro una clase, que se identifica mediante un atributo. Este atributo puede tomar diferentes valores discretos, los cuales representan una clase. El objetivo es predecir la clase de nuevos registros de los cuales se desconoce su clase
- Regresión: Las técnicas de regresión son modelos predictivos, cuyo objetivo es el de generar una función real para asignar un valor real a nuevos registros. La diferencia con modelos de clasificación es el tipo de datos; en un modelo de regresión son números reales.
- Árboles de decisión: Son una serie de decisiones o condiciones organizadas en forma jerárquica, a modo de árbol. Existen varias técnicas que se pueden generalizar como modelos predictivos y entre ellos los más usados son árboles de decisión o clasificación y árboles de regresión.
- Asociación: Técnicas que generan modelos predictivos bajo el objetivo de encontrar relaciones no explicitas entre datos categóricos. Lo importante de esta técnica es que no debe existir necesariamente algún tipo de relación entre los datos, es decir una relación causa efecto, por lo que puede o no existir una causa para que los datos estén relacionados.
- Clustering o segmentación: Son técnicas que generan modelos descriptivos y consisten en obtener grupos naturales a partir de los datos. La diferencia con técnicas de clasificación es que no trabaja con clases (etiquetas), sino se busca analizar los datos para encontrar esas clases.
- Redes neuronales: Una de las técnicas de mayor interacción con otras disciplinas como la computación y la toma de decisiones, las redes neuronales representan una técnica muy potente que permite modelar problemas muy complejos entre los cuales puede haber interacciones no lineales entre atributos. Representada por grafos dirigidos con nodos y arcos que representan su dirección o relación. Esta técnica puede trabajar con datos numéricos o con datos categóricos pero estos se tiene que enumerar.

4.2.3.4 Software de minería de datos

Parte del continuo crecimiento de la minería de datos se debe al desarrollo de software que en esencia permita el tratamiento de cantidades masivas de datos así como el análisis de los mismos todo ello con una interfaz gráfica amigable y un proceso semiautomático e inductivo.

Hoy en día existe una gran apertura comercial de aplicaciones que ofrecen paquetes de minería de datos como Microsoft SQL Server, IBM, ORACLE, SAS, Teradata, SPSS entre otras, pero también existen aplicaciones de uso libre como WEKA, XL- Sipina, Tanagra, R que dan el uso de las diferentes técnicas de minería de datos además de que buscan el autoaprendizaje de los usuarios.

4.2.3.4.1 WEKA (Waikato Environment Knowledge Analysis)

Para el análisis de los datos del producto “Regiones socioeconómicas de México” del Capítulo 7 se utilizó el paquete WEKA, que está disponible bajo la licencia pública general de (GNU) Software conocido por su inducción al autoaprendizaje y desarrollo de técnicas de minería de datos.

Sus ventajas además de ser de uso libre y de gran compatibilidad ya que está implementado en Java, lo cual lo hacen funcionar casi en cualquier plataforma. De uso inductivo por su interfaz gráfica es un software que contiene una extensa cantidad de técnicas de procesamiento de datos y modelado. Otra característica importante es el acceso a base de datos con SQL con la conexión JDBC (*Java Data Base Conectivity*).

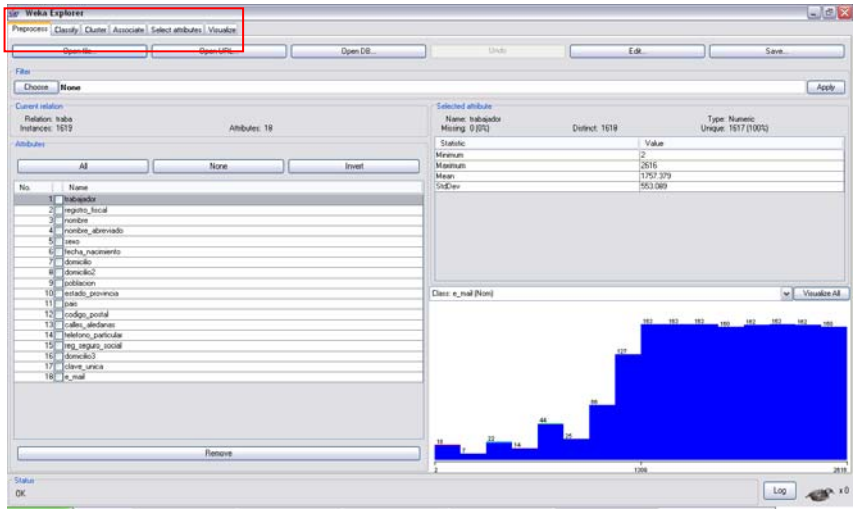
WEKA soporta varias técnicas de minería de datos como son clasificación, regresión, *clustering* y visualización entre las más representativas.

WEKA contiene cuatro módulos o interfaces de trabajo que son:

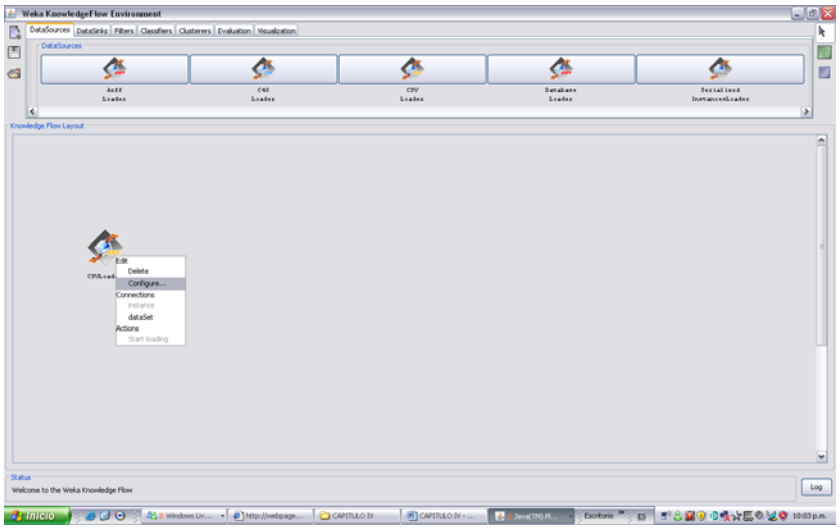
- Simple CLI: Interfaz simple de la línea de comandos (*Simple Command-Line Interface*), se trata de una consola que permite acceder a todas las opciones de WEKA desde la línea de comandos (Java).
- Explorer: Dispone de varios paneles que acceden al banco de trabajo de la aplicación, desde aquí se pueden exportar datos de tipo csv, acceder a una base de datos y filtrar datos con los diferentes algoritmos de filtrado de la aplicación, además de eliminar registros o atributos, también se puede observar las distribuciones de los diferentes atributos de los datos y otros indicadores descriptivos como máximo, mínimo, media, mediana entre otras.

Esta suite del programa es de gran funcionalidad para las fases de integración y recopilación ya que puede importar un archivo .csv o conectarse a una base de datos; además apoya la fase de selección, limpieza y trasformación de los datos, dando un panorama descriptivo muy objetivo que permita saber sobre los datos y sus principales indicadores y cualquier tipo de desviación como pueden ser datos faltantes.

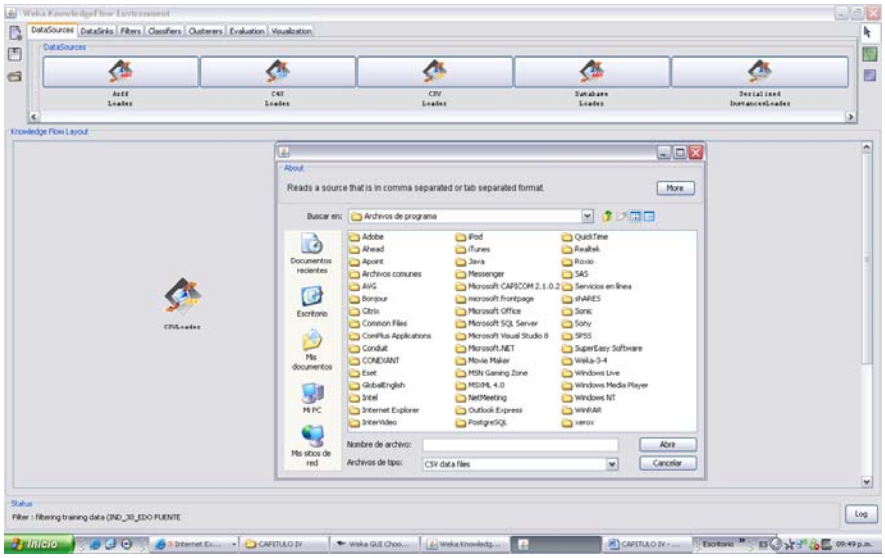
Como se señala en el siguiente cuadro ilustrativo el módulo explorer tiene acceso a diferentes técnicas de minería de datos como son "*Classify*", "*Cluster*", "*Associate*", "*Visualize*"; En el capítulo 7 se presenta su uso y se ilustra más a detalle la potencialidad de algunas técnicas.

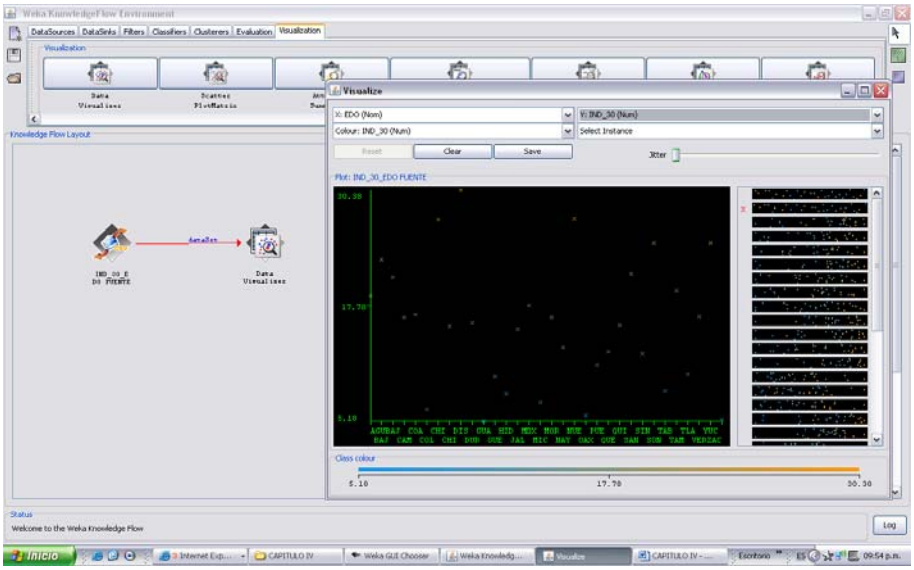


- KnowledgeFlow: Al igual que con el módulo explorer permite ejecutar técnicas de minería de datos, puedes hacer los mismos procesos e incluso otros más complejos; además te presenta una interfaz gráfica del desarrollo de tu proceso llamada “Layout”; es decir puedes visualizar gráficamente la relación o relaciones que se presentan y técnicas que estás utilizando, conforme ejecutas tu proceso además se refleja gráficamente los atributos de los algoritmos que se utilizaron lo cual sirve como control por si se quiere simular bajo otros atributos o supuestos el mismo conjunto de datos.



- Datasources: Presenta las distintas fuentes que puedes importar a WEKA, un archivo propio de WEKA, un archivo separado por comas o una extracción de una base de datos; seleccionas la fuente de los datos y haces click sobre el layout, se presentara la fuente, sabiendo que tipo de fuente representa tus datos ahora hay que cargarlos al banco de trabajo con click derecho sobre el botón de la fuente y seleccionando “configure”, donde te desplegara la ruta y el tipo de archivo de tu fuente.





Con los datos cargados puedes aplicar gran cantidad de algoritmos y técnicas de minería de datos, filtros, experimentos, visualizar mediante gráficas los datos y sus relaciones, lo cual es otra herramienta de gran apoyo.

Este módulo en particular tiene ventajas sobre los demás, las más relevantes son:

- Existen técnicas que no se pueden realizar con la suite explorer.
- Construye relaciones más abstractas y complejas.
- Describe el proceso claramente.
- Puedes automatizar procesos y aplicar diferentes parámetros a un mismo proceso.

En el Capítulo 7 se desarrollan e implementan de manera práctica algunos algoritmos y técnicas de WEKA, por lo cual es importante mencionar sobre su potencialidad siendo software de uso libre y código abierto que hoy representa una opción en el mercado de análisis y minería de datos compitiendo con aplicaciones como SPSS, *Data Miner y Enterprise Guide de SAS, Intelligent Miner* de IBM todos estos reconocidos Internacionalmente.

4.2.3.5 Aplicaciones de minería de datos y su relación con otras disciplinas

La minería de datos representa un campo interdisciplinario que por su naturaleza y entorno puede implementarse en cualquier sector, lo mismo puede implementarse en una organización pública que en una organización privada aunque esta última es donde tiene mayor auge por que su desarrollo constituye hoy una herramienta muy robusta y que se traduce en una ventaja competitiva.

A continuación se citan algunas de las muchas aplicaciones que puede tener la minería de datos:

Negocios

- Hábitos de compra: Una de las aplicaciones más conocidas de la minería de datos es el análisis de los hábitos de compra, con diferentes técnicas hoy se tiene información respecto a asociaciones de compras, secuencias, lo que puede generar una estrategia de una promoción, un paquete de algunos productos y servicios, pronostico de ventas, temporalidad de un producto, su rotación; entre otras cosas que brindan a personas de negocios suficiente información para tomar mejores decisiones.
- Conocimiento de clientes y prospectos: La competitividad que se presenta en los diferentes mercados hace necesario buscar a toda costa la lealtad de los clientes. El conocer sobre hábitos, y comportamientos se puede lograr a través de métodos descriptivos y predictivos, lo que representa un valor agregado para poder anticipar y cubrir nuevas necesidades; con está información podemos determinar prospectos y mantener la lealtad de los que ya son nuestros clientes.
- Prevención y riesgos: El conocimiento de clientes y prospectos ayuda a predecir comportamiento y evaluar riesgos, además de considerar nuevas políticas de crédito, de seguridad y en general políticas operativas; esto resulta muy útil en la comparación de poblaciones para considerar posibles transformaciones y desviaciones que pueden representar pérdidas y cambios en los portafolios analizados.
- Recursos humanos: Uno de los factores que forja una organización es su personal; el reclutamiento y evaluación de personal se pueden analizar con técnicas de minería de datos, muchas de estas sirven para evaluar tendencias y patrones sobre conductas; también se puede aplicar para generar perfiles sobre características deseadas de personal, aspectos demográficos y otros datos que puedan determinar el desarrollo personal y profesional de un individuo.

Minería Web

Sin duda alguna el desarrollo Web trajo consigo una era o un hecho inconmensurable a la humanidad, su crecimiento y explotación son motivos de múltiples proyectos de investigación; con el tiempo y tras convertirse no solamente en un canal sino también en un negocio, han surgido indicadores como: el trafico en línea, tiempo de navegación promedio y perfiles de cibernautas. Al resultar hoy un mercado con apertura mundial resulta de vasto interés su estudio.

Con una cantidad masiva de transacciones resulta ser un medio idóneo para aplicar un proceso de minería de datos. El poder segmentar, asociar, establecer secuencias, agrupaciones y realizar pronósticos sobre la Web son procesos necesarios que constituyen toda una área en una organización. El comercio electrónico "*e-commerce*" es ya otro canal, el poder comprar en línea, solicitar una prueba de manejo, agendar tus próximas vacaciones, resulta un área de oportunidad para cualquier organización; su estudio, desarrollo e interpretación son hechos que pueden mantener o hacer crecer o terminar con una organización en mercados tan estrechos.

Salud y Ciencia

En la salud procesos de minería de datos tienen gran auge en la Biología, Genética, Epidemiología por citar solo algunas. El estudio de pacientes y enfermedades, su clasificación, asociaciones y susceptibilidades lo hacen un campo de aplicación con gran potencial.

Por otra parte en ciencia podemos hablar de proyectos de ingeniería eléctrica, la robótica y sistemas expertos donde tienen gran auge técnicas de clasificación, árboles de decisión y las redes neuronales.

Con esto se puede concluir con que el campo de aplicación de la minería de datos es aquel en donde existan grandes cantidades de datos que mediante técnicas analíticas y procesos de automatización generen información útil y comprensible para tomar decisiones.

Su relación se establece directamente con disciplinas como Estadística, Base de datos, Informática, Toma de decisiones entre otras.

4.2.4 Evaluación e interpretación de patrones

La evaluación e interpretación de patrones que genera la minería de datos son parte del proceso KDD. Como tal son una fase del KDD pero en la práctica son dos actividades independientes que tienen tareas especificas y que convergen a generar una decisión en cuanto a la información y/o conocimiento encontrado en el proceso KDD.

- La evaluación: busca validar analíticamente el modelo y técnicas utilizadas en la fase de minería de datos, es decir evalúa dependiendo el modelo utilizado su calidad y/o resultados. Por ejemplo para un modelo de clasificación se considera el número de clasificaciones correctas entre el total de registros clasificados.

- Interpretación: La precisión de un modelo no garantiza la representación correcta de la realidad, por lo cual se tiene que tener pleno conocimiento del entorno y naturaleza del problema que se pretende modelar y tener la certeza de que los resultados son correctamente interpretados.

Los resultados que hasta aquí se tienen del proceso KDD pueden ser quizás una muestra o parte de un universo que resulta impractico evaluar en su totalidad pero que por resultados históricos o parámetros poblacionales y una correcta interpretación podemos definir si el modelo es el correcto o si existe algún error en el proceso.

4.2.5 Difusión, uso y monitorización del conocimiento

Una vez validado el modelo su difusión y uso deben de comenzar a dar un conocimiento. Hay que recordar que un proceso KDD quizás lo realice un individuo o un grupo de individuos expertos y que el usuario o usuarios finales pueden carecer de conocimientos técnicos por lo que es importante una difusión tanto técnica como conceptual detallada; probablemente con manuales o con glosarios de terminología que oriente y de información a los usuarios finales para que se le de un uso correcto y que explote todo su potencial.

El uso y su monitorización deben considerar que un modelo y el evento que representa ese modelo evolucionan, por lo cual se tiene que tener una continua monitorización que determine si la información y resultados son los idóneos o si se tiene que reconstruir, realinear el modelo o el proceso completo.

Muchos de estos cambios pueden depender del área donde se aplique o desarrolle un modelo, por ejemplo en el ámbito de la salud podemos tener un modelo de clasificación de síntomas que presenta un paciente o una serie de pacientes y que conforman un cuadro para determinar un diagnostico, pero la naturaleza y las condiciones de vida se transforman y cambian lo cual pude alterar ese tipo de clasificaciones y finalmente el diagnostico no será el correcto.

Otro ejemplo de monitorización y uso de los resultados; quizás pueden ser solicitudes de crédito donde se considere únicamente la liquidez de pago, que se determina a partir de los ingresos de un trabajo formal y encontramos que la población económicamente activa se ha transformado y que hoy día las personas entre 20 y 23 años que trabajan o tienen ingresos de manera independiente o informal tienen mucho mayor capacidad adquisitiva que las personas con un empleo formal de entre 30 y 35 años. El modelo tendría que considerar un ajuste o realineación para establecer nuevas políticas y evaluar el riesgo de conceder créditos o hacerlos mas accesibles para personas entre 20 y 23 años.

En el Capítulo 7 se presenta de manera práctica muchos de estos conceptos, técnicas, aplicaciones y una visión con resultados de lo que genera el descubrimiento de conocimiento en base de datos KDD.

Capítulo 5

5 Gestión de la relación con clientes (CRM) (*Customer Relathionship Managment*)

Resumen: Se define el concepto CRM (*Customer Relationship Managment*), antecedentes, área de desarrollo, características principales y su relación principalmente con sistemas de almacenamiento de datos y el proceso KDD para convertir datos en información. Los diferentes tipos de CRM, y como visualizarlos para integrar una estrategia. Usos y tendencias a corto plazo.

Palabras clave: Gestión de relación con clientes (CRM), estrategia, personalización

Como antecedente principal se tiene después de la Segunda Guerra Mundial la implementación del concepto de calidad y calidad total; en ese tiempo se fabricaba y se producía en masa por lo cual se deja el servicio de contacto personal y se ejerce un servicio anónimo, y de una calidad estándar; pero es hasta los 90´s que se da un giro importante y paralelo al desarrollo de los sistemas administradores de bases de datos, se considera como objetivo primario la satisfacción del cliente y con la posibilidad de tener datos específicos se busca generar un trato personalizado o sobre un grupo con características similares.

En los 90´s con numerosas aplicaciones que brindan herramientas de gran alcance y dinamismo, sistemas de información estructurados y orientados a dar conocimiento y soportar mejores decisiones se desarrolla el concepto de CRM llamado así por sus siglas en inglés (*Customer Relationship Management*), ó en español “Gestión de relación con clientes”

5.1 Definición

El CRM es una estrategia o un conjunto de estrategias apoyadas de tecnología, principalmente de sistemas administradores de bases de datos para identificar, atraer y retener a prospectos y clientes, buscando satisfacer necesidades con el objetivo de crear una relación para buscar su fidelidad.

"Es el conjunto de estrategias de negocio, marketing, comunicación e infraestructuras tecnológicas, diseñadas con el objetivo de construir una relación duradera con los clientes, identificando, comprendiendo y satisfaciendo sus necesidades.", según: [AECEM; (2007)]

"Es una estrategia de negocios y no una simple solución de software. Los principios básicos del CRM:

- Alinear la organización alrededor de los clientes
- Compartir información de los clientes en la empresa
- Extraer información de diferentes fuentes para entender mejor a los clientes y anticipar sus necesidades." según: [Warrilow; (2007)]

5.2 Características del CRM

A continuación se citan características del CRM su estructura y desarrollo:

- Es una estrategia o una serie de estrategias no es software o un conjunto de aplicaciones.
- Su desarrollo e implementación dependen en gran parte de la cultura y visión de la organización que pretende implementarlo.
- Considera un enfoque global de la organización y no sólo de un área o departamento.
- Todas sus acciones deben converger a satisfacer las necesidades de los clientes y anticiparse a futuras necesidades, implica el desarrollo de una cultura de mejora continua.
- Implica cambiar procesos.
- El CRM es interactivo y predictivo no reactivo

- Se relaciona principalmente con la mercadotecnia, ventas y servicios lo cual implica canales en pleno desarrollo; Internet y aplicaciones tecnológicas de comunicación, venta y servicio implican una continua revisión de los procesos apegados a nuevas tecnologías.

5.3 El CRM y su relación con otras disciplinas

La mercadotecnia, informática, bases de datos, estadística, minería de datos, servicio al cliente, Business Intelligence (BI), entre otras son disciplinas de gran interacción con el CRM.

El CRM no es software; es una o varias estrategias que se complementan de más disciplinas que tienen como objetivo general crear una relación con los clientes.

No se pueden generar estrategias (CRM), si no tienes una base de datos que te permita obtener un conjunto de datos con características específicas; no puedes obtener un conjunto de clientes que representa un comportamiento homogéneo de compra en los últimos tres meses sin el procesamiento de datos y técnicas estadísticas o de minería de datos.

En la actualidad existen numerosas alternativas para desarrollar proyectos de gran jerarquía, el uso de software libre brinda una herramienta con tanto potencial como herramientas comerciales; entonces una estrategia de CRM la puede implementar cualquier organización independientemente de su jerarquía y ámbito de negocio.

5.4 Tipos de CRM

El CRM se puede clasificar en tres tipos:

1.- CRM analítico: Es la parte del CRM que tiene como objetivo conocer más y mejor a los clientes, todo esto apoyado de técnicas estadísticas y minería de datos, con herramientas tecnológicas principalmente bases de datos que permitan identificar y conocer las necesidades y hábitos de los clientes y explotarlas para llevar a cabo acciones. Una vez con información

conceptualizada para definir la potencialidad del cliente y su comportamiento se decide el canal y acción a seguir y ejercer a través del CRM operativo.

2.- CRM operativo: El CRM operativo constituye los diferentes canales de comunicación y contacto con el cliente los cuales pueden ser mediante correo directo, telemarketing, cara a cara o las diferentes vías que hoy brinda Internet: e-mail marketing, encuestas en línea y *chat*, entre las más usadas.

3.- CRM colaborativo: Consiste en Integrar los diferentes canales del CRM operacional y facilitar su interacción con los clientes además puede considerarse una fase de interacción entre el CRM operativo y el CRM analítico que busca coordinación, cooperación y cohesión de equipo para explotar y compartir información.

5.4.1 CRM analítico

El CRM analítico comprende las distintas técnicas, herramientas y tecnología mediante las cuales podamos explotar los datos y obtener conocimiento básicamente de tres tipos:

- Socio-demográficos: Se refiere básicamente a aspectos de identidad como son nombre, edad, género, residencia, estado civil, ocupación y/o profesión entre los más importantes. Este tipo de datos ayudan a conocer más de un individuo lo que facilita generar un perfil, una clasificación o segmentación de un mercado.
- Transaccionales: Este tipo de datos explican las transacciones o hechos mediante las cuales se determinan las relaciones que existen y que pueden existir; es decir sabemos si han comprado algún tipo de producto o servicio, o si tienen algún interés en algún producto, si buscan informes, cotizaciones, lo cual crea un perfil de cliente o de cliente potencial.

- Comportamiento: La asociación de diferentes hábitos y el conocimiento de gustos y pasatiempos ayudan para poder anticipar diferentes necesidades y dar mejor servicio al cliente, el saber que le gusta hacer los fines de semana, si le gusta el cine o la pizza o si prefiere descansar, ayuda a generar un mejor perfil y a satisfacer necesidades ya que puedes vender o prestar un servicio pero también puedes ofrecer otro que complemente o este relacionado con sus pasatiempos y hábitos.

5.4.2 CRM operativo

El CRM operativo es la implementación, desarrollo y ejecución de los diferentes medios o canales de comunicación a través de los cuales se busca una relación de la organización con los clientes y principalmente son:

- Correo directo: Quizás es el medio o canal de comunicación de mayor antigüedad, se caracteriza por ser un medio controlado respecto a un medio masivo como la televisión o la radio. Su éxito depende de la lista de clientes o prospectos a los cuales se les envíe la comunicación lo cual depende de la recopilación y gestión de la base de datos.
- Telemarketing: Es el canal de mayor apertura del CRM, el cual ha venido siendo explotado en México desde finales de los 90´s y que consiste en un contacto directo vía telefónica operador-cliente. Por su costo es frecuentemente utilizada este tipo de estrategias de CRM, a finales de los 90´s era la más común pero un abuso en su práctica y la comercialización de datos confidenciales a afectado severamente su implementación y sobre todo sus resultados.
- E-mail marketing o *emailling*: Representa hoy en día el medio de mayor velocidad, penetración, uso y el de menor costo. Consiste en mandar un correo a direcciones electrónicas de una base de datos, mediante un software que genera el envío masivo. Este puede estar personalizado y cuenta con grandes beneficios para ser considerado como un medio para implementar una campaña, lanzar un nuevo producto, o realizar promociones. La desventaja del correo directo o e-mailing es que la veracidad de los datos es relativa lo cual ocasiona que no se trate con individuos lo que implica la no

relación con individuos o si existe esa relación que la información de ese individuo corresponda a su verdadera identidad. Además del bloqueo de clientes que evitan este tipo de comunicación clasificándolo como correo basura.

5.4.3 CRM colaborativo

El CRM colaborativo facilita la interacción organización-cliente a través de todos los canales de comunicación además de dar soporte y coordinar a los empleados y equipos de servicios. El objetivo es integrar recursos humanos, procesos y datos para servir mejor a los clientes.

5.4.4 Usos y tendencias del CRM

Actualmente representa gran parte de las estrategias de una organización comercial o privada, pero en México el CRM en organizaciones públicas y en algunas organizaciones privadas tiene un desarrollo muy lento o casi nulo, esto debido a la estructura y a la falta de presupuesto.

La mejor alternativa para su implementación y desarrollo es el uso de software libre; la implementación de estructuras de almacenamiento de datos en municipios e incluso en estados son una necesidad que no se esta cumpliendo.

Las tendencias del CRM tienen como futuro marcar sectores de población finitos y homogéneos, a los cuales se les pueda gestionar dependiendo de la dependencia con programas de acción pública por parte de organizaciones estatales y con promociones y programas casi personalizados por parte de organizaciones privadas.

El desarrollo de sistemas administradores de bases de datos (DBMS) y del descubrimiento de conocimiento en bases de datos (KDD), han sido fundamentales para el CRM; recientemente nuevas áreas de estudio apuntan a sistemas de información geográfica (*Geographic Information Systems*), lo cual representa un área de desarrollo más para el CRM; el poder tener acceso a información transaccional, demográfica y además un indicador geográfico hacen más robusto una estrategia y sus resultados.

Capítulo 6

6. Business intelligence (BI)

Resumen: El continuo desarrollo de la tecnología en el área del almacenamiento de datos y su procesamiento ha generado cambios que hace algunos años parecían un mito; la conectividad, movilidad y representación de información con la mayoría de los sistemas manejadores de bases de datos representan hoy una herramienta para tomar decisiones asertivas y oportunas; gran parte de esas herramientas se conocen hoy como Business Intelligence; en este Capítulo se presentan sus antecedentes, definición, componentes y desarrollo; sus tendencias a conformar herramientas y sistemas multimedia que permitan representar indicadores e información diversa de grandes sistemas de almacenamiento; consolidando nuevas aplicaciones como aprendizaje en línea (*e-learning*), comercio electrónico (*e-commerce*), tableros de control, blogs, agendas personalizadas, entre otras.

Palabras clave: Business Intelligence, tableros de control (dashboard), scorecard, e-learning, e-commerce.

Aunque la traducción es “Inteligencia de negocios” se usará el término Business Intelligence (BI), ya que es comúnmente adoptado por el entorno profesional y por el académico lo cual resulta de gran ayuda para su investigación, búsqueda de referencias y material complementario.

Considerando que el término “Business intelligence” ha tenido una gran aceptación y comercialización en los últimos años no se puede considerar reciente ya que desde los 60´s se viene manejando bajo diferentes estructuras, por ejemplo antes las computadoras centrales (mainframes), eran los receptáculos de datos de donde se extraían grandes cantidades de datos para su análisis, estos procedimientos eran costosos y muy tardados además de ser programadores expertos quienes lo realizaban, por lo cual la información no era oportuna y para la gente indicada.

El auge del business intelligence es a partir de la década de los 90´s con la comercialización de numerosas aplicaciones por parte de las principales compañías desarrolladoras de software de base de datos con el objetivo de generar información a partir de datos almacenados y dirigido principalmente a gente encargada de tomar decisiones; estas se caracterizan por ser aplicaciones de fácil acceso, con herramientas visuales sobre grandes cantidades de datos y ejecución en tiempo real.

6.1 Definición business intelligence

Business intelligence (BI) es un concepto que integra los procesos, herramientas, y tecnología para convertir datos en información, información en conocimiento y conocimiento como soporte de decisiones para conducir de forma eficaz las actividades de una organización.

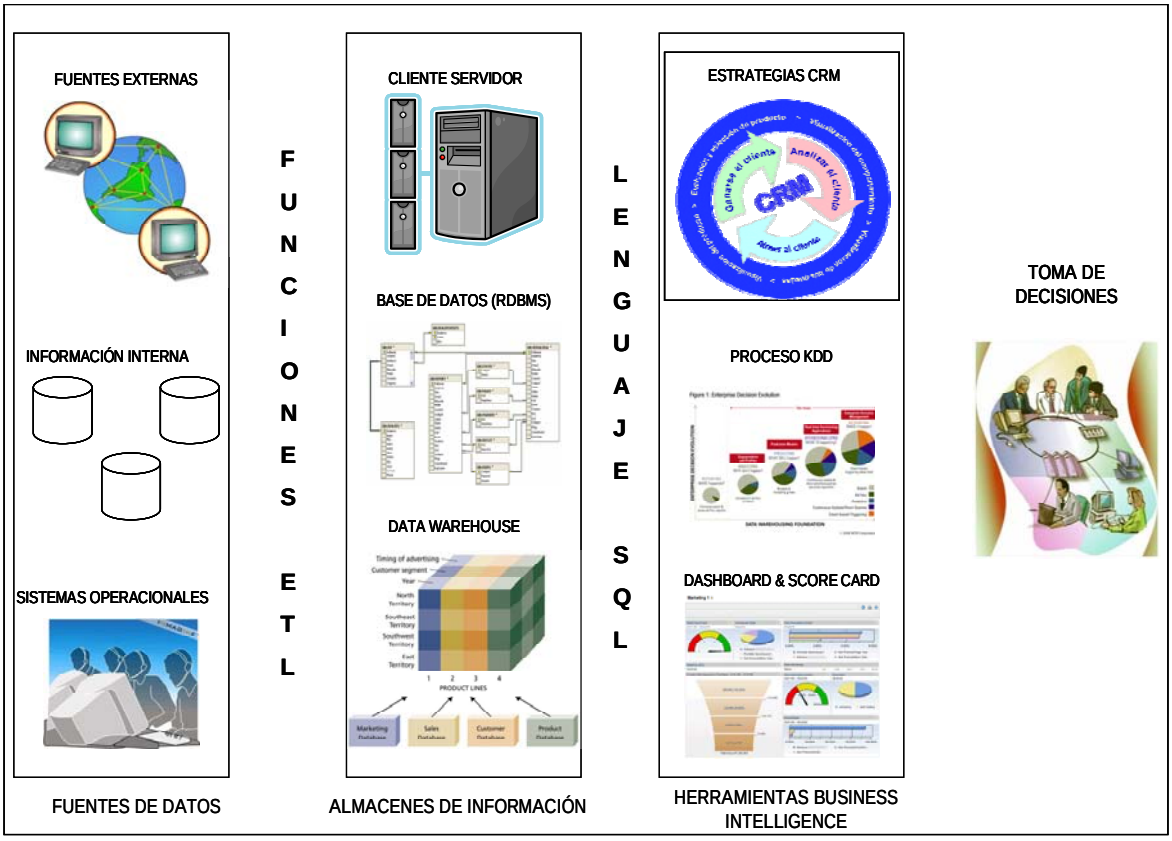
"El uso de información que permite a las organizaciones dirigir de la mejor forma, decidir, medir, gestionar y optimizar el alcance de la eficiencia y los resultados financieros."

"Business intelligence (BI) es un término paraguas que abarca los procesos, las herramientas, y las tecnologías para convertir datos en información, información en conocimiento y planes para conducir de forma eficaz las actividades de los negocios. BI abarca las tecnologías de datawarehousing los procesos en el 'back end10', consultas, informes, análisis y las herramientas para mostrar información (herramientas de BI) y los procesos en el '*front end*", según: [Eckerson; 2005]

Existen diversas definiciones de BI pero todas convergen a un factor determinante, el uso de los datos a través de diversos procesos y herramientas, permitiendo obtener información de problemas y hechos que intervengan directa o indirectamente la efectividad de los resultados de una organización.

6.2 Componentes del business intelligence

El trabajo ha presentado estructuras y componentes genéricos, con el objetivo de explicar el funcionamiento y la interacción de procesos herramientas y aplicaciones; el BI es conceptualmente un Proceso de integración. A continuación se presenta un esquema genérico del concepto y herramientas de business litelligence.



- Fuentes de datos: Los datos se pueden generar desde sistemas operacionales, procesos internos y externos un data warehouse, o extractos de bases de datos.
- Sistemas operacionales: Son aquellos sistemas o puntos de origen de datos de una organización; pueden ser un punto de venta, un *call center*, un sitio Web, etc.

- Procesos o fuentes internas: Son aquellos procesos que representan una referencia del funcionamiento, desarrollo, planeación y control de una organización, por ejemplo el análisis de desempeño por punto de venta, por zona, por empleado, o un inventario.
- Fuentes externas: Principalmente indicadores sobre la organización sobre su participación en el mercado, la competencia, tendencias, por ejemplo estudios de mercado, censos, encuestas, datos sociodemográficos de sectores de la población, etc.
- Diseño e implementación de almacenes de información: Quizás representa el proceso y/o tarea de mayor desafío para las organizaciones, y que involucra decisiones en todas las áreas de la organización.
- Tecnología: Generalmente siempre limitada por cuestión económica debe considerar la proyección del sistema de información, su horizonte de vida, seguridad, conectividad, plataforma concurrencia y uso con el objetivo de tener un impacto positivo y de apertura para la organización.
- Modelo de datos: La representación de un problema y/o situación operativa de una organización es la parte medular del concepto BI y se debe hacer un análisis exhaustivo para considerar todas las posibles variables y hechos que involucre la organización.
- Base de datos (data warehouse): Una base de datos o un data warehouse dentro del concepto de BI es la parte central y la fuente de los datos.
- Funciones ETL (Procesos de extracción transformación y carga de datos): Consisten en procesos de selección, limpieza y transformación de los datos, técnicamente es lo mismo que se presento en el Capítulo 3 respecto a la carga de los datos a un data warehouse un ejemplo pueden ser extracciones de datos periódicamente para hacer reportes, pronósticos, análisis de inventarios etc.
- Herramientas de BI: Representan un panorama general y abstracto sobre procesos complejos que involucran toda la operación y sus principales indicadores sobre una organización.

- Descubrimiento de conocimiento: El resultado de procesos y uso de herramientas representa un número o números, un porcentaje de perdida o ganancia, perfiles de usuarios, clientes, prospectos; principalmente indicadores claves en la planeación, desarrollo, situación financiera y pronósticos de una organización.
- Herramientas de visualización: Principalmente gráficas de diversos tipos, a diferencia de una gráfica común estas herramientas pueden representar una cantidad masiva de datos en tiempo real y bajo diferentes condiciones, dimensiones y escalas. Otra herramienta de mucho uso son los árboles de decisión que asocian probabilidades y decisiones a diferentes escenarios y estrategias.
- *Dashboards*: También conocidos como tableros de control, son diversos reportes consolidados y personalizados en uno solo, principalmente representan indicadores relacionados a un proceso o área en particular y pueden ser conteos sobre datos, distribuciones y frecuencias poblacionales, flujos y comportamientos entre otras; de gran uso por su potencial para determinar desviaciones o errores en procesos entre otras funcionalidades.
- *Scorecards* (tarjetas de puntuación): Pueden ser sistemas de información de una población, variables demográficas, modelos estadísticos que evalúen comportamiento, inclusive lealtad, de gran uso en el área de crédito y cobranza principalmente en el sector bancario y de seguros, representa una herramienta potencial para determinar riesgo y características poblacionales.

6.3 ¿Qué representa el business intelligence en los actuales manejadores de bases de datos comerciales?

Es tal el desarrollo del business intelligence que prácticamente todos los DBMS comerciales cuentan con módulos o herramientas que permiten análisis sobre los datos almacenados.

En los 90´s un DBMS cumplía con aplicaciones estándar sobre administración y gestión de los datos, con funciones estadísticas descriptivas y orientadas a expertos con conocimientos técnicos avanzados. En la actualidad la gran mayoría de los DBMS ofrece una serie de aplicaciones de análisis sobre los datos, implementando técnicas estadísticas complejas con una interfaz gráfica y con herramientas de visualización, orientados a la consulta y de respuesta inmediata sobre gran variedad de datos y un horizonte de tiempo tan largo como se tenga almacenado. Herramientas de visualización, multiplataforma, sobre lenguaje de alto nivel, son parte del desarrollo del business intelligence y prácticamente son parte de los componentes principales de todos los DBMS.

6.4 Tendencias del business intelligence en el mercado

Definitivamente las expectativas del business intelligence son de gran potencial sobre su desarrollo. La información y conocimiento que tiene como origen los datos deben ser el eje de cualquier organización respecto a su futuro y operación.

El reto de los principales desarrolladores de herramientas de business intelligence se encuentra en implementar herramientas de análisis que permitan combinar e interactuar con fuentes de información estructuradas y no estructuradas; y por no estructuradas se hace referencia a la gama de canales que brinda Internet como son los blogs´, chat, e-mail, comunidades, archivos multimedia entre las más representativas.

La explotación de sistemas de almacenamiento con herramientas de visualización sobre relaciones y propiedades de grandes cantidades de datos como son los dashboards & scorecards, que permiten obtener información de toda la organización a detalle en tiempo real son el futuro a corto plazo.

Otro factor determinante es el desarrollo de programas de uso libre como sucedió en los 90´s con el desarrollo de distintos DBMS con gran parte de las herramientas y usos potenciales respecto a los comerciales.

La integración de los procesos y herramientas con el desarrollo tecnológico deben brindar una opción de herramientas de business intelligence de acceso casi a cualquier persona de una organización o de manera personal para llevar una administración y gestión de sus principales actividades.

Se deben considerar como nuevos mercados al e-business, el e-learning basados en la comunicación remota y en la interacción electrónica de información, servicios y objetos; todo ello basado en sistemas de almacenamiento de datos, y explotados con herramientas de business intelligence bajo un enfoque global sencillo y estructurado.

Capítulo 7

7 Desarrollo de una base de datos con postgreSQL para la aplicación del proceso de descubrimiento en bases de datos (*knowledge discovery in databases, KDD*)

Resumen: Se presentan de manera práctica los diferentes conceptos que se han mencionado en el desarrollo de este trabajo; se implementa un sistema de almacenamiento de datos que desarrolla metodologías y técnicas explicadas; mostrando la obtención de información útil y previamente desconocida a través del proceso KDD, todo esto apoyado con aplicaciones de uso libre.

Se investigó sobre un tema de interés general como son condiciones de bienestar básicamente sociales y económicas a nivel estatal de la República Mexicana; con el objetivo de que los indicadores, resultados y conocimiento obtenido sean de interés general y muestren el potencial de la información y su explotación en una organización.

7.1 Justificación

La principal expectativa del trabajo es que sea útil, y que cualquier persona pueda interpretar sus resultados por lo cual se buscó un tema de interés general. Se decidió sobre condiciones de bienestar de la República Mexicana, considerando indicadores sociales y económicos principalmente.

7.2 Planteamiento del problema

En la actualidad la población y la sociedad Mexicana sufren cambios en general, los cuales empíricamente los atribuimos a factores como la globalización, la economía, la inseguridad e incluso factores políticos, por lo cual surgió la expectativa de generar un conocimiento con más detalle de estos indicadores poblacionales, económicos y sociales.

El Consejo Nacional de Población (CONAPO) realizó una proyección sobre indicadores básicos referentes a desarrollo social de la República Mexicana a nivel estatal considerando el periodo de 1990 a 2030; estos indicadores consideran aspectos de la población como: nacimientos, defunciones, fecundidad, crecimiento social entre otros, por año de 1990 a 2030. También CONAPO realizó una estimación de "Índices de marginación de México 2005" basado en el II conteo de población y vivienda realizado en 2005 por el INEGI y en la "Encuesta Nacional de ocupación y empleo 2005".

Por otra parte el Instituto Nacional de Estadística Geografía e Informática (INEGI), desarrolló en 2004 el producto "Regiones socioeconómicas de México", a partir del XII Censo Nacional del año 2000; este es un producto de altas dimensiones ya que cuenta con información diversa a nivel nacional, entidad, municipios y AGEB* y que su principal objetivo es presentar diferencias y similitudes respecto a indicadores relacionados con el bienestar de la población que son de manera general: ocupación, educación, salud, vivienda y empleo. "Para ello se forman siete estratos (distintos entre sí), donde los elementos clasificados en un mismo grupo tienen en promedio características similares, es decir, son homogéneos. Los estratos se ordenan de tal forma que en el estrato 7 se encuentran las entidades federativas que respecto al total de indicadores considerados presentan en promedio la situación relativa más favorable, por el contrario, el estrato 1 se compone de las unidades que en promedio presentan la situación relativa menos favorable".

Para los objetivos del trabajo sólo se realizaron las extracciones por estado si se desea consultar a más detalle por municipio o AGEB, según sea el caso se puede consultar el sitio WEB del INEGI: www.inegi.org.mx

Recapitulando se cuenta con:

- Proyecciones desde 1990 hasta 2030 (CONAPO)
- Estratificación basada en indicadores de bienestar que representa la situación de cada estado respecto a la República Mexicana (XII censo INEGI, 2000)
- Indicadores de marginación correspondientes al año 2005 (II Conteo de población y vivienda, CONAPO)

7.3 Objetivo

Encontrar información que ha determinado los principales cambios en la población, sociedad y economía respecto a las estimaciones e indicadores generados por CONAPO e INEGI respectivamente

7.4 Usos y alcances

Con información tan diversa, este proyecto puede tener grandes alcances y potencial para desarrollo en diferentes organizaciones tanto a nivel público como privado.

A continuación se mencionan sólo algunas de las posibles líneas de aplicación y desarrollo de esta investigación:

- Muestreo a nivel nacional y por entidad.
- Apoyo y/o focalización de programas públicos y políticas sociales.
- Perfiles que permitan saber más acerca del comportamiento y estructura socioeconómica de diferentes regiones de la República Mexicana.
- Niveles e índices de desempleo, salud, educación, vivienda, delincuencia; patrones y tendencias.
- Estudios de mercado.

- Escenarios y potencialidad de estrategias comerciales.
- Estudios de información geográfica.

A continuación se definen las diferentes etapas y proceso del trabajo desde el análisis y diseño de un sistema de almacenamiento de datos con postgreSQL, la extracción y carga de los datos, su explotación y ciclo de procesamiento para implementar el proceso KDD, el uso y desarrollo de varias técnicas de minería de datos y establecer la existencia de patrones y tendencias que originan el descubrimiento de conocimiento para finalmente evaluar el conocimiento y considerarlo como base para nuevas acciones y quizás como nuevas líneas de investigación.

7.5 Análisis y diseño de un sistema de almacenamiento de datos

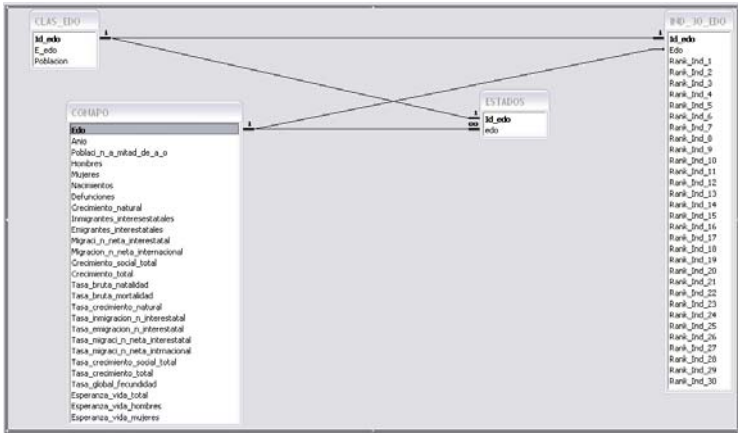
Es el soporte del proyecto ya que de ello depende su desarrollo y finalmente su uso y explotación por lo que se recomienda mantener en todo momento un control documentado que defina las responsabilidades, avances y tiempos de implementación.

El análisis y diseño de un sistema de almacenamiento de datos debe ser exhaustivo y objetivo; en el Capítulo 3 se mencionó la estructura de Zachman como base del diseño de un data warehouse, considerando en una matriz todos los aspectos de una organización que deben ser considerados; a continuación se establece dicha estructura con el sistema a desarrollar.

Regiones socioeconómicas de México	Datos (Qué)	Procesos (Cómo)	Redes (Dónde)	Personas (Quién)	Tiempo (Cuándo)	Motivación (Por qué)
Ámbito (Contexto)	Indicadores de bienestar por estado de la República Mexicana	Extracción y transformación de datos en conocimiento	Sector público y privado	INEGI, CONAPO, dependencias públicas y privadas	Desde el año 1990 con proyecciones hasta 2030	Interés general que pueda ser interpretado y útil
Modelo (Conceptual)	Representar la realidad de los mexicanos	Diagrama entidad-relación y de flujo de procesos	PostgreSQL: arquitectura cliente servidor	INEGI, CONAPO; entre otras	Inidcadores desde 1990 hasta 2030	Hipótesis: Descubrimiento de conocimiento potencialmente útil
Modelo organizacional	Relacionar, comparar y generar indicadores	Integración de las diferentes entidades a nivel nacional	Plataformas y recursos de INEGI, CONAPO, SEP IFE entre otras	Codependencia, e interaccion de las diferentes entidades	Continúo	Consolidar entidades que representen el desarrollo de México
Modelo de sistema (Lógico)	Modelo relacional	Levantamiento, carga y explotación de datos	Cobertura hasta los lugares de mayor marginación	Organizaciones públicas y privadas	Continúo	Representar antecedentes, actualidad y proyecciones
Modelo tecnológico (Físico)	Base de datos	Extracción de diversas fuentes e importación a PostgreSQL	Plataforma Windows	Conocimiento en base de datos	Desarrollo	Control, procesamiento y explotación de datos
Componentes	Entidades (Tablas, catalogos) relaciones (Estratos, proyecciones)	Carga, transformación y procesamiento de datos	PostgreSQL, WEKA, hojas de cálculo	Procesamiento y descubrimiento de conocimiento	Desarrollo	Uso libre y desarrollo intelectual
Sistema Funcional	Datos estructurados, consistentes, validos que representen la realidad	Integración, modelado, validación y comparación de resultados	PostgreSQL, WEKA, Hojas de cálculo	Conocimiento en procesamiento, e interpretación de datos	Desarrollo	Desarrollo intelectual y profesional

7.6 Script en postgreSQL para la creación de un sistema de almacenamiento de datos

Relaciones



Restricciones (*Constrain*s)

TABLA	CONSTRAIN
ESTADOS	CONSTRAINT estados_pkey PRIMARY KEY (id_edo)
CLAS_EDO	CONSTRAINT clas_edo_id_edo_fkey FOREIGN KEY (id_edo) REFERENCES estados (id_edo) MATCH SIMPLE ON UPDATE NO ACTION ON DELETE NO ACTION
CONAPO	CONSTRAINT conapo_pkey PRIMARY KEY (Edo)REFERENCE estados_

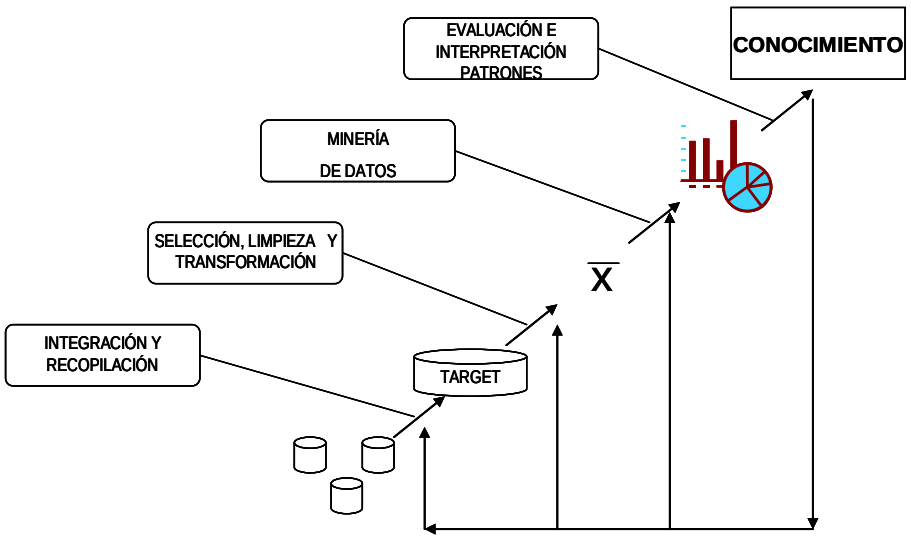
7.7 Proceso de descubrimiento de conocimiento en bases de datos (*knowledge discovery in databases*, KDD)

El proceso KDD busca obtener información previamente desconocida y potencialmente útil, en este caso particular sobre las estimaciones e indicadores de CONAPO e INEGI; no se pretende cuestionar los resultados sino encontrar información que represente una herramienta alternativa para comparar y entender los resultados sobre la situación poblacional, económica y social de cada estado de la República Mexicana.

Objetivo: Encontrar información que permita entender la situación poblacional económica y social de los estados de la República Mexicana, considerando los resultados del producto "Regiones socioeconómicas de México" (INEGI; 2004), las proyecciones de 1990 a 2030 realizadas por CONAPO y las estimaciones de "Índices de marginación de México 2005", realizadas por CONAPO en base al II conteo de población y vivienda realizado por el INEGI en 2005.

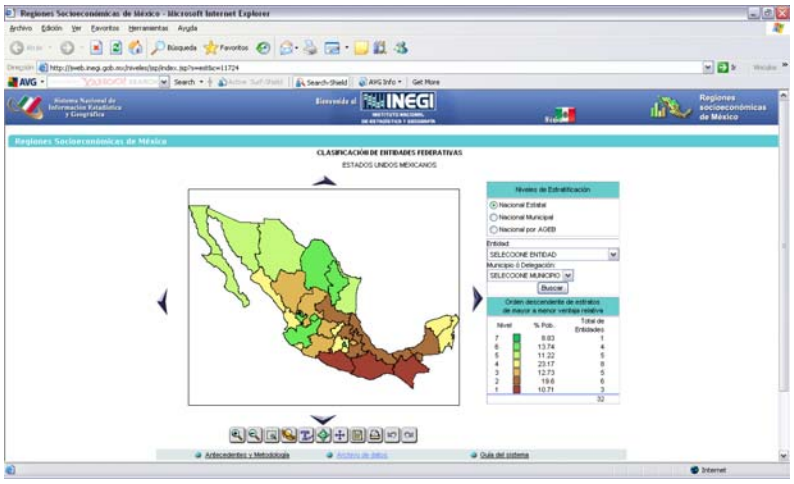
A continuación se presenta un esquema general de etapas del proceso de descubrimiento de conocimiento en bases de datos (KDD), que es el proceso que se desarrolló sobre los datos de INEGI y CONAPO.

PROCESO GENÉRICO DE DESCUBRIMIENTO DE CONOCIMIENTO
KNOWLEDGE DISCOVERY IN DATABASES (KDD)



7.7.1 Recopilación e integración

El proceso de recopilación consistió en hacer extracciones de datos de los sitios Web de cada institución, en este caso el INEGI y CONAPO. Los datos estaban por estado y en formato Excel por lo cual se tuvieron que realizar procedimientos almacenados para automatizar la importación y el procesamiento de los datos. Finalizando con tablas relacionadas por el campo ID_EDO (Identificador de estado), como lo muestra el modelo de datos, teniendo datos a nivel estatal y por año correspondiente.

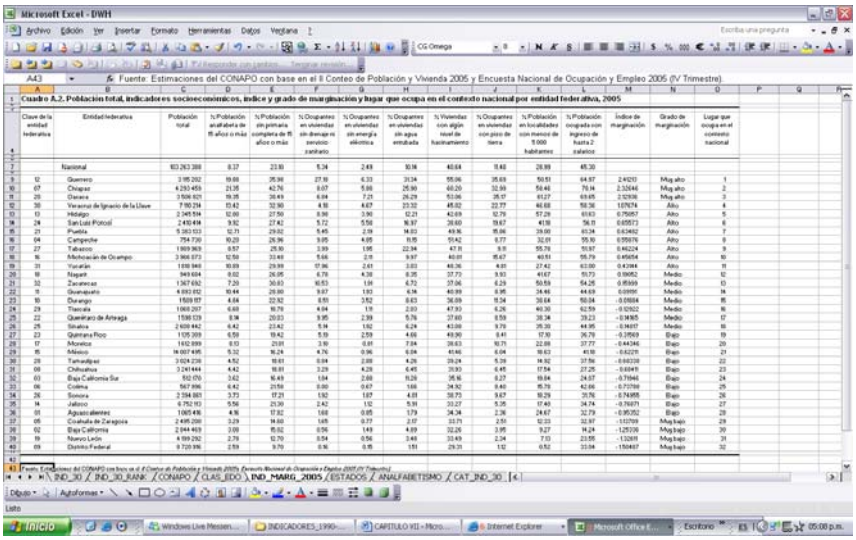


Esta imagen corresponde al sitio Web del INEGI del producto “Regiones socioeconómicas de México” (http://www.inegi.org.mx/est/contenidos/espanol/sistemas/regsoc/default.asp?c=5688), de donde se obtuvieron datos por estado para generar una tabla llamada CLAS_EDO, la cual tiene el identificador del estado, su población a mitad de año según el XII censo 2000 y el estrato al que corresponde según la estratificación de “Regiones socioeconómicas de México”

Los datos están disponibles a nivel estatal, municipal y por AGEB; en este caso se hicieron las extracciones por estado.

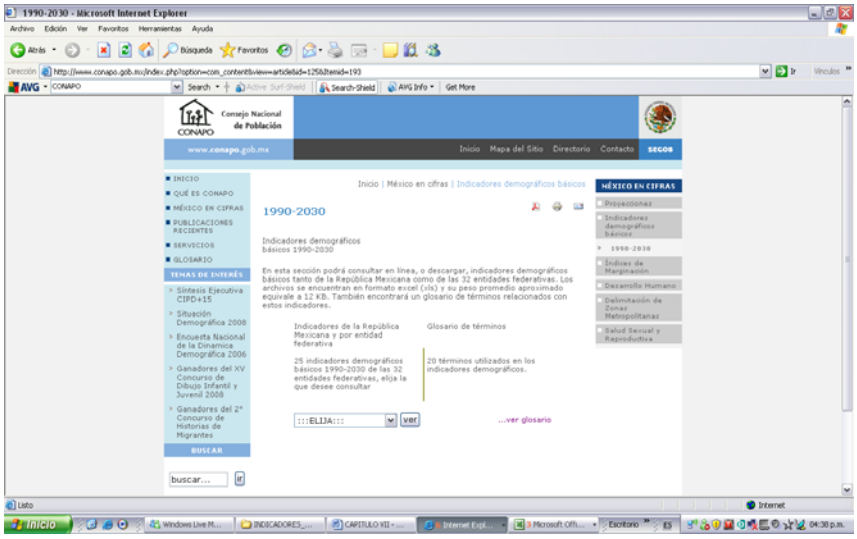
Otra fuente de datos fue el sitio Web del Consejo Nacional de Población (CONAPO) www.conapo.gob.mx, en este sitio se cuenta con una gran cantidad de investigaciones, indicadores proyecciones entre otras publicaciones de México.

La imagen muestra la fuente de la extracción de estimaciones realizadas por CONAPO de la República Mexicana llamadas “Índices de marginación de México” del año 2005 basadas en el II conteo de población y vivienda realizado por el INEGI en el 2005 y en la encuesta nacional de ocupación y empleo 2005.

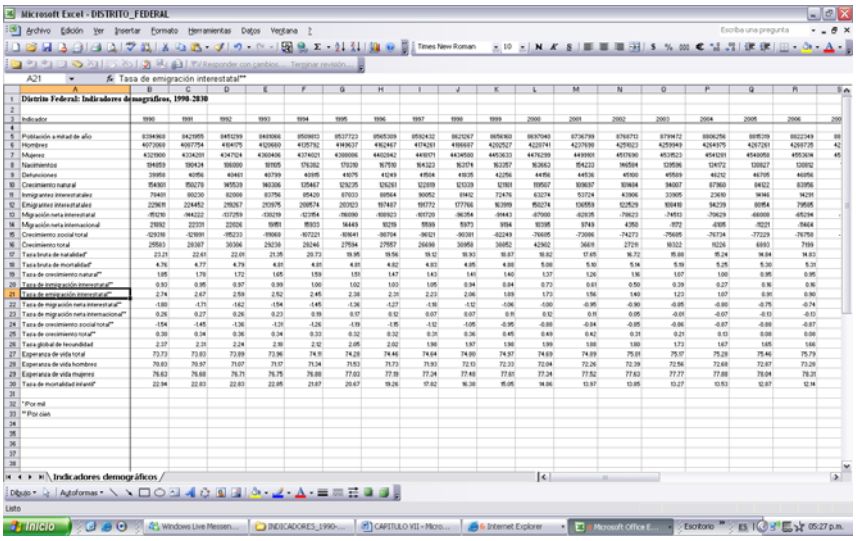


Son datos en formato Excel y en este caso es una sola hoja que contiene los Índices de marginación por estado correspondientes al año 2005

Del mismo sitio de CONAPO también se extrajeron del apartado indicadores-proyecciones una estimación realizada por CONAPO respecto a indicadores demográficos tanto de la Republica Mexicana como por estado que comprende desde 1990 hasta 2030



Como se puede observar los datos están disponibles por estado en formato Excel.



7.7.2 Selección limpieza y transformación

Es importante mencionar que como son reportes o datos que son para el público en general cuentan con un formato, signos de puntuación y presentación que no facilitan su procesamiento con diferentes aplicaciones, en este caso se tuvo que hacer un trabajo exhaustivo de limpieza para eliminar espacios, caracteres especiales y tramas.

En cuanto a la transformación en este caso ya se tiene un nivel de agregación por estado y año según sea el caso; aunque es importante mencionar aspectos importantes para el almacenamiento de datos como pueden ser la recodificación de variables para que no ocupen un espacio innecesario dentro de la base de datos, es este caso un ejemplo claro es el nombre del estado cuyos nombres rebasan incluso los 20 caracteres por lo cual se optó por generar una relación por el campo ID_EDO y ligarla directamente al nombre del estado; también se generaron catálogos para que en lugar de poner la variable por ejemplo "Porcentaje de población en viviendas con agua entubada en el ámbito de la vivienda" quedara: "IND_01".

A continuación se muestra el catalogo correspondiente a la tabla "IND_MARG_2005", y como se muestra en el modelo de datos.

CAMPO	DESCRIPCIÓN	TIPO	TABLA
ID_EDO	CLAVE DE LA ENTIDAD FEDERATIVA (2 DÍGITOS)	CHAR (2)	IND_MARG_2005
EDO	EDO	CHAR (21)	IND_MARG_2005
POB	POBLACIÓN TOTAL	INT8	IND_MARG_2005
IND_01	% POBLACIÓN ANALFABETA DE 15 AÑOS O MÁS	FLOAT4	IND_MARG_2005
IND_02	% POBLACIÓN SIN PRIMARIA COMPLETA DE 15 AÑOS O MÁS	FLOAT4	IND_MARG_2005
IND_03	% OCUPANTES EN VIVIENDAS SIN DRENAJE NI SERVICIO SANITARIO	FLOAT4	IND_MARG_2005
IND_04	% OCUPANTES EN VIVIENDAS SIN ENERGÍA ELÉCTRICA	FLOAT4	IND_MARG_2005
IND_05	% OCUPANTES EN VIVIENDAS SIN AGUA ENTUBADA	FLOAT4	IND_MARG_2005
IND_06	% VIVIENDAS CON ALGÚN NIVEL DE HACINAMIENTO	FLOAT4	IND_MARG_2005
IND_07	% OCUPANTES EN VIVIENDAS CON PISO DE TIERRA	FLOAT4	IND_MARG_2005
IND_08	% POBLACIÓN EN LOCALIDADES CON MENOS DE 5 000 HABITANTES	FLOAT4	IND_MARG_2005
IND_09	% POBLACIÓN OCUPADA CON INGRESO DE HASTA 2 SALARIOS MÍNIMOS	FLOAT4	IND_MARG_2005
IND_10	ÍNDICE DE MARGINACIÓN	FLOAT4	IND_MARG_2005
IND_11	GRADO DE MARGINACIÓN	CHAR (8)	IND_MARG_2005
IND_12	LUGAR QUE OCUPA EN EL CONTEXTO NACIONAL	INT2	IND_MARG_2005

7.7.3 Selección de las variables

El proceso y metodología de "Regiones socioeconómicas", se focalizó a cubrir cinco indicadores principales:

- Vivienda
- Ocupación (Hacinamiento)
- Salud
- Educación
- Empleo

El método utilizado fue la clasificación en base al algoritmo de centros finales que busca formar grupos similares entre sí (homogéneos), y a su ves lo mas distintos posibles entre ellos. La determinación de estos 5 indicadores genero 30 variables que entraron en el análisis clasificadas de la siguiente manera:

CAMPO	VARIABLES REGIONES SOCIOECONÓMICAS DE MÉXICO	INDICADOR
IND_01	% DE POBLACIÓN EN VIVIENDAS CON AGUA ENTUBADA EN EL ÁMBITO DE LA VIVIENDA	INFRAESTRUCTURA DE LA VIVIENDA
IND_02	% DE POBLACIÓN EN VIVIENDAS CON ENERGÍA ELÉCTRICA	INFRAESTRUCTURA DE LA VIVIENDA
IND_03	% DE POBLACIÓN EN VIVIENDAS CON DRENAJE	INFRAESTRUCTURA DE LA VIVIENDA
IND_04	% DE POBLACIÓN EN VIVIENDAS CON PISO DIFERENTE DE TIERRA	CALIDAD DE LA VIVIENDA
IND_05	% DE POBLACIÓN EN VIVIENDAS CON PAREDES DE MATERIALES DURABLES	CALIDAD DE LA VIVIENDA
IND_06	% DE POBLACIÓN EN VIVIENDAS CON TECHOS DE MATERIALES DURABLES	CALIDAD DE LA VIVIENDA
IND_08	% DE POBLACIÓN EN VIVIENDAS CON SERVICIO SANITARIO EXCLUSIVO	EQUIPAMIENTO DE LA VIVIENDA
IND_09	% DE POBLACIÓN EN VIVIENDAS QUE USAN GAS O ELECTRICIDAD PARA COCINAR	EQUIPAMIENTO DE LA VIVIENDA
IND_10	% DE POBLACIÓN EN VIVIENDAS CON REFRIGERADOR	EQUIPAMIENTO DE LA VIVIENDA
IND_11	% DE POBLACIÓN EN VIVIENDAS CON RADIO, RADIOGRABADORA O TELEVISIÓN	EQUIPAMIENTO DE LA VIVIENDA
IND_12	% DE POBLACIÓN EN VIVIENDAS CON TELÉFONO	EQUIPAMIENTO DE LA VIVIENDA
IND_13	% DE POBLACIÓN EN VIVIENDAS CON AUTOMÓVIL O CAMIONETA PROPIOS	EQUIPAMIENTO DE LA VIVIENDA
IND_28	% DE HIJOS SOBREVIVIENTES DE MUJERES DE 20 A 34 AÑOS DE EDAD	SALUD
IND_14	% DE POBLACIÓN CON DERECHOHABIENCIA A SERVICIOS DE SALUD	SALUD
IND_15	% DE POBLACIÓN DE 15 AÑOS Y MÁS ALFABETA	EDUCACIÓN
IND_16	% DE NIÑOS DE 6 A 14 AÑOS QUE ASISTEN A LA ESCUELA	EDUCACIÓN
IND_17	% DE ADOLESCENTES DE 12 A 17 AÑOS QUE ASISTEN A LA ESCUELA	EDUCACIÓN
IND_18	% DE POBLACIÓN DE 15 AÑOS Y MÁS CON INSTRUCCIÓN POSTPRIMARIA	EDUCACIÓN
IND_29	SEGREGACIÓN DE GÉNERO EN TÉRMINOS DE ALFABETISMO	EDUCACIÓN
IND_07	% DE POBLACIÓN EN VIVIENDAS SIN HACINAMIENTO	HACINAMIENTO
IND_19	% DE POBLACIÓN OCUPADA FEMENINA	EMPLEO
IND_20	% DE POBLACIÓN ECONÓMICAMENTE ACTIVA ENTRE 20 Y 49 AÑOS	EMPLEO
IND_21	PERCEPTORES POR CADA 100 PERSONAS	EMPLEO
IND_22	% DE POBLACIÓN OCUPADA QUE PERCIBE MÁS DE DOS Y MEDIO SALARIOS MÍNIMOS	EMPLEO
IND_23	% DE POBLACIÓN OCUPADA QUE PERCIBE MÁS DE CINCO SALARIOS MÍNIMOS	EMPLEO
IND_24	% DE POBLACIÓN EN HOGARES QUE PERCIBEN MÁS DE $10.42 DIARIOS POR PERSONA.	EMPLEO
IND_25	% DE POBLACIÓN OCUPADA QUE SON TRABAJADORES FAMILIARES SIN PAGO	EMPLEO
IND_26	% DE POBLACIÓN OCUPADA EN EL SECTOR TERCIARIO FORMAL	EMPLEO
IND_27	% DE POBLACIÓN OCUPADA QUE SON PROFESIONISTAS O TÉCNICOS	EMPLEO
IND_30	% DE POBLACIÓN ECONÓMICAMENTE INACTIVA DE 65 AÑOS Y MÁS QUE ES JUBILADA O PENSIONADA	EMPLEO

Respecto al desarrollo del estudio realizado por CONAPO “Índices de marginación México 2005” estas son las variables mediante las cuales se construye el “Índice de marginación” son:

VARIABLES "INDICES DE MARGINACIÓN MÉXICO 2005"	INDICADOR
% POBLACIÓN ANALFABETA DE 15 AÑOS O MÁS	EDUCACIÓN
% POBLACIÓN SIN PRIMARIA COMPLETA DE 15 AÑOS O MÁS	EDUCACIÓN
% OCUPANTES EN VIVIENDAS SIN DRENAJE NI SERVICIO SANITARIO	INFRAESTRUCTURA DE LA VIVIENDA
% OCUPANTES EN VIVIENDAS SIN ENERGÍA ELÉCTRICA	INFRAESTRUCTURA DE LA VIVIENDA
% OCUPANTES EN VIVIENDAS SIN AGUA ENTUBADA	INFRAESTRUCTURA DE LA VIVIENDA
% OCUPANTES EN VIVIENDAS CON PISO DE TIERRA	CALIDAD DE LA VIVIENDA
% VIVIENDAS CON ALGÚN NIVEL DE HACINAMIENTO	HACINAMIENTO
% POBLACIÓN OCUPADA CON INGRESO DE HASTA 2 SALARIOS MÍNIMOS	EMPLEO
% POBLACIÓN EN LOCALIDADES CON MENOS DE 5 000 HABITANTES	*POBLACIÓN**

*Como se puede observar para la construcción del índice de marginación se atribuye una variable al volumen poblacional que explica inequidad y falta de asignación de recursos para su desarrollo.

A continuación se presenta un esquema conceptual de la marginación, en donde se establecen las causas principales de este fenómeno de inequidad y exclusión para un grupo de la población.

Como se muestra en el esquema no existen indicadores referentes a la salud; variable que si considera "Regiones socioeconómicas de México", para determinar un "índice de bienestar", favorable o desfavorable en un conjunto poblacional, que bien no es lo mismo que el "índice de marginación", pero que puede ser interpretado de la misma manera.

Respecto a la selección de variables para tratar de explicar la situación poblacional, económica y social de los estados de la República Mexicana son:

- Vivienda (Infraestructura de la vivienda)
- Ocupación (Hacinamiento)
- Educación (Enfocados a condiciones de analfabetismo en individuos mayores a 15 años)
- Empleo

7.8 Algunas técnicas de minería de datos: tendencias y patrones

La minería de datos (DM), ha tenido gran desarrollo en los últimos años, ello ha traído cualquier cantidad de programas y empresas que han desarrollado aplicaciones sobre esta área y una comunidad creciente de usuarios y expertos en el tema; por lo cual es importante mencionar que oficialmente existen dos metodologías para su implementación que son:

- CRISP-DM: Su acrónimo:(*Cross Industry Standard Process for Data Mining*): metodología que es considerada como *standard* ya que no es desarrollada o implementada por una aplicación en particular, y que define el proceso de minería de datos en cuatro grandes fases que son:
    - Comprensión y conocimiento del problema
    - Conocimiento de los datos
    - Modelado
    - Evaluación e implementación del modelo.

- SEMMA: Su acrónimo: (*Sample Explore Modify Model Assesment*), se define como el proceso de selección, exploración, modificar o transformar y modelar de grandes cantidades de datos para descubrir patrones desconocidos. Desarrollada por SAS *Institute* empresa cuya herramienta de minería de datos "*Data Miner*", es una de las mas robustas del mercado.

La siguiente tabla muestra la representación de la minería de datos respecto a sus técnicas y resultados que pueden ser de dos tipos:

- Descriptivo
- Predictivo.

MINERÍA DE DATOS	MODELOS				
	PREDICTIVO (SUPERVISADO)		DESCRIPTIVO (NO SUPERVISADO)		
TÉCNICAS	CLASIFICACIÓN	PREDICCIÓN	AGRUPACIÓN	ASOCIACIÓN	CORRELACIÓN
Redes neuronales	✓	✓	✓		
Árboles de decisión	✓	✓	✓		
Kohonen			✓		
Regresión líneal & logarítmica		✓			✓
Regresión logística	✓				✓
K Means			✓		
Asociaciones				✓	
Análisis factorial					✓
Análisis discriminante		✓			

Continuando con el desarrollo del proceso KDD (Knowledge Discovery in Databases), ya se cuenta con los datos integrados y consistentes y se ha determinado la selección de las variables a estudiar.

A continuación se cita información relevante respecto a ambos productos; “Regiones socioeconómicas de México 2000” e “Índices de marginación México 2005”; tratando de situar en el contexto de los resultados de ambos trabajos y de la información potencial que se obtendrá mediante minería de datos.

Respecto a los resultados de "Regiones socioeconómicas de México", se tiene la siguiente clasificación:

Se consideraron 30 variables que explican cinco indicadores que son: educación, empleo, hacinamiento, salud y vivienda.

La interpretación de la clasificación es la siguiente: se clasifica mejor, a mejor condiciones respecto a las 30 variables, vistas desde el sentido positivo; es decir una variable se interpreta de la siguiente manera: % de la población en viviendas con agua entubada; así se considera que a mayor cantidad de población en viviendas con agua entubada mejores condiciones de bienestar tienen respecto a esa variable.

Clasificación final de "Regiones socioeconómicas de México" basados en el XII censo realizado por el INEGI en el año 2000:

"REGIONES SOCIOECONÓMICAS DE MÉXICO" INEGI XII CENSO, 2000					
ESTRATO	# ESTADOS	ESTADO	POBLACIÓN ESTADO	POBLACION ESTRATO	% ESTRATO
1	3	CHIAPAS OAXACA GUERRERO	3,920,892 3,438,765 3,079,649	10,439,306	10.71%
2	6	VERACRUZ-LLAVE PUEBLA SAN_LUIS_POTOSÍ HIDALGO TABASCO CAMPECHE	6,908,975 5,076,686 2,299,360 2,235,591 1,891,829 690,689	19,103,130	19.60%
3	5	GUANAJUATO MICHOACÁN_DE_OCAMPO DURANGO ZACATECAS TLAXCALA	4,663,032 3,985,667 1,448,661 1,353,610 962,646	12,413,616	12.73%
4	8	MÉXICO SINALOA YUCATÁN MORELOS QUERÉTARO_DE_ARTEAGA NAYARIT QUINTANA_ROO COLIMA	13,096,686 2,536,844 1,658,210 1,555,296 1,404,306 920,185 874,963 542,627	22,589,117	23.17%
5	5	CHIHUAHUA TAMAULIPAS BAJA CALIFORNIA SONORA BAJA CALIFORNIA_SUR	3,052,907 2,753,222 2,487,367 2,216,969 424,041	10,934,506	11.22%
6	4	JALISCO NUEVO_LEÓN COAHUILA_DE_ZARAGOZA AGUASCALIENTES	6,322,002 3,834,141 2,298,070 944,285	13,398,498	13.74%
7	1	DISTRITO_FEDERAL	8,605,239	8,605,239	8.83%
TOTAL	**32**	**REPÚBLICA MEXICANA**	**97,483,412**	**97,483,412**	**100%**

Condiciones poblacionales de México en el año 2000:

- Había 97,483,412 habitantes en el año 2000
- De los habitantes de la República Mexicana en el 2000, al menos 1 de cada 10 vivía en condiciones totalmente desfavorables respecto a vivienda, hacinamiento, educación, salud y empleo.
- El sureste mexicano, presentó la región geográfica del país con las condiciones más desfavorables siendo Chiapas, Oaxaca y Guerrero los estados con la mayoría de la población en esa situación.
- En el 2000, el Distrito Federal con 8,605,239 habitantes representando aproximadamente el 8.83% de toda la población del país fue el estado con mejores condiciones de bienestar en su población.

El "Índice de marginación México 2005", representa una clasificación ordinal respecto al Índice de marginación y su discriminante es básicamente la infraestructura de la vivienda, el empleo, la ocupación (hacinamiento), educación y población en localidades con menos de 5,000 habitantes y tuvo la siguiente clasificación:

"ÍNDICE DE MARGINACIÓN MÉXICO 2005" CONAPO, II CONTEO DE POBLACIÓN Y VIVIENDA, 2005					
ESTRATO	# ESTADOS	ESTADO	POBLACIÓN ESTADO	POBLACION ESTRATO	% ESTRATO
Muy alto	3	GUERRERO CHIAPAS OAXACA	3,115,202 4,293,459 3,506,821	10,915,482	10.57%
Alto	8	VERACRUZ-LLAVE HIDALGO SAN_LUIS_POTOSÍ PUEBLA CAMPECHE TABASCO MICHOACÁN_DE_OCAMPO YUCATÁN	7,110,214 2,345,514 2,410,414 5,383,133 754,730 1,989,969 3,966,073 1,818,948	25,778,995	24.96%
Medio	7	NAYARIT ZACATECAS GUANAJUATO DURANGO TLAXCALA QUERÉTARO_DE_ARTEAGA SINALOA	949,684 1,367,692 4,893,812 1,509,117 1,068,207 1,598,139 2,608,442	13,995,093	13.55%
Bajo	10	QUINTANA_ROO MORELOS MÉXICO TAMAULIPAS CHIHUAHUA BAJA CALIFORNIA_SUR COLIMA SONORA JALISCO AGUASCALIENTES	1,135,309 1,612,899 14,007,495 3,024,238 3,241,444 512,170 567,996 2,394,861 6,752,113 1,065,416	34,313,941	33.23%
Muy bajo	4	COAHUILA_DE_ZARAGOZA BAJA CALIFORNIA NUEVO_LEÓN DISTRITO_FEDERAL	2,495,200 2,844,469 4,199,292 8,720,916	18,259,877	17.68%
TOTAL	32	REPUBLICA MEXICANA	103,263,388	103,263,388	100.00%

Condiciones poblacionales y de marginación de México en el año 2005:

- Había 103,263,388 habitantes en el año 2005
- De toda la población de la República Mexicana en el 2005, al menos 1 de cada 10 vivía en condiciones de marginación "Muy alto"

- El sureste mexicano, presentó la región geográfica del país con las condiciones de marginación calificada como “Muy alto” siendo Chiapas, Oaxaca y Guerrero los estados con la mayoría de la población en esa situación.
- Además del Distrito Federal la región geográfica calificada con índices de marginación “Muy bajos” es el noroeste de la República Mexicana, representando aproximadamente el 17.68% de la población en 2005. el Distrito Federal, Coahuila de Zaragoza, Baja California y Nuevo León.

7.8.1 Clasificación

Considerando que el objetivo es encontrar conocimiento respecto a información ya conocida como son los resultados del producto “Regiones socioeconómicas de México” y el “Índice de marginación México 2005”, se trabajo con un árbol de clasificación (J48), que es una técnica de minería de datos predictiva ya que asigna una clase, pero que también es descriptiva por su estructura o flujo que resulta fácil de interpretar, y mediante el cual se pueden encontrar asociaciones o reglas de las variables; además buscando, relevancia de alguna o algunas de las variables involucradas en los indicadores, diferencias y similitudes que no sean el resultado de la metodología aplicada por INEGI y CONAPO respectivamente, y buscando algún tipo de patrón o tendencia sobre los datos.

Respecto a los datos del XII censo nacional, se observa que el sureste mexicano es la región geográfica con las condiciones menos favorables en cuanto a vivienda, ocupación (hacinamiento), empleo, educación, salud, lo cual genera un nivel de desarrollo social limitado o nulo que se ve reflejado en altos índices de marginación lo cual ratifica CONAPO con los resultados de “Índices de marginación México 2005”.

Con esta información se puede responder a varias preguntas; por ejemplo, considerando las condiciones de marginación:

- ¿Cuáles son las más críticas?

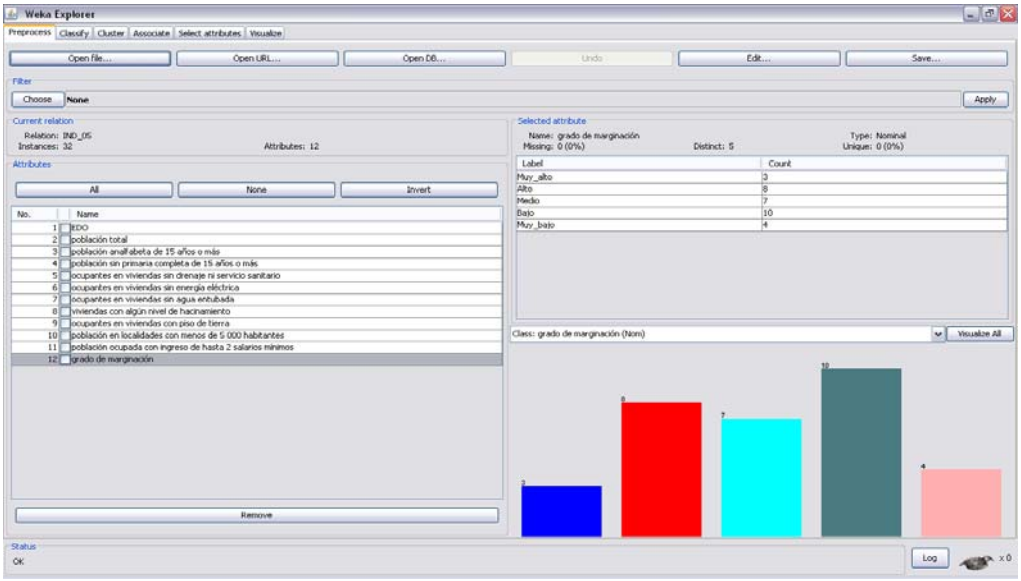
- ¿En que área o sector hay que focalizar acciones o programas que se vean reflejadas en un cambio positivo para esa población?
- Se pueden plantear nuevos programas tanto de prevención, como de acción para evitar marginación.

Mediante un árbol de clasificación se busca obtener información respecto a una variable o variables que sean altamente significativas o discriminantes para establecer una situación de marginación o no; lo cual puede ayudar a identificar estados con esas características y establecer decisiones o programas a seguir según sea el caso.

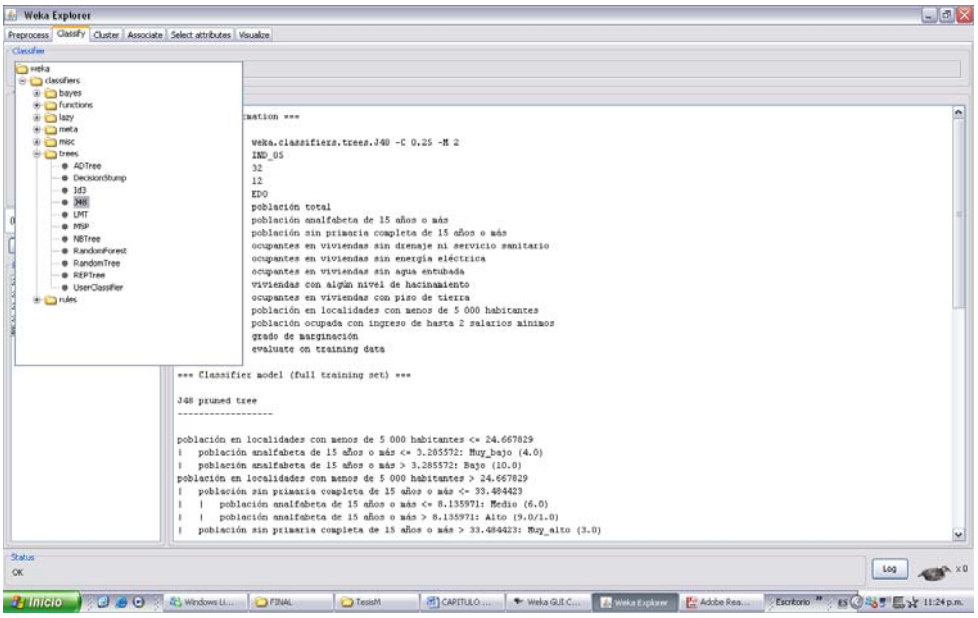
Una de las principales razones sobre usar un método de clasificación es que puede servir como un método predictivo es decir con las diferentes clasificaciones que resulten se puede evaluar otro año o período y además generara información descriptiva que genere información sobre un estado o una región de la República Mexicana, donde se pueden encontrar patrones, condiciones similares y estilos de vida, información geográfica de manera general.

Con el paquete WEKA se cargaron ambas tablas, la tabla de “Regiones socioeconómicas de México” e “Índices de marginación México 2005”

Se puede observar en el apartado de WEKA “preprocess”, el área de trabajo, los campos que integran la tabla, así como algunos de sus principales estadísticos descriptivos y su distribución respecto a la variable seleccionada.



En la siguiente pantalla podemos observar los diferentes métodos de clasificación que tiene WEKA: bayesianos, de reglas, funciones, no supervisados y de árboles entre otros; como ya se mencionó se uso el árbol "J48", que es una técnica de minería de datos supervisada.



El algoritmo J48 es una versión del algoritmo de árboles de clasificación "C4.5" (Quilan), y la razón por la cual se considero este algoritmo es principalmente por que además de obtener conocimiento de las diferentes variables también se busca encontrar cuales son las de más peso o jerarquía. Es decir mas haya de demostrar la clasificación de cada estado de la República Mexicana respecto a los índices de bienestar o marginación, se busca identificar que los hace considerablemente diferentes.

El método J48 funciona determinando un nodo raíz, el cual tiene asociados todos los elementos de la muestra de entrenamiento y mediante los cuales se determinaran cuales son los nodos que conformaran las diferentes clases buscando una minima variabilidad entre clases, este método es recursivo y una ves que se determina un nodo o "hijo", se realiza el mismo proceso para cada rama determinando nuevas ramas o "hijos" y terminando el proceso cuando cada hoja tenga la misma clase.

Las ventajas de este método es que es muy fácil de interpretar admite datos de tipo nominal y numéricos, su criterio de división es principalmente obtener información que no necesariamente representa una regla o relación explicita, por lo cual para este trabajo será de gran utilidad para encontrar reglas, asociaciones, diferencias, relaciones y sobre todo conocimiento sobre todos los indicadores involucrados en la situación poblacional, económica y social de la República Mexicana.

Como se puede observar en la siguiente pantalla, al utilizar un algoritmo en WEKA se presenta la siguiente área de trabajo:

[1] Output, presentara los principales resultados del algoritmo, como pueden ser la matriz de confusión, coeficientes de una regresión, error absoluto, error relativo, y la efectividad del algoritmo.

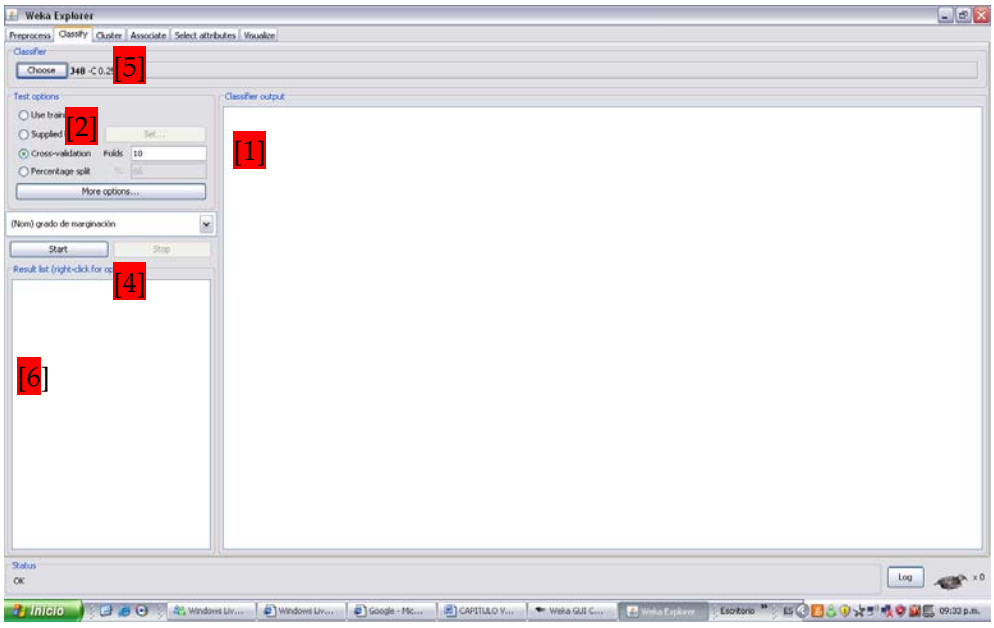
[2] Las opciones de prueba del algoritmo, las cuales se utilizan para construir el algoritmo y posteriormente probar su efectividad.

[3] Opciones de configuración para salida de datos y herramientas de visualización

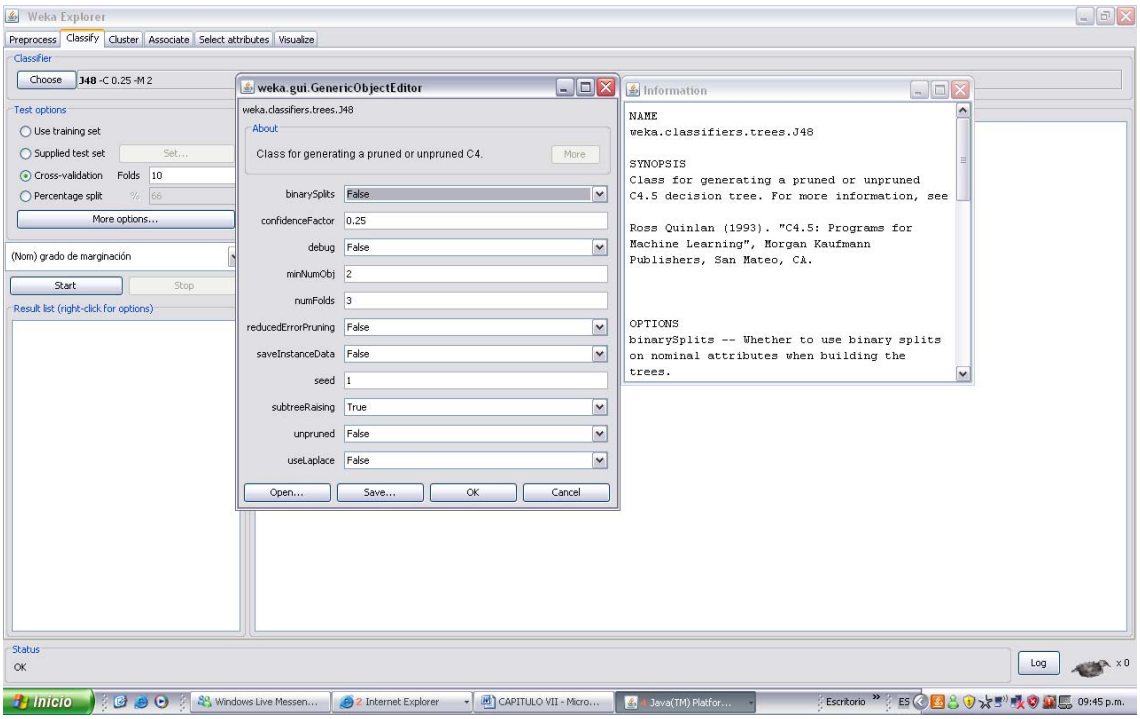
[4] Variable dependiente o clasificatoria.

[5] Los parámetros del algoritmo; los cuales siempre aparecen con valores por default que son los óptimos para el desempeño del algoritmo pero que los puedes ajustar según las necesidades.

[6] El área de resultados mediante los cuales puedes visualizar el árbol, o los diferentes resultados según corresponda a cada algoritmo.



Otra de las funcionalidades de WEKA, es que presenta referencias y una breve reseña del algoritmo, dando click derecho sobre el algoritmo utilizado, lo cual te ayuda a entender los supuestos bajo los cuales se desarrolló su funcionamiento y los parámetros del algoritmo.



Como se mencionó el objetivo no pretende cuestionar los resultados de las clasificaciones realizadas por INEGI y CONAPO, sino encontrar información que explique y ayude a comprender la situación de la República Mexicana en este caso, se comenzara analizar la información correspondiente a "Índices de marginación de México 2005", ya que son sólo 10 variables que se dividen básicamente en cinco indicadores:

- Educación
- Ocupación (Hacinamiento)
- Vivienda
- Empleo
- Localidades aisladas (menos de 5,000 habitantes)

Lo cual es más fácil de analizar e interpretar que el caso de “Regiones socioeconómicas de México 2000”, que maneja 30 indicadores y que se dividen básicamente en cinco indicadores:

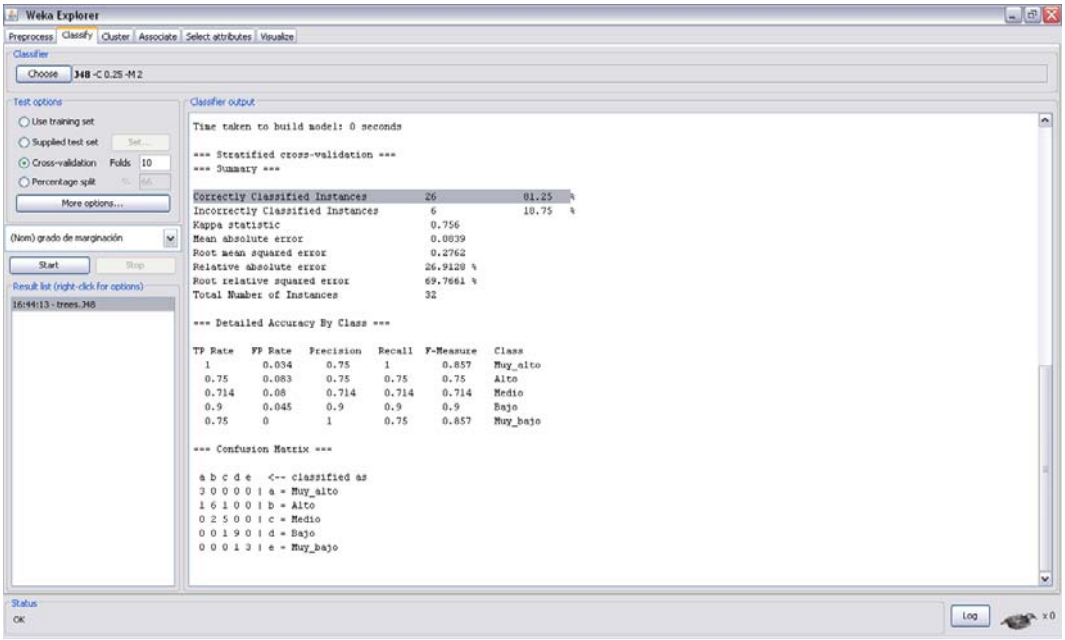
- Educación
- Ocupación (Hacinamiento)
- Vivienda
- Empleo
- Salud

Es importante mencionar que los resultados obtenidos tienen relación, e incluso pueden ser interpretados de la misma manera y complementarios (El índice de marginación puede complementar la región económica a la cual se asignó el estado), pero bajo ningún supuesto son comparables ya que no se habla ni de la misma población (censo vs conteo), ni tampoco de metodologías y técnicas.

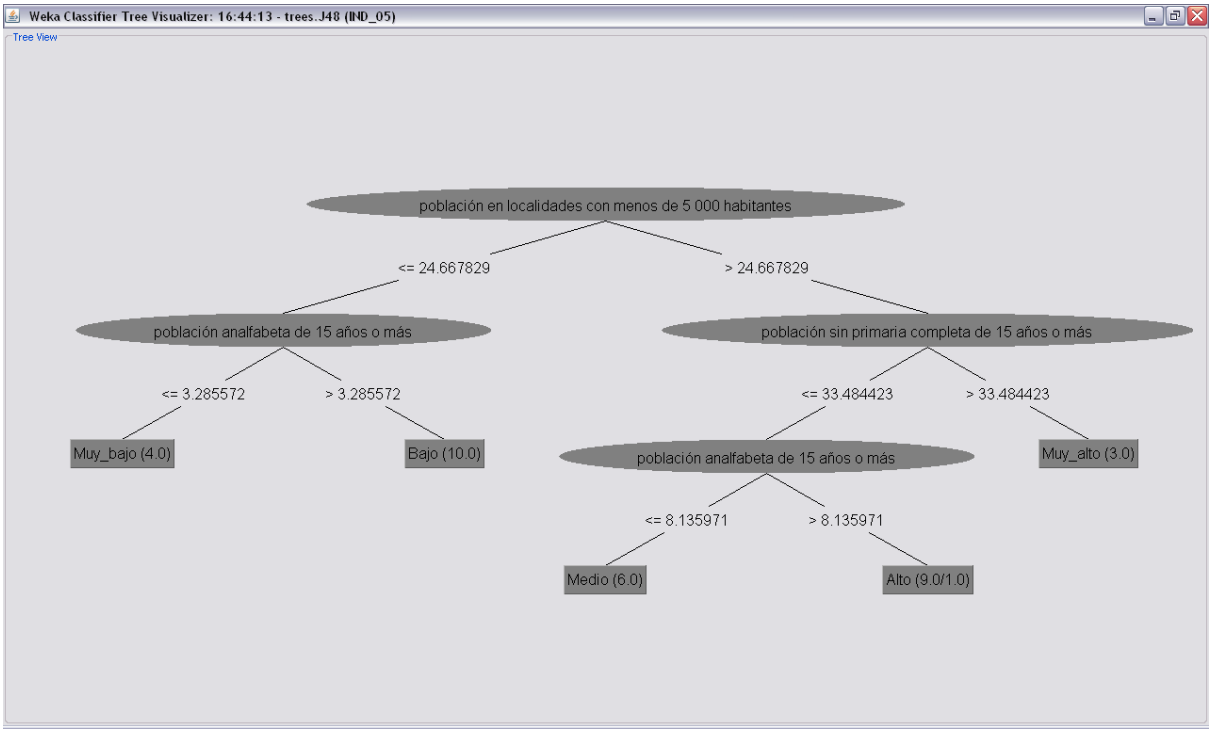
Adicionalmente con los resultados obtenidos de ambas clasificaciones y conclusiones conjuntas se pretenden marcar ciertos escenarios, respecto a la proyección de CONAPO que considera información desde 1990 hasta 2030 y ver si existen desviaciones, o grandes cambios en las condiciones demográficas de la República Mexicana.

Considerando los parámetros por default del algoritmo J48 se corre para los datos correspondientes a la tabla “Índice de marginación 2005”, en este caso se busca predecir el índice de marginación de cada estado de la República Mexicana que esta clasificado como: “Muy alto”, “Alto”, “Medio”, “Bajo” y “Muy bajo” y obtenemos los siguientes resultados:

Observando los resultados se observa que con una validación cruzada de 10 elementos tiene el 81.25%; es decir de los 32 estados de la República Mexicana, se clasifican correctamente 26 y considerando la matriz de confusión como una interpretación de la clasificación (todos los elementos que están en la diagonal principal son las clasificaciones correctas), se observa que los estados con un "Índice de marginación", "Muy alto", los clasifica correctamente, por lo cual si se observa el árbol y analiza esa rama se puede determinar cual es la variable o variables determinantes para que esos estados sean considerados con "Altos", índices de marginación.



La interpretación del árbol de clasificación es muy sencilla y se lee en forma jerárquica (de arriba hacia abajo); buscando la rama de "Muy alto" encontramos que un estado con "Muy alto" índice de marginación tiene más del 24.66% de su población viviendo en localidades con menos de 5,000 habitantes, de los cuales el 33.48% tienen 15 años o más y no cuentan con la primaria completa. Lo cual quiere decir según los datos del II conteo de población y vivienda 2005, que Oaxaca, Chiapas y Guerrero que tienen casi 11,000 millones de habitantes en conjunto tienen al menos 1,000,000 de habitantes en zonas con menos de 5,000 personas, mayores de 15 años que no tienen la primaria terminada y que probablemente no la terminen nunca.



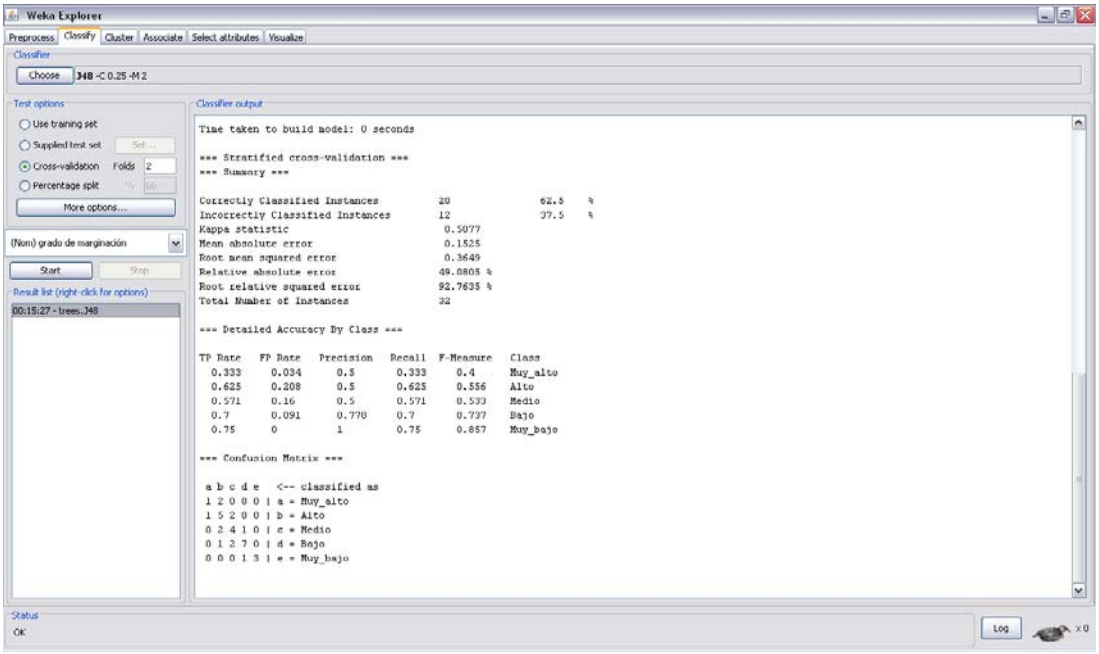
Como se observa el algoritmo clasifica correctamente a los estados con índices de marginación "Muy alto", dada la funcionalidad y automatización de WEKA puedes simular bajo diferentes criterios, filtrando alguna variable o variables, cambiando los parámetros del algoritmo.

Uno de los principales criterios para comparar este algoritmo es variar la muestra de prueba de la clasificación; es decir variar según corresponda a las necesidades la cantidad de evaluaciones que hay que hacer sobre los datos para determinar la clasificación óptima.

Se tiene que para la primera corrida se tuvo una prueba cruzada dividiendo las instancias en 10, considerando que son 32 Instancias (Estados), se tienen que se realizaron 10 evaluaciones teniendo 10 estados como datos de prueba y el resto que son el complemento de 32 como datos de entrenamiento para construir el modelo.

De acuerdo a la primera corrida que dio como resultado el 81.25% de efectividad y mediante el árbol se puede interpretar una diferencia significativa respecto a si se consideran con un índice de marginación ú otro; mas que nada para los de índices de marginación “Muy alto”; lo cual quiere decir que se tienen datos significativamente diferentes para índices de marginación “Alto” y “Muy Alto”, pero que para los demás índices las diferencias son mínimas, por lo cual se puede concluir que si se disminuye el número de evaluaciones se estaría a clasificando con menos precisión. Es decir si el parámetro “*Folds*” disminuye, por ejemplo a 2 esto dividirá la población en 2, 16 estados y 16 estados respectivamente, con una de las divisiones construirá el modelo y con la otra lo probara, lo cual implica que exista más volatilidad y consecuentemente menos precisión en el resultado.

Disminuyendo el número de evaluaciones a dos se tiene lo siguiente:



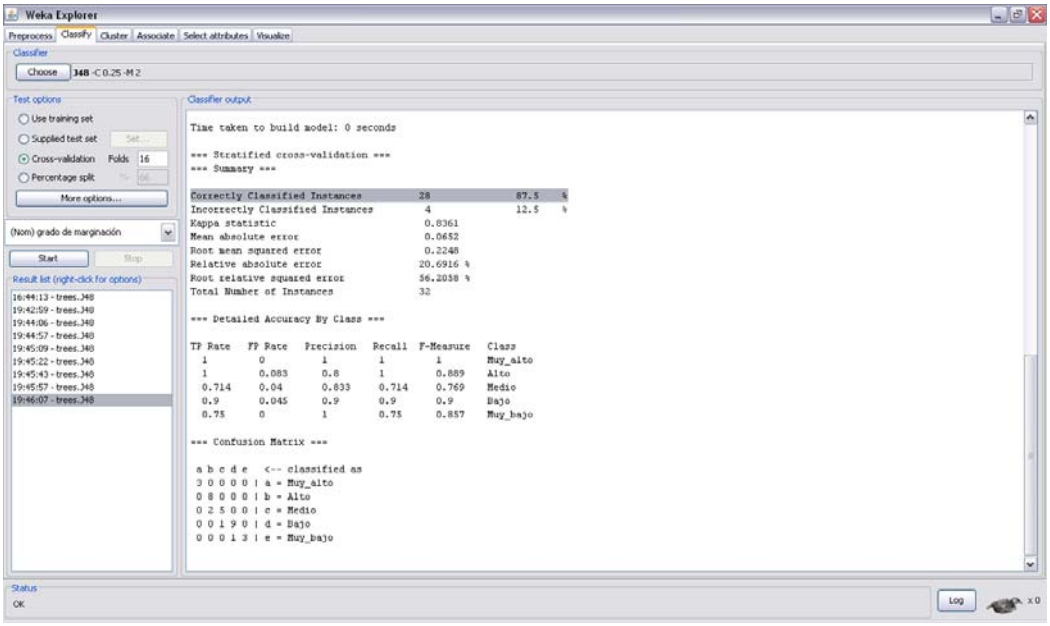
Una efectividad de solo el 62.5%, es decir sólo clasificando correctamente a 20 de los 32 estados.

¿Cómo se pueden obtener mejores resultados del clasificador?

Ya se observó que a menor cantidad de divisiones para probar el clasificador, disminuye su precisión, lo cual indica que hay que aumentar el número de divisiones, dado que son 32 estados, entonces la validación cruzada para 10, 11, 12, esto es dividir la población en 10, 11, 12 estados obteniendo resultados similares ya que se tendría un clasificador con 3 conjuntos de 10-12 estados, y el resto que son entre 22-20 estados para probar la clasificación; por lo cual vale la pena observar el desempeño del algoritmo con 16-17 casos para construir el modelo y los otros 16-15 probando el modelo.

Probando para 16 estados de validación se tiene la división de los 32 estados entre 16 teniendo dos conjuntos uno de 16 estados para construir la clasificación y los otros 16 estados para probarla.

Y se obtiene lo siguiente:

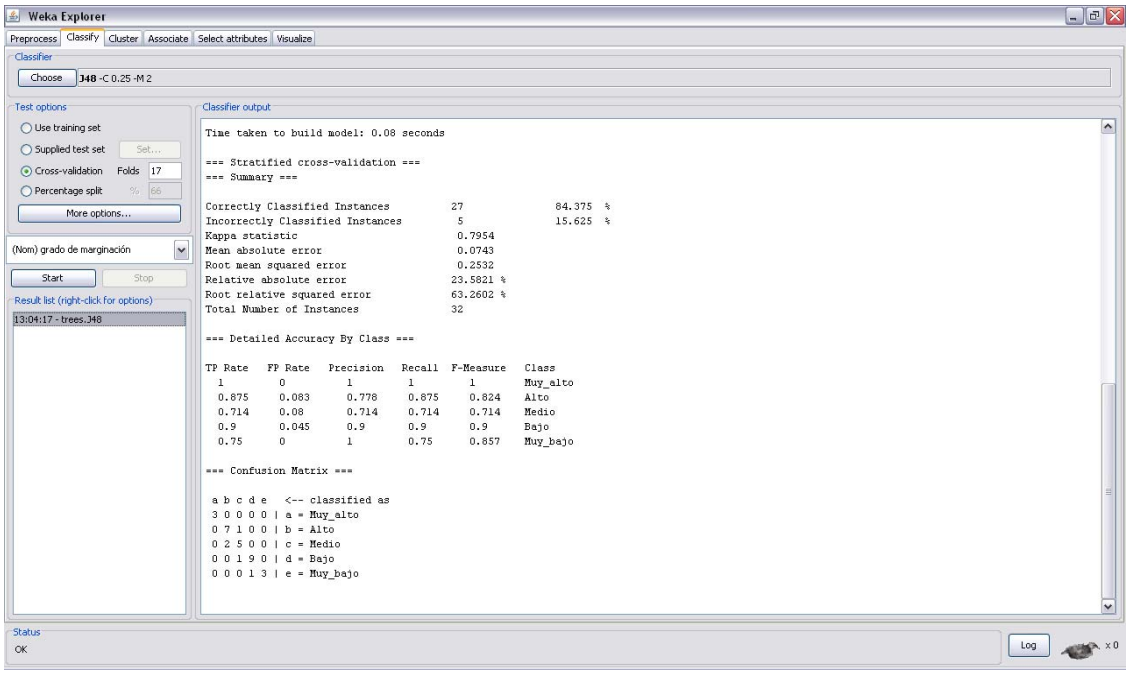


Una efectividad del 87.5%, es decir 28 de 32 estados son clasificados correctamente respecto a su índice de marginación, y el árbol de clasificación es el siguiente

Se puede observar que clasifica correctamente a los estados con “Muy alto” y “Alto” que en total son 11 estados y que clasifica a un estado considerado con índice de marginación “Medio”, como “Alto”.

Para una validación de 17 casos de prueba se tiene la división de la población que es de 32 estados, entre 17 con lo cual se tiene un conjunto de 17 construyendo el modelo y los otros 15 estados probando la clasificación.

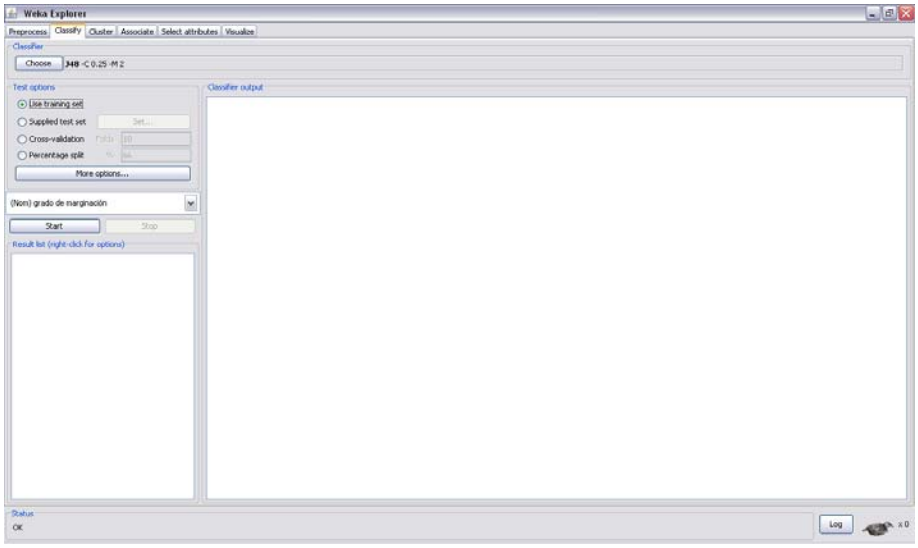
Se obtiene una efectividad del 84.37% es decir clasifica correctamente 27 de los 32 estados. Y respecto a la matriz de confusión se mantiene clasificando correctamente a los estados con "Muy alto" índice de marginación.



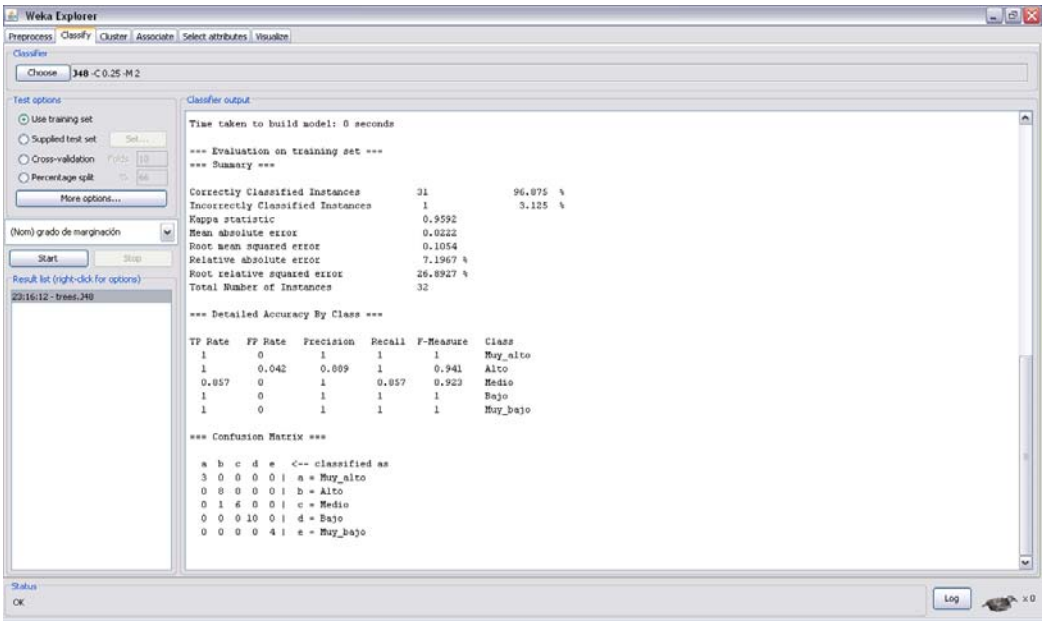
Con lo cual se puede concluir que la clasificación optima para predecir el índice de marginación de cada estado de la República Mexicana mediante el algoritmo J48 es una validación de 16 estados; es decir construir o entrenar el modelo con 16 estados y probarlo con los restantes 16 estados, teniendo como variable discriminante principal:

- % de población en localidades con menos de 5,000 habitantes
- % de población analfabeta de 15 años o más.

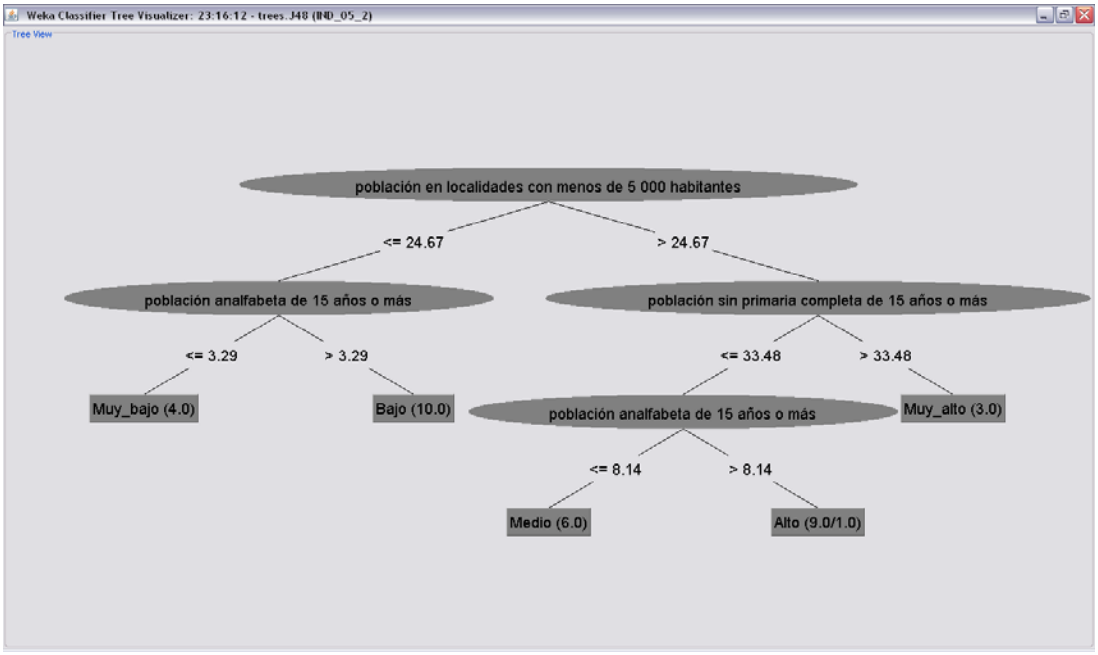
Teniendo un modelo de clasificación basado en el II conteo de población y vivienda realizado por el INEGI en 2005 y la encuesta nacional de ocupación y vivienda, a continuación se muestra la clasificación y su efectividad implementándolo sobre los datos con los cuales se construyó. En este caso como se está probando con los datos con los cuales fue desarrollado en la parte de *test options* (opciones de prueba), se selecciona "*Use training set*"



Las estadísticas e indicadores del algoritmo muestran que se clasificaron 31 estados correctamente lo que representa el 97%. Y como se puede observar en la matriz de confusión el estado que se clasificó incorrectamente, se le clasificó con índice de marginación "Alto", mientras que la clasificación de CONAPO fue de "Medio".



Finalmente se muestra el árbol de clasificación para cada estado respecto a su índice de marginación.

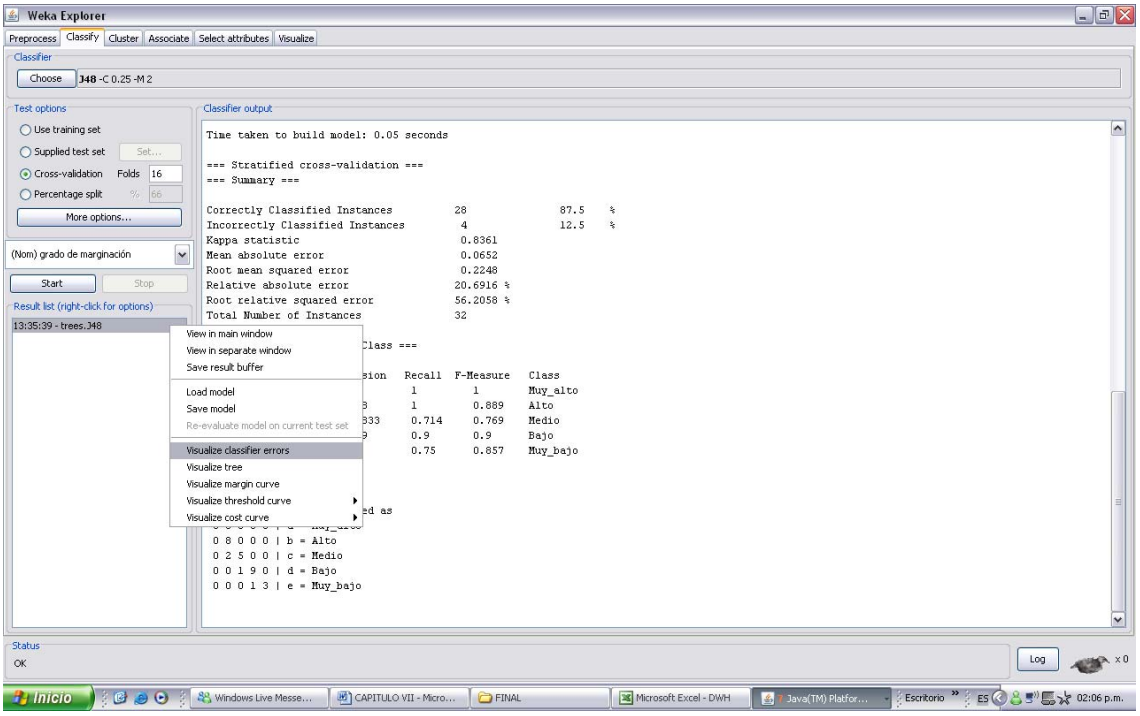


Otra de las grandes funcionalidades que tiene WEKA; es la opción de visualizar la clasificación realizada mediante un ambiente gráfico y cruzar con todas las variables involucradas, lo cual resulta una poderosa herramienta para obtener información adicional de la que genera explícitamente el modelo de clasificación.

A continuación se presenta un análisis del estado que el algoritmo clasificó incorrectamente respecto a CONAPO y su estimación de "Índices de marginación de México 2005".

Estado	Índice de marginación CONAPO 2005	Índice de marginación Clasificación (Árbol)
Guanajuato	Medio	Alto

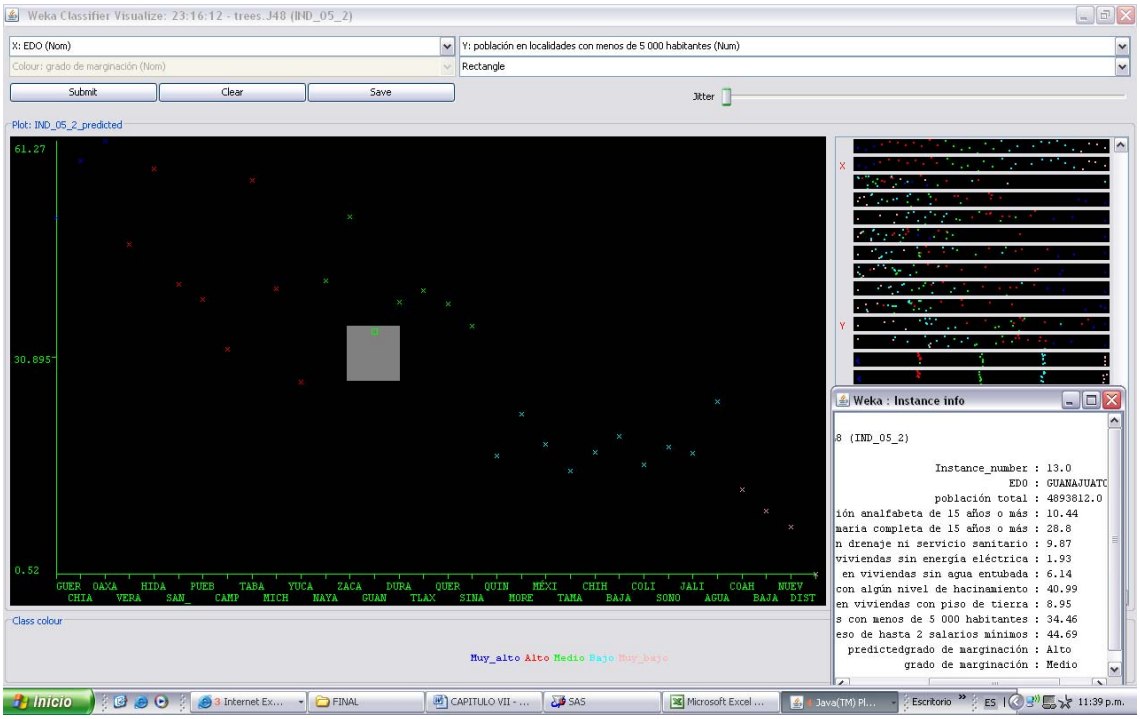
WEKA en el área de resultados (*list results*), con click derecho tiene las siguientes opciones:



Mediante la herramienta "*Visualize*" se puede observar de manera gráfica los errores de la clasificación. Con gran potencial esta herramienta permite observar los diferentes cruces respecto a todas las variables involucradas, mismas que pueden ser seleccionadas para observarlas en el eje X ó Y, también se pueden observar las clases por colores diferentes respecto a la clasificación. En este caso se seleccionan los colores respecto a las clases de la clasificación "Muy alto", "Alto", "Medio", "Bajo","Muy bajo".

Guanajuato

Como se puede observar la gráfica muestra los colores que corresponden a cada clase de la clasificación; en el eje X se presentan los estados y en el eje Y representa la variable "% de población en localidades con menos de 5,000 habitantes" y se observa en la gráfica que los errores de la clasificación se presentan mediante un cuadro y las clasificaciones correctas con una x.

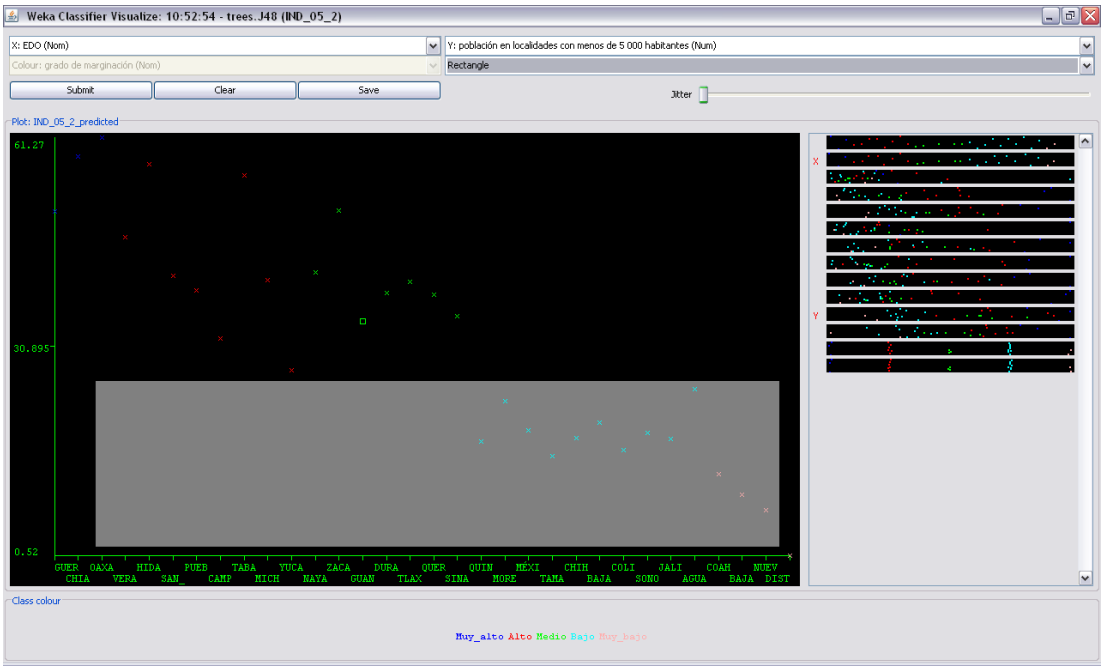


Con click derecho sobre cada instancia se presentan todos sus datos correspondientes, que se muestran en la parte inferior derecha.

En este caso para el estado de Guanajuato se observa la diferencia de clasificación “Medio” a “Alto” y la razón principal es que respecto al modelo presenta:

Características principales de estados con índices de marginación “Medio”	Guanajuato
% de población en localidades con menos de 5,000 habitantes>24.66%	34.46% (Cumple)
% de población sin primaria completa de 15 años o más <=33.48%	28.8% (Cumple)
% de población analfabeta de 15 años o más<=8.14%	10.44% (No cumple)

Además la gráfica genera información muy valiosa, es claro y se observa la relación que existe entre la variable "% de la población en localidades con menos de 5,000 habitantes", y su interpretación donde se puede observar que a menor cantidad de población viviendo en localidades de estas características menor índice de marginación.



Se puede observar que todos los estados que tienen índices de marginación "Medio", "Alto", y "Muy alto" tienen al menos el 30% de su población viviendo en localidades con menos de 5,000 habitantes; naturalmente localidades aisladas no tienen oportunidad de obtener recursos, ni de educación, ni los brindados por el gobierno tanto federal como estatal como pueden ser tuberías, agua, electricidad y educación entre otras.

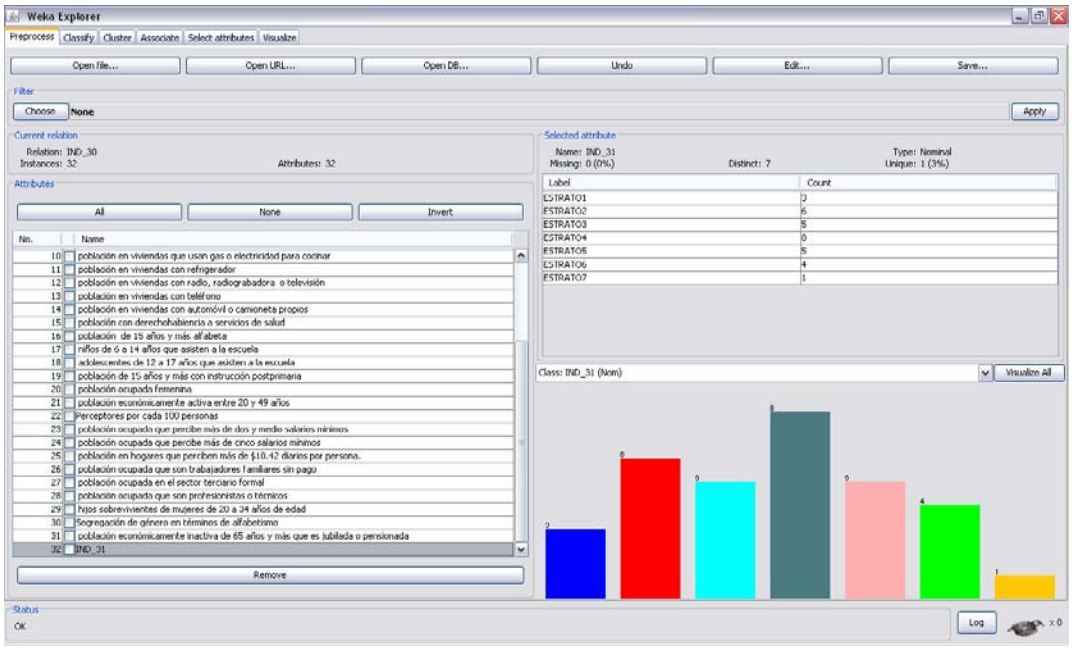
La oportuna detección y control de estas localidades representan áreas de oportunidad para cada estado y sus programas sociales y de desarrollo para que eviten un mayor aislamiento de estas comunidades.

Entonces se puede concluir que los principales factores para determinar un índice de marginación son:

- Aislamiento de gran parte de la población
- Niveles de educación muy bajos ó niveles de analfabetismo muy altos.

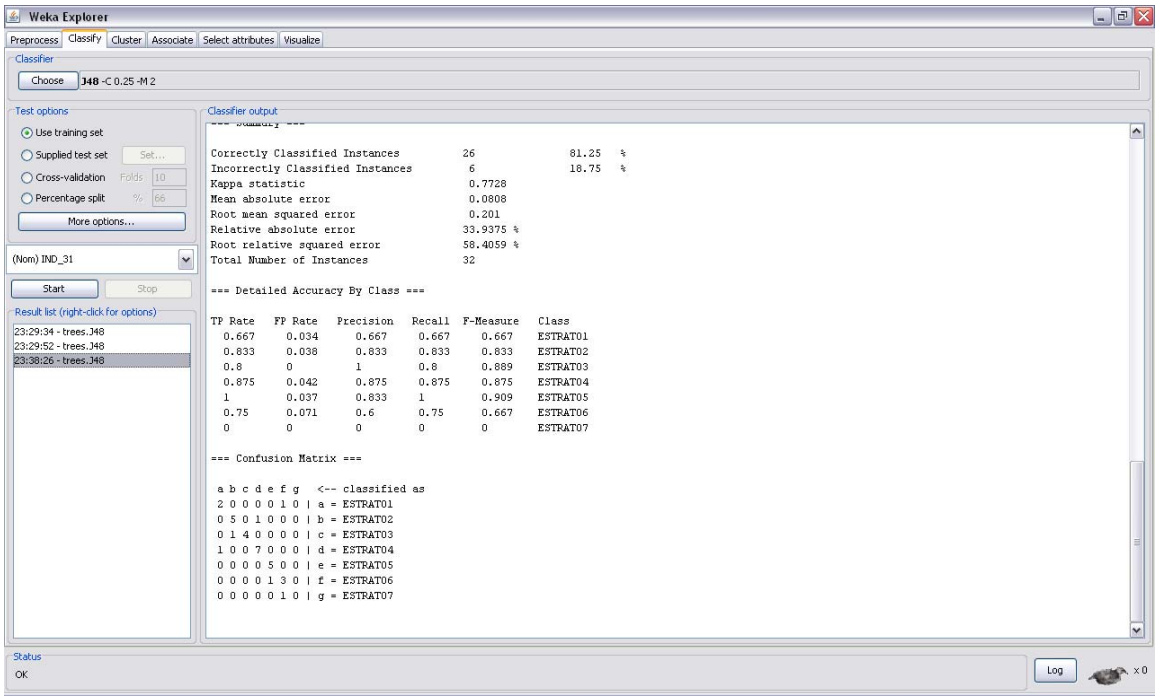
Se realizó el mismo procedimiento para los datos del producto “Regiones socioeconómicas de México 2000”.

La imagen muestra la importación de los datos correspondientes a “Regiones socioeconómicas de México 2000”, se pueden ver los 30 indicadores y la distribución de los 32 estados respecto a las siete regiones determinadas (Estratos considerados por la metodología del INEGI).

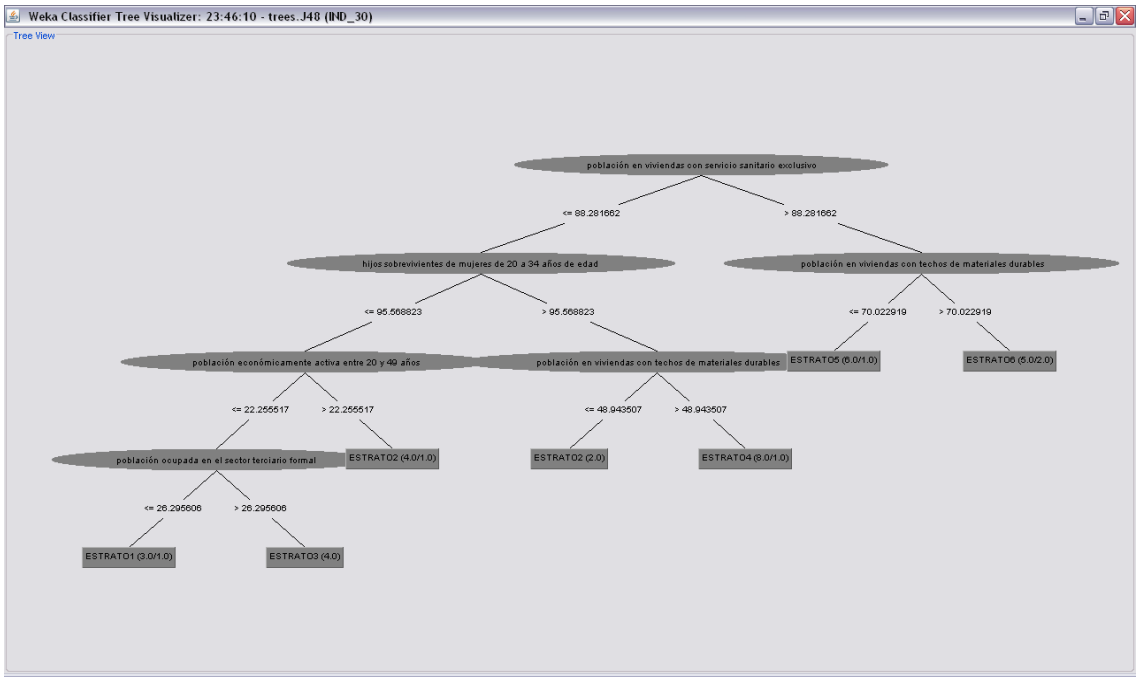


En la pestaña "*Classify*", se carga el algoritmo "J48", y en el área de resultados (*Classifier Output*), presenta los principales indicadores del algoritmo y la matriz de confusión.

Como se puede observar clasifica correctamente a 26 de los 32 estados lo cual representa el 81.25%



Considerando los 30 indicadores este es el árbol resultante.

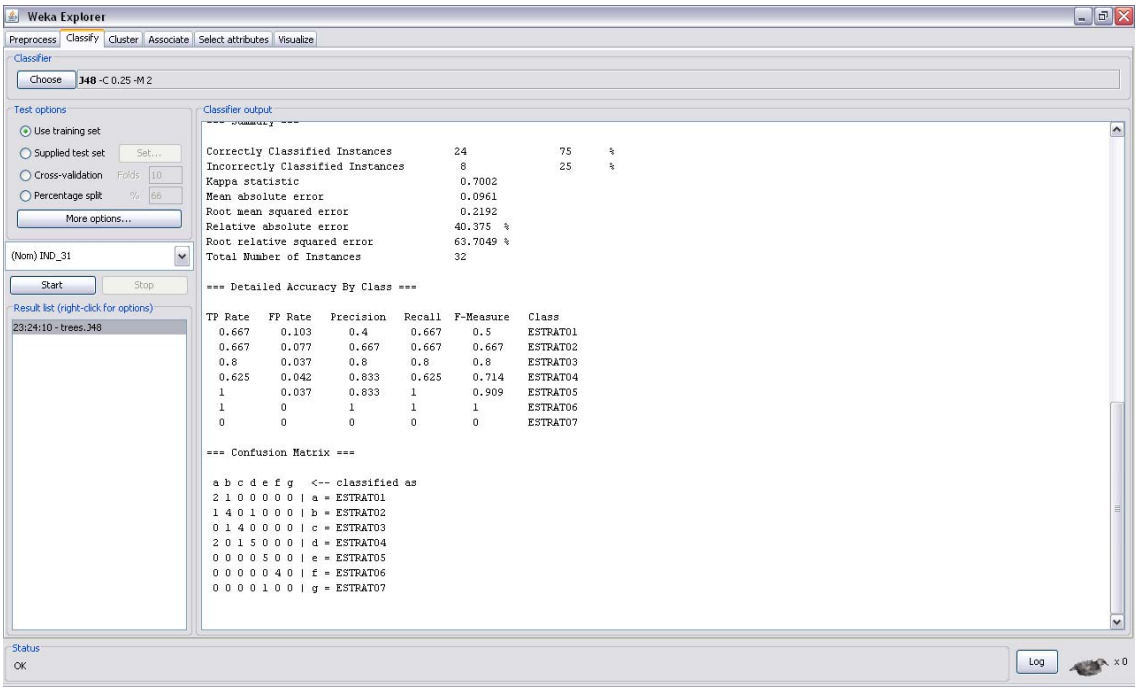


Como se puede observar el árbol no es tan fácil de interpretar, como pudiera ser, lo cual se debe principalmente a la cantidad de variables sobre el mismo indicador, es decir se tienen 30 variables que representan básicamente 5 Indicadores: vivienda, empleo, ocupación, salud y educación, donde casi cada indicador tiene entre 4 y 7 variables. Por lo cual se consideró la eliminación de variables que pudieran tener una fuerte correlación, pero además también se consideró el juicio y conocimiento del problema, es decir se trato de tener la representación desde lo más marginal o que representa las condiciones menos favorables hasta lo que mejor representen las condiciones favorables.

Por ejemplo se hizo un análisis de correlación multivariada para los 30 indicadores y se obtuvo que las variables más representativas para determinar un nivel favorable o desfavorable, fueron en un nivel de significancia decreciente sólo 3 variables:

- % de población en viviendas con teléfono (IND_12)
- % de población ocupada femenina (IND_19)
- % de población ocupada que son profesionistas o técnicos (IND_27)

Se probó el algoritmo con sólo estos 3 indicadores y se obtuvieron los siguientes resultados:



El algoritmo sólo tiene el 75% de efectividad es decir clasifica 25 estados de 32 correctamente y se puede observar en la matriz de confusión que clasifica correctamente a los estados de regiones socioeconómicas consideradas en los estratos 5 y 6 es decir por encima del promedio ó que se consideran bajo situaciones favorables; y no representa mucha diferenciación entre los estratos (1, 2, 3 y 4) considerados con las condiciones menos favorables.

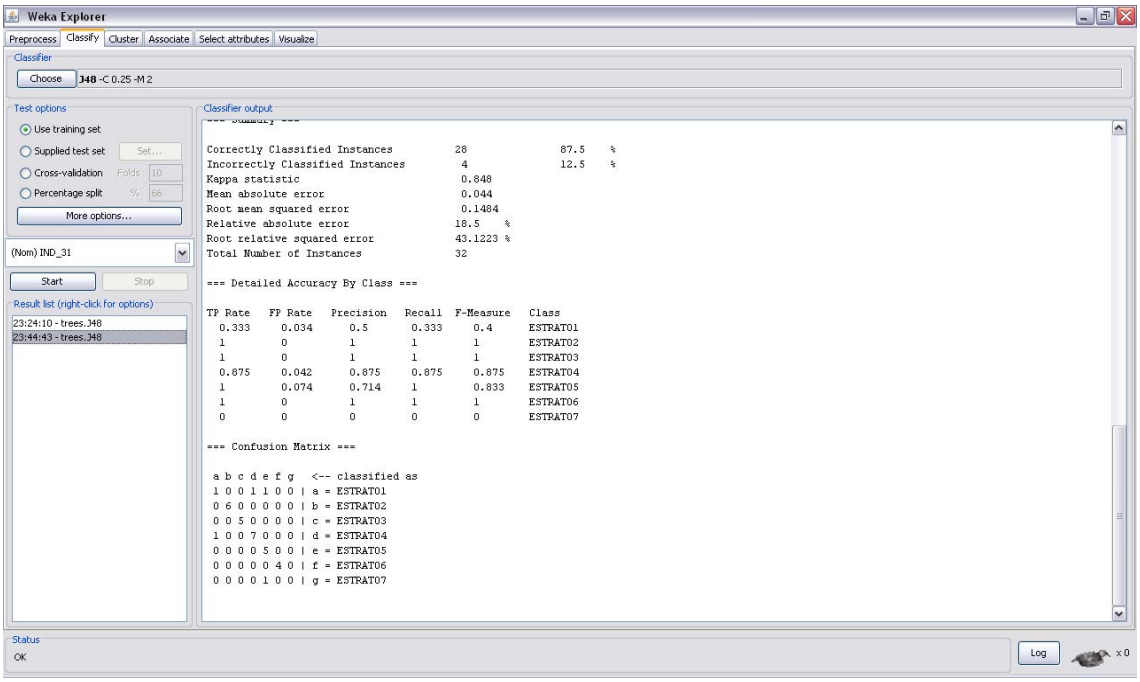
Analizando además de las variables que mejor puedan discriminar un nivel de vida entre favorable y desfavorable, también se trato de considerar al menos a una variable de cada indicador. Los criterios para determinar las variables fueron:

- Variabilidad (Desviación estándar, coeficiente de variación), sobre cada indicador
- Correlación
- Considerar al menos una variable por cada indicador: vivienda, hacinamiento, ocupación, educación, empleo (No fue condición, pero si criterio).

Finalmente se obtuvieron las siguientes variables con las cuales se implementó el algoritmo de clasificación J48.

VARIABLE	INDICADOR
% DE POBLACIÓN EN VIVIENDAS CON AGUA ENTUBADA EN EL ÁMBITO DE LA VIVIENDA	VIVIENDA
% DE POBLACIÓN EN VIVIENDAS CON DRENAJE	VIVIENDA
% DE POBLACIÓN EN VIVIENDAS CON PISO DIFERENTE DE TIERRA	VIVIENDA
% DE POBLACIÓN EN VIVIENDAS CON PAREDES DE MATERIALES DURABLES	VIVIENDA
% DE POBLACIÓN EN VIVIENDAS CON SERVICIO SANITARIO EXCLUSIVO	VIVIENDA
% DE POBLACIÓN EN VIVIENDAS CON REFRIGERADOR	VIVIENDA
% DE POBLACIÓN DE 15 AÑOS Y MÁS ALFABETA	EDUCACIÓN
% DE POBLACIÓN OCUPADA FEMENINA	EMPLEO
% DE POBLACIÓN EN HOGARES QUE PERCIBEN MÁS DE $10.42 DIARIOS POR PERSONA.	EMPLEO
% DE POBLACIÓN OCUPADA QUE SON PROFESIONISTAS O TÉCNICOS	OCUPACIÓN
% DE POBLACIÓN EN VIVIENDAS SIN HACINAMIENTO	HACINAMIENTO

Obteniendo los siguientes resultados:

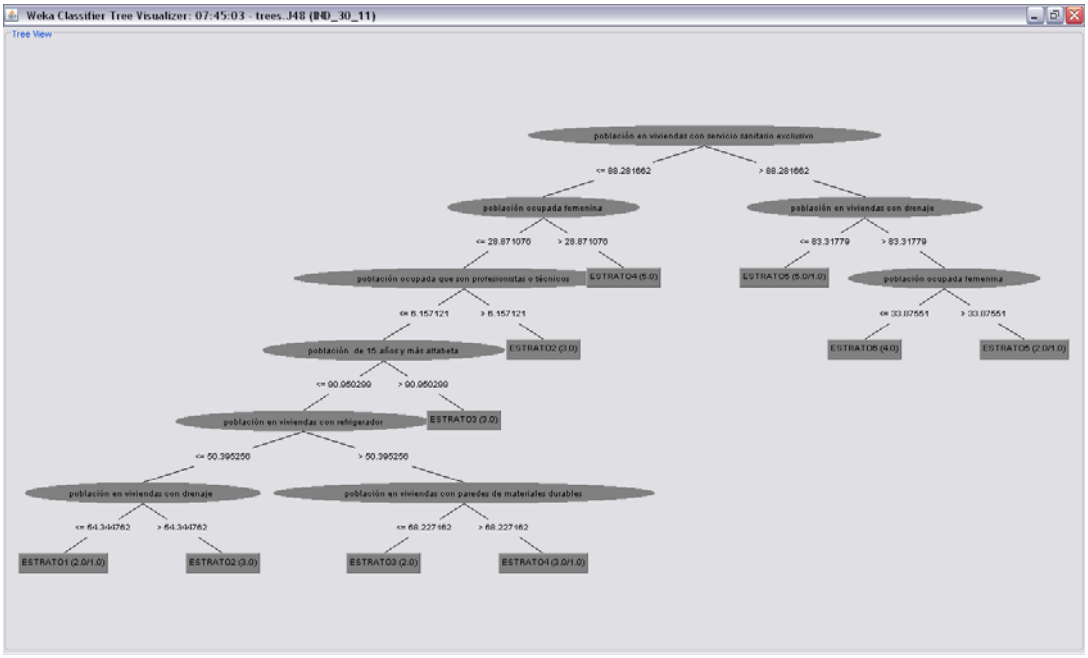


Una efectividad del 87.5%, es decir se clasificó correctamente a 28 estados de 32. Se puede observar mediante la matriz de confusión que de los cuatro estados que clasificó incorrectamente, 2 son del estrato 1 y otro es del estrato 7, esto puede tener las siguientes explicaciones.

No existen diferencias significativas entre esos estados para estar considerados en esos estratos; es decir estos estados incorrectamente asignados a otro estrato, bien pueden ser representados por esos estratos.

El algoritmo clasifica correctamente los estratos en donde existe mayor población, lo cual indica que al no tener parámetros de comparación estos estados (Chiapas, Guerrero y el Distrito Federal) los clasifica incorrectamente.

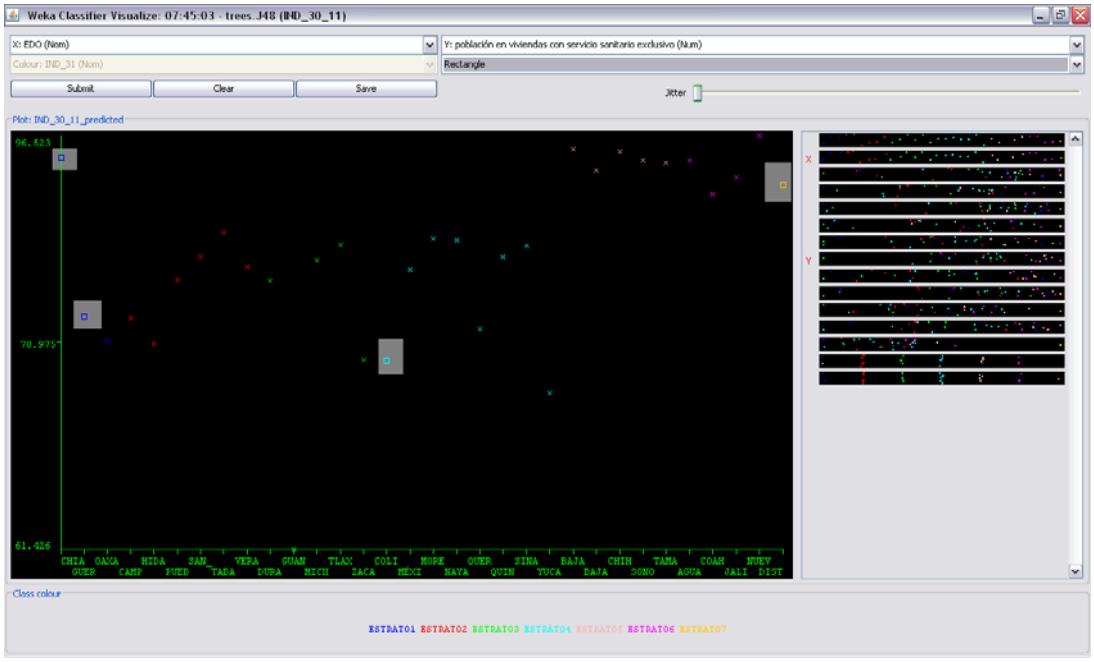
El árbol de clasificación para “Regiones socioeconómicas de México 2000”, generado por el algoritmo J48 con una efectividad del 87.5% (clasifica correctamente 28 de 32 estados).



Considerando los estados representados por estratos con características menos favorables se presenta un análisis de los cuatro estados clasificados incorrectamente.

Respecto a la variable:

"% de población en viviendas con servicio sanitario exclusivo" según los datos del XII censo nacional de vivienda se tiene la siguiente distribución:



Se tiene en el eje X a los estados y en el eje Y se representa la variable "% de población en viviendas con sanitario de uso exclusivo"; se tienen 7 estratos representados por diferentes colores y los estados clasificados correctamente se muestran con una "x" y los estados clasificados incorrectamente se muestran con un cuadro además se les ubicó con una trama en color gris.

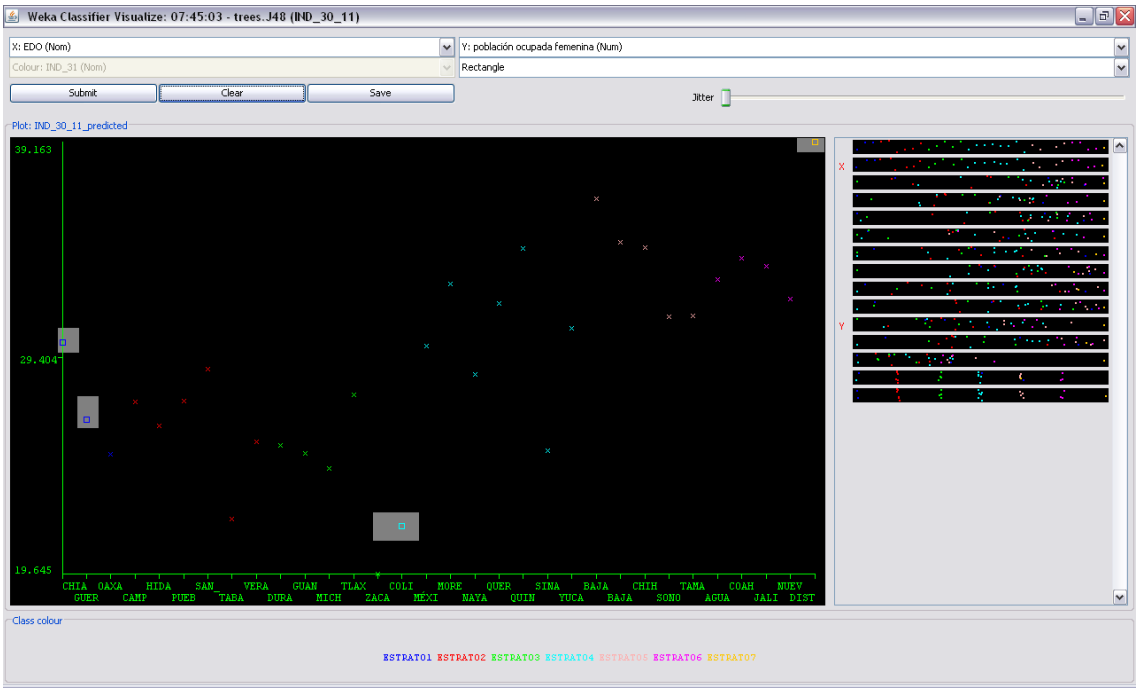
Con esta variable se puede observar la diferencia entre los estados pertenecientes a los estratos 5, 6 y 7 considerados con condiciones más favorables.

- Más del 92% de la población de los estados clasificados en los estratos 5, 6, y 7 habitan viviendas con sanitario exclusivo.
- De los estados clasificados en los estratos 1, 2, 3 y 4 en promedio el 83% de su población habitan viviendas con sanitario exclusivo; con lo cual se puede concluir que existe una diferencia significativa respecto a esa variable relacionada a la vivienda.

Mediante esta variable también se puede observar que estados considerados en el estrato 1 como Chiapas y Guerrero, presentan mayor porcentaje de su población "con servicio de sanitario exclusivo" respecto a otros como Colima considerado con características más favorables.

Continuando con la estructura del árbol respecto a los estados considerados incorrectamente, en la clasificación hacia situaciones menos favorables (Estratos 1, 2, 3 y 4) la variable jerárquica que considero el árbol de clasificación es "% de población ocupada femenina"

Su distribución según el XII censo de población y vivienda 2000:



Se puede observar que es significativa en estados clasificados en condiciones favorables (Estratos 5, 6 y 7), pero que para los estados clasificados en condiciones menos favorables no hay gran significancia, esto también se puede apreciar en las ramas del árbol.

Estados clasificados en el estrato 1 tienen mejores condiciones que estados clasificados en estratos 3 y 4; lo que si se puede observar es que Colima clasificado en el estrato 4 presenta el menor "% de población ocupada femenina" al igual que la variable de peso jerárquico anterior; es decir mediante estas dos variables se puede concluir:

Existen diferencias significativas entre los estados considerados en condiciones más favorables (Estratos 5, 6 y 7) respecto a estados considerados en condiciones menos favorables (Estratos 1, 2, 3 y 4).

- Considerando la división entre estratos con condiciones favorables (Estratos 5, 6 y 7), y la de condiciones menos favorables (Estratos 1, 2, 3 y 4), no se presenta una separación significativa entre ellos, es decir son muy similares, por ejemplo el Distrito Federal (Estrato 7) no tiene una diferencia significativa con los otros estados clasificados en estratos de condiciones favorables (Estratos 5 y 6), e incluso en la variable "% de población en viviendas con sanitario de uso exclusivo", tiene menos población que otros estados.
- De los cuatro estados clasificados incorrectamente que son Chiapas, Guerrero, Colima y el Distrito Federal se puede concluir, considerando estas dos variables de mayor jerarquía según el árbol de clasificación, que son estados que tienen características diferentes a las que describen en promedio al estrato asignado por "Regiones socioeconómicas de México", por lo cual es comprensible la incorrecta clasificación del algoritmo.

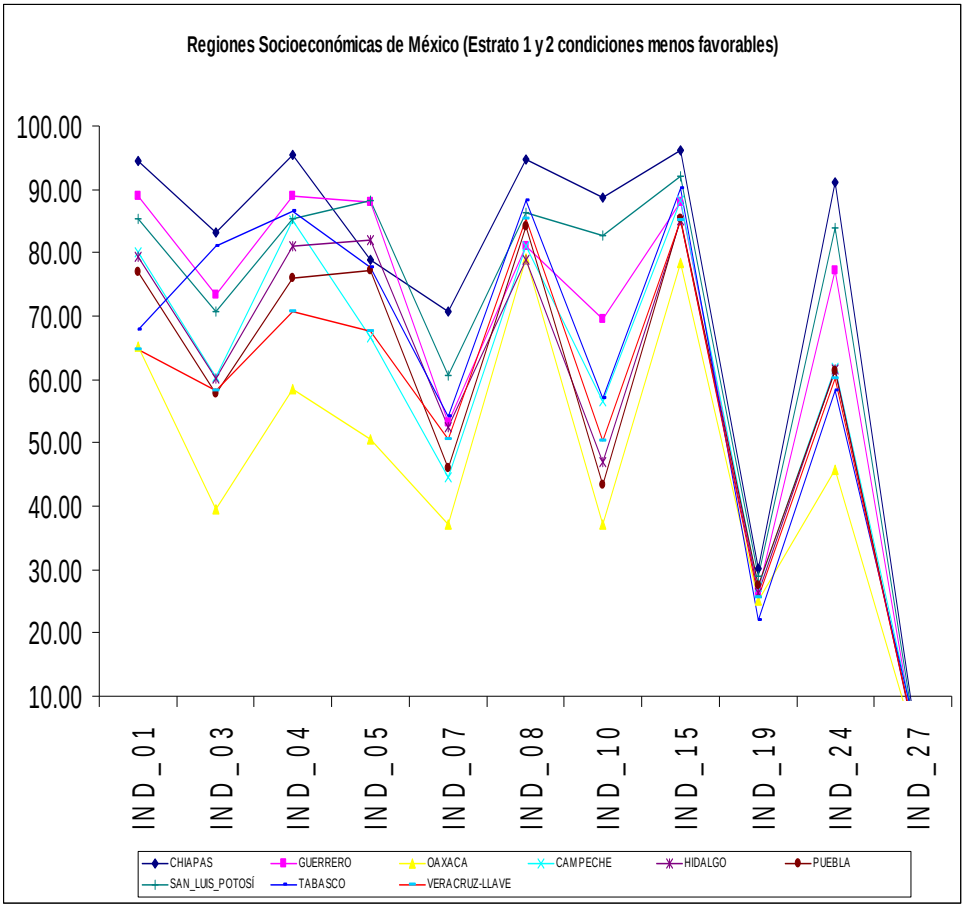
Análisis descriptivo de las clasificaciones incorrectas:

A continuación se presenta una tabla que presenta el "ranking", de cada estado para las diferentes variables.

Chiapas y Guerrero (Estrato 1, características menos favorables a nivel nacional)

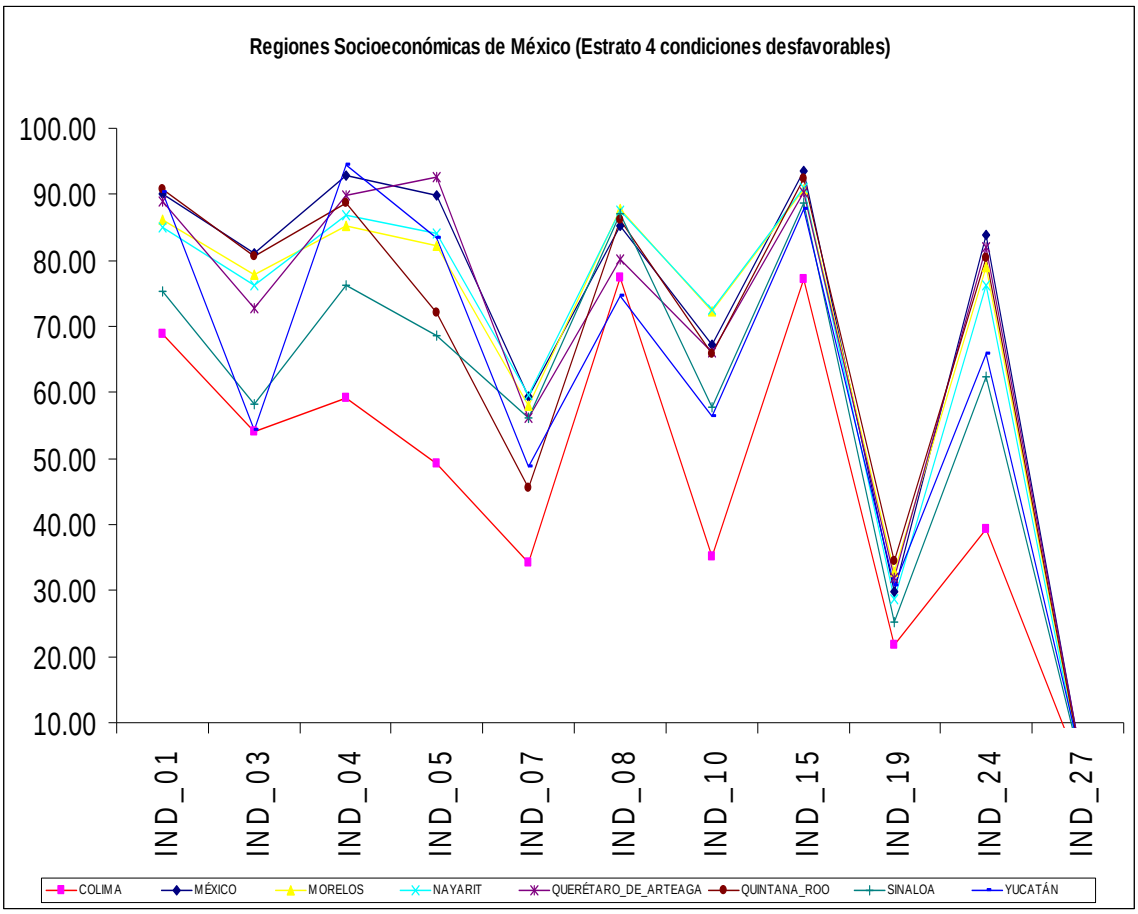
La siguiente gráfica representa los estratos 1 y 2 considerados como los que tienen las condiciones menos favorables a nivel nacional, entre los cuales se encuentran dos de los estados que el algoritmo clasificó mal, Chiapas y Guerrero respecto a los 11 indicadores con los que se construyó el modelo de clasificación.

Con base a cada serie que representa un estado se observa que Oaxaca es el estado con menor % en todos los indicadores, es decir es significativamente diferente a todos los demás, mientras que Chiapas y Guerrero que de acuerdo a "Regiones socioeconómicas de México" tienen características similares a Oaxaca, tienen ambos porcentajes incluso por encima de los mostrados por estados que se clasifican en mejores condiciones.



Colima (Estrato 4, condiciones desfavorables)

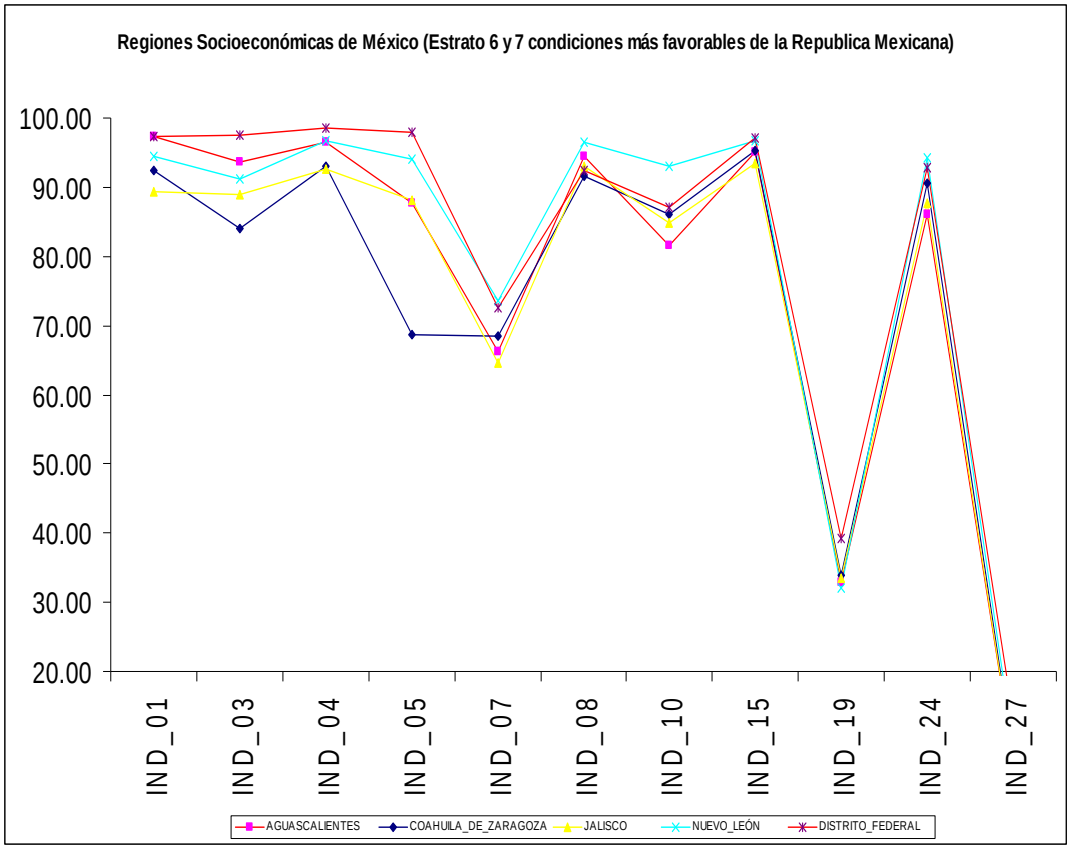
La siguiente gráfica representa los estados clasificados en el estrato 4 considerados con características desfavorables a nivel nacional, entre los cuales se encuentran el estado de Colima, estado que el algoritmo clasificó mal y el cual es el más contrastante respecto a los 11 indicadores con los que se construyó el modelo de clasificación. La clasificación le asignó el estrato 1 siendo el de mayor diferencia respecto a "Regiones socioeconómicas de México 2000", que le asignó el estrato 4.



Se observa claramente que Colima presenta los porcentajes más bajos respecto a todo el estrato, lo cual explica el por que la clasificación respecto a los once indicadores, lo clasifica en el estrato 1 considerado el de características menos favorables a nivel nacional.

Distrito Federal (Estrato 7, condiciones más favorables a nivel nacional)

La siguiente gráfica representa los estados clasificados en el estrato 6 y 7 considerados con las características más favorables a nivel nacional; el Distrito Federal es el único estado que sólo conforma un estrato; es decir según la metodología de "Regiones socioeconómicas de México 2000", es el estado que respecto a vivienda, educación, hacinamiento, salud y ocupación, presenta diferencias significativas de su población respecto a los otros 31 estados de la República Mexicana.



La gráfica muestra como los 5 estados considerados, en este caso en los estratos 5 y 6 que se consideran como los que presentan las condiciones más favorables a nivel nacional; no muestran gran diferencia y fluctúan entre el promedio de cada indicador. Aunque es importante mencionar que la mayor separación de estos estados es referente a los indicadores referentes a la "Infraestructura de la vivienda, y su calidad" (Indicadores 1, 3, 4 y 5), y es el Distrito Federal el que tiene a la mayor parte de su población con estas características respecto a los demás estados.

Evaluación de los resultados

Respecto a la clasificación del árbol que se desarrolló en base al producto "Regiones socioeconómicas de México 2000" basado en el XII censo general de población y vivienda realizado por el INEGI se puede concluir:

El producto considera 5 indicadores principales: educación, empleo, hacinamiento salud y vivienda.

Existen diferencias significativas en los indicadores de vivienda, entre los estados con las condiciones más favorables (Estratos 5, 6 Y 7), y los estados con condiciones menos favorables (Estratos 1, 2, 3 y 4), estos indicadores de vivienda se dividen en 3:

- Infraestructura de la vivienda
- Calidad de la vivienda
- Equipamiento de la vivienda

Se implementó un modelo de clasificación mediante el algoritmo J48 considerando 11 de las 30 variables usadas en la metodología de "Regiones socioeconómicas de México 2000" obteniendo el 88% de efectividad (Clasifica correctamente 28 de 32 estados), el cual puede servir para evaluar nuevos periodos de observación, quizás la actualización de "Regiones socioeconómicas 2010", correspondiente al XIII censo nacional de población y vivienda ó un conteo que considere las mismas variables o el sentido de las mismas.

- Oaxaca es el estado que presenta las condiciones menos favorables a nivel nacional y en 2005 tenía casi el 4% del total de la población de la República Mexicana.
- Colima es de los estados con menos población (0.6%), del total de la República Mexicana en 2005, aunque presenta condiciones en su población muy similar o incluso por debajo de las de los estados considerados menos favorables respecto a educación, empleo, hacinamiento, salud y vivienda.

- El Distrito Federal que en 2005 representaba casi el 9% del total de la población de la República Mexicana, es significativamente diferente a los demás 31 estados en cuanto a infraestructura de la vivienda (condiciones urbanas), pero no muestra diferencias significativas en indicadores de empleo, hacinamiento, ocupación, y salud.

Respecto al árbol de clasificación considerado para los datos correspondientes al "Índice de marginación México 2005" que tiene como clasificación índices considerados como: "Muy alto", "Alto", "Medio", "Bajo" y "Muy bajo" se concluye lo siguiente:

Existen dos indicadores altamente discriminantes en esta clasificación que son:

- Población (% de población en localidades con menos de 5, 000 habitantes)
- Educación (% población analfabeta de 15 años ó más).

La clasificación discrimina muy bien entre:

- Índices de marginación "Medio", "Alto" y "Muy alto"
- Índices de marginación "Bajo" y "Muy bajo"

Se implementó un modelo de clasificación mediante el algoritmo J48 con un 97% de efectividad (Clasificó correctamente 31 de 32 estados), con 10 indicadores a partir de los datos del II conteo de población y vivienda con el que se desarrolló la estimación por parte de CONAPO de "Índices de marginación México 2005", esta clasificación puede servir para evaluar nuevos periodos de observación, quizás la actualización de "Índices de marginación México 2009", el III conteo de ocupación y empleo, condiciones de marginación para regiones de un estado o un municipio.

- El estado que la clasificación asignó incorrectamente fue Guanajuato el cual presenta altos índices de analfabetismo y altos porcentajes de personas mayores de 15 años sin terminar la primaria lo cual es característica de estados con índices de marginación altos.

Ya se mencionó que "Regiones socioeconómicas de México 2000" y la estimación de "Índices de marginación México 2005", no pueden ser comparables estadísticamente, pero pueden representar descriptivamente cambios o tendencias sobre la situación poblacional, educativa, de vivienda y empleo de la República Mexicana.

La siguiente tabla muestra las clasificaciones de "Regiones socioeconómicas México 2000", realizada por INEGI y la estimación de "Índices de marginación México 2005".

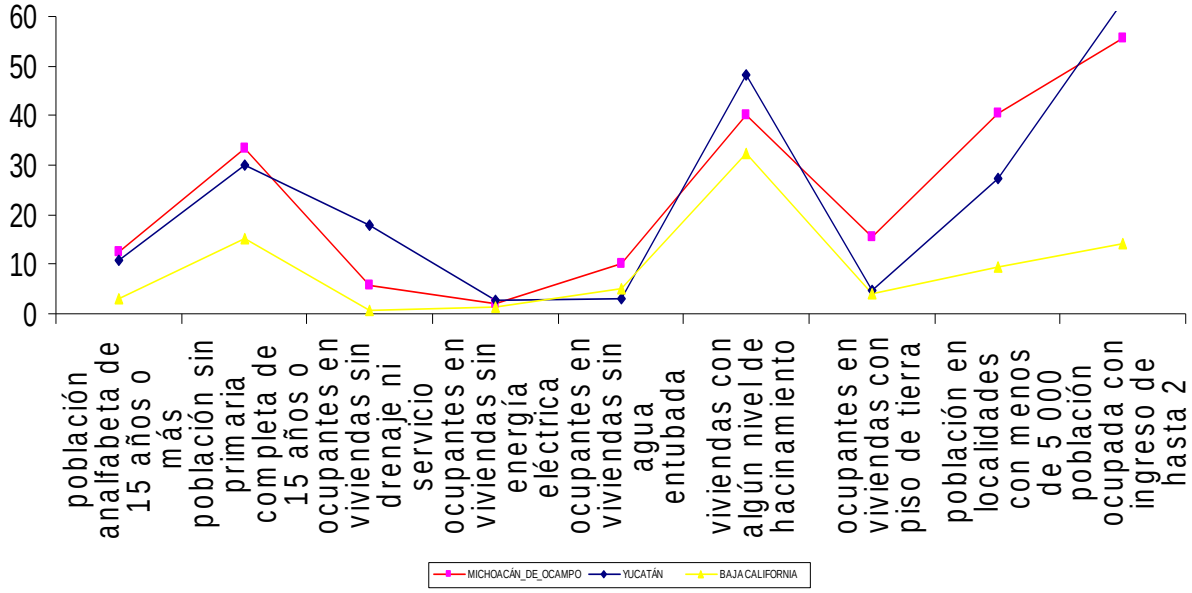
ESTADOS DE LA REPÚBLICA MEXICANA	REGIONES SOCIOECONÓMICAS DE MÉXICO 2000	ÍNDICES DE MARGINACIÓN MÉXICO 2005
CHIAPAS	ESTRATO1	MUY ALTO
GUERRERO	ESTRATO1	MUY ALTO
OAXACA	ESTRATO1	MUY ALTO
CAMPECHE	ESTRATO2	ALTO
HIDALGO	ESTRATO2	ALTO
PUEBLA	ESTRATO2	ALTO
SAN_LUIS_POTOSÍ	ESTRATO2	ALTO
TABASCO	ESTRATO2	ALTO
VERACRUZ-LLAVE	ESTRATO2	ALTO
DURANGO	ESTRATO3	MEDIO
GUANAJUATO	ESTRATO3	MEDIO
MICHOACÁN_DE_OCAMPO	ESTRATO3	ALTO
TLAXCALA	ESTRATO3	MEDIO
ZACATECAS	ESTRATO3	MEDIO
COLIMA	ESTRATO4	BAJO
MÉXICO	ESTRATO4	BAJO
MORELOS	ESTRATO4	BAJO
NAYARIT	ESTRATO4	MEDIO
QUERÉTARO_DE_ARTEAGA	ESTRATO4	MEDIO
QUINTANA_ROO	ESTRATO4	BAJO
SINALOA	ESTRATO4	MEDIO
YUCATÁN	ESTRATO4	ALTO
BAJA CALIFORNIA	ESTRATO5	MUY BAJO
BAJA CALIFORNIA_SUR	ESTRATO5	BAJO
CHIHUAHUA	ESTRATO5	BAJO
SONORA	ESTRATO5	BAJO
TAMAULIPAS	ESTRATO5	BAJO
AGUASCALIENTES	ESTRATO6	BAJO
COAHUILA_DE_ZARAGOZA	ESTRATO6	MUY BAJO
JALISCO	ESTRATO6	BAJO
NUEVO_LEÓN	ESTRATO6	MUY BAJO
DISTRITO_FEDERAL	ESTRATO7	MUY BAJO

Se puede observar que los estados con cambios contrastantes entre una clasificación y otra y considerando los diferentes años son:

- Michoacán de Ocampo
- Yucatán
- Baja California

Lo cual mediante la clasificación obtenida con los datos de CONAPO se puede explicar y visualizar fácilmente con la siguiente gráfica; la cual considera todos los indicadores con que se estimo el índice de marginación.

Indicadores para considerar un Nivel de Marginación CONAPO 2005



- Como muestra la gráfica para el estado de Baja California el 3% de su población es "analfabeta de 15 años o más", mientras que se puede observar el contraste de Michoacán de Ocampo con 13% y Yucatán el 11% de su población son "analfabetas de 15 años o más"

Otro indicador importante es % de población en localidades con menos de 5,000 habitantes teniendo Baja California el 9% de su población con estas características, mientras que Yucatán tiene el 21% y Michoacán de Ocampo el 41% de su población vive bajo condiciones de aislamiento.

Hasta ahora se han analizado las diferentes comparaciones entre las diferentes clasificaciones tanto “Regiones socioeconómicas México 2000”, como “Índices de marginación México 2005” se han presentado los estados con mayor contraste para cada modelo y su respectiva clasificación.

Proyecciones, demográficas de la República Mexicana del periodo 1990 al 2030

A continuación se presenta un análisis sobre los indicadores de la proyección realizada por CONAPO que considera el periodo 1990 - 2030.

La tabla muestra los datos obtenidos por el INEGI del XII censo nacional de población y vivienda del año 2000, los datos del II conteo de población y vivienda del año 2005; es decir los datos correspondientes al año 2000 y 2005 son reales y oficiales, mientras que los datos correspondientes al año 2009, 2010, 2029 y 2030 son proyecciones realizadas por CONAPO.

ESTADOS DE LA REPÚBLICA MEXICANA	POB INEGI 2000	RANK 'INEGI_2000	POB MARG 2005	RANK_MARG 2005	POB PROY 2009	RANK 'PROY 2009	POB PROY 2010	RANK 'PROY 2010	POB PROY 2029	RANK 'PROY 2029	POB PROY 2030	RANK 'PROY 2030
MÉXICO	13,096,686	32	14,007,495	32	14,837,208	32	15,031,728	32	17,977,249	32	18,088,060	32
DISTRITO_FEDERAL	8,605,239	31	8,720,916	31	8,841,916	31	8,846,752	31	8,607,989	31	8,575,089	31
VERACRUZ-LLAVE	6,908,975	30	7,110,214	30	7,278,690	30	7,294,895	30	7,373,828	29	7,362,776	29
JALISCO	6,322,002	29	6,752,113	29	7,016,595	29	7,070,555	29	7,771,091	30	7,787,954	30
PUEBLA	5,076,686	28	5,383,133	28	5,651,371	28	5,705,519	28	6,500,237	28	6,527,495	28
GUANAJUATO	4,663,032	27	4,893,812	27	5,044,735	27	5,067,217	27	5,272,644	26	5,270,383	25
CHIAPAS	3,920,892	25	4,293,459	26	4,507,177	26	4,553,358	26	5,264,508	25	5,290,229	26
NUEVO_LEÓN	3,834,141	24	4,199,292	25	4,448,068	25	4,502,035	25	5,364,079	27	5,398,387	27
MICHOACÁN_DE_OCAMPO	3,985,667	26	3,966,073	24	3,964,009	24	3,949,377	24	3,560,338	21	3,533,061	21
OAXACA	3,438,765	23	3,506,821	23	3,550,788	23	3,548,623	23	3,411,814	20	3,397,575	20
CHIHUAHUA	3,052,907	21	3,241,444	22	3,391,617	22	3,422,047	22	3,826,672	23	3,838,176	23
TAMAULIPAS	2,753,222	20	3,024,238	20	3,193,017	21	3,230,307	20	3,802,915	22	3,824,091	22
BAJA CALIFORNIA	2,487,367	18	2,844,469	19	3,165,776	20	3,252,690	21	4,983,934	24	5,074,986	24
GUERRERO	3,079,649	22	3,115,202	21	3,140,529	19	3,134,433	19	2,902,323	18	2,883,660	18
SINALOA	2,536,844	19	2,608,442	18	2,652,451	18	2,655,951	18	2,617,000	16	2,608,651	16
COAHUILA_DE_ZARAGOZA	2,298,070	16	2,495,200	17	2,628,942	17	2,655,187	17	3,041,360	19	3,054,774	19
SONORA	2,216,969	14	2,394,861	15	2,510,562	16	2,532,639	16	2,832,331	17	2,841,311	17
SAN_LUIS_POTOSÍ	2,299,360	17	2,410,414	16	2,484,949	15	2,495,513	15	2,595,863	15	2,595,169	15
HIDALGO	2,235,591	15	2,345,514	14	2,421,606	14	2,433,563	14	2,568,702	14	2,569,852	14
TABASCO	1,891,829	13	1,989,969	13	2,050,514	13	2,060,628	13	2,164,835	10	2,164,863	10
YUCATÁN	1,658,210	12	1,818,948	12	1,921,959	12	1,945,840	12	2,369,479	12	2,388,286	12
QUERÉTARO_DE_ARTEAGA	1,404,306	9	1,598,139	10	1,720,556	11	1,750,965	11	2,280,322	11	2,303,496	11
MORELOS	1,555,296	11	1,612,899	11	1,674,795	10	1,687,396	10	1,852,247	9	1,856,004	9
DURANGO	1,448,661	10	1,509,117	9	1,550,417	9	1,555,688	9	1,583,244	8	1,580,639	8
ZACATECAS	1,353,610	8	1,367,692	8	1,379,752	8	1,377,708	8	1,286,010	5	1,278,576	5
QUINTANA_ROO	874,963	4	1,135,309	7	1,314,062	7	1,361,821	7	2,392,141	13	2,450,833	13
AGUASCALIENTES	944,285	6	1,065,416	5	1,141,946	6	1,159,304	6	1,446,108	7	1,458,116	7
TLAXCALA	962,646	7	1,068,207	6	1,134,844	5	1,149,653	5	1,396,287	6	1,406,950	6
NAYARIT	920,185	5	949,684	4	969,540	4	971,913	4	987,540	4	986,329	4
CAMPECHE	690,689	3	754,730	3	795,982	3	805,182	3	961,040	3	967,262	3
COLIMA	542,627	2	567,996	2	600,924	2	608,535	2	728,695	1	733,205	1
BAJA CALIFORNIA_SUR	424,041	1	512,170	1	565,400	1	579,189	1	820,959	2	831,837	2
ESTADOS CON DISMINUCIÓN DE SU POBLACIÓN												
ESTADOS CON INCREMENTO DE SU POBLACIÓN												

Los datos representan la población del año correspondiente y su ranking respecto a su situación poblacional; es decir a mayor ranking mayor población. La tabla está ordenada respecto al estado con mayor población respecto a la proyección del año 2009.

Información y contrastes relevantes de la situación poblacional de los estados de la República Mexicana:

- Se puede observar que los estados más poblados son: México, Distrito Federal, Veracruz, Jalisco y Puebla, además Jalisco se proyecta como el tercer estado más poblado de la República Mexicana desde el año 2029, con más población que Veracruz que según el INEGI es el tercer estado más poblado actualmente. Según la proyección estos cinco estados representarán el 40% del total de la población de México en 2030.
- Se puede observar que todos los estados con aumento de población según la proyección son estados del norte de la República Mexicana a excepción de Quintana Roo, que al igual que Baja California son los dos estados con mayor aumento de la población a largo plazo según las proyecciones.
- De los estados con mayor contraste disminuyendo o aumentando su población considerando las proyecciones 2009 vs 2030 son:

    - Michoacán de Ocampo disminuyendo 11% su población.
    - Quintana Roo aumentando 87% su población
    - Baja California aumentando 60% su población.

- De los estados con menor cantidad de población del total de la República Mexicana:

    - Quintana Roo es el que presenta el contraste más relevante ya que de ser de los 5 estados menos poblados según la proyección será uno de los 20 estados más poblados en 2030.
    - Colima es el estado menos poblado y seguirá siendo según las proyecciones hasta 2030

Conclusiones

Actualmente se puede establecer que los datos relacionados a una organización deben ser considerados como uno de sus principales activos. La planeación, integración e implementación de sistemas de almacenamiento de datos, son parte fundamental en cualquier organización, con lo que se debe garantizar la validez de los datos.

El concepto data warehouse resulta ser una herramienta fundamental ya que representa la relación integrada de toda la información de una organización; esta información puede ser interna y externa. Contando con una base de datos, el proceso de descubrimiento de conocimiento en bases de datos (KDD), representa una herramienta de gran potencial al obtener información previamente desconocida que tenga una aportación y que sea útil para la optimización de procesos, disminución de costos, hacer más efectivos programas y estrategias entre muchas ventajas más.

La principal aportación de este trabajo motiva el uso de software libre. En este caso se utilizó WEKA como aplicación de técnicas de minería de datos y postgreSQL como sistema manejador de base de datos, pero no son los únicos, existen otros como “Tanagra”, “Sipina”, “R” “Orange”, entre muchos otros. El uso de este tipo de software es un área de oportunidad para aquellas organizaciones carentes de recursos para manejar aplicaciones comerciales.

Una de las motivaciones para realizar este trabajo fue demostrar el potencial que existe al utilizar software de uso libre. Aunque México hoy esta por debajo de países como Argentina, Chile e incluso Perú en su uso; el software libre debe ser apoyado principalmente por organizaciones públicas como universidades por ejemplo. El software libre tiene ventajas y desventajas dado que muchos procesos que se realizan de manera inmediata con software comercial hay que desarrollarlos en el caso de software libre, es decir hay que programarlas, lo cual involucra tiempo y razonamiento.

Este trabajo puede servir como base de estudio para una nueva área de desarrollo que son "Sistemas de información geográficos" (GIS) *Geographical Information Systems*; la información obtenida del análisis sobre indicadores de la República Mexicana muestra como existen diferencias y similitudes en cuanto a zonas geográficas lo cual hoy es un área de desarrollo. Los sistemas de información geográficos combinados con modelos econométricos y técnicas de minería de datos presentan un nuevo enfoque llamado análisis geoespacial (*Geospatial Analysis*), área en donde también existe software libre llamado "GEoDA" (http://geodacenter.asu.edu), que permite realizar análisis a nivel espacial (regional, municipal o estatal).

Finalmente al análisis desarrollado a partir del producto "Regiones socioeconómicas de México, 2000" (INEGI, 2000), y la estimación de "Índices de marginación México 2005" (CONAPO, 2005), presenta resultados que pueden ser de gran utilidad para organizaciones públicas teniendo como premisa las principales características de algunas regiones de la República Mexicana, la transformación de la población, la esperanza de vida que se ha mantenido en crecimiento, la integración y desaparición de estratos socioeconómicos, indicadores de analfabetismo, desempleo y condiciones referentes a la vivienda de los mexicanos pueden ser utilizados para generar nuevos programas sociales o apoyar los existentes.

Bibliografía

Arnold R. Robert; Hill, Harold; Nichols V. Aylmer (1975): “Sistema moderno de procesamiento de datos”; Ed. Limusa.

Burch Jhon G Jr; Strater Felix R. Jr. (1986): “Sistemas de información teoría y práctica”; Ed. Limusa.

Cood Frank E. (1970): “A relational model of data for large shared of data banks”; Addison Wesley Publishing Company

Eibe Frank; Witten Ian H; (2005): “Data mining: practical machine learning tools and techniques”; Ed. Elsevier. Second edition.

Fayyad, Usama; Piatetsky-Shapiro Gregory and Padhraic Smyth; (1996): ”From data mining to knowledge discovery in databases”; American Association for Artificial Intelligence.

Inmon, William H. (1996): “Building the data Warehouse”; Wiley Computer, New York.

Lockhart, Thomas (1996a): “Tutorial de PostgreSQL”; Desarrollado por Postgres global development group

Lockhart, Thomas (1996b): “Guía del Programador”; Desarrollado por Postgres global development group

Martin, James (1975): “Organización de las bases de datos”; Ed. Prentice Hall.

Matías Hernández Sergio Alejandro (2009). “Manual de Curso-Taller Manipulación de Datos con el paquete estadístico R”; Universidad Nacional Autónoma de México, Dirección general de asuntos de personal académico, Facultad de Estudios Superiores Acatlán

Matías Hernández Sergio Alejandro (2008). ”Manual de Curso-Taller Bases de Datos PostgreSQL Básico”; Universidad Nacional Autónoma de México, Dirección general de asuntos de personal académico, Facultad de Estudios Superiores Acatlán

Momjian, Bruce (2000): “PostgreSQL: introduction and concepts”. Editorial Addison.Wesley.

Piatetsky-Shapiro Gregory (2000): “Knowledge discovery in databases: 10 years after”; Knowledge Stream Partners

Sitios Web consultados

AECEM, (2007): “Asociación española de comercio electrónico y marketing relacional” http://www.aecem.org/

Eckerson, Wayne; (2005): “The Five Dimensions of Business Intelligence” http://www.tdwi.org/Publications

Warrilow, Emma; (2007): http://www.emmawarrillow.com/

Apéndice

1. Historia PostgreSQL

Uno de los primeros intentos por implementar un motor de base de datos relacional se dio en la Universidad de Berkeley en California Estados Unidos y el proyecto se llamo Ingres a finales de los años 70´s.

Retomando el proyecto Ingres a inicios de los 80´s en la universidad de Berkeley se trabajo con bases de datos bajo el modelo relacional representando los datos mediante tablas y relaciones entre estas; llamando a este proyecto "*Post-ingres*" o "*Postgres*". En 1988 ya se contaba con una versión utilizable y es en 1989 cuando se publica la primera versión para una pequeña comunidad de usuarios. En años continuos se publicaron nuevas versiones básicamente apegados al modelo relacional pero con mayor apertura a la comunidad de usuarios que demandaba más características, es entonces que en 1994 antes de la publicación de versión 4 el proyecto terminó y se disolvió el grupo de desarrollo.

El proyecto de Ingres y Postgres se desarrollaron bajo la licencia BSD (*Berkeley System Distribution*) la cual en esencia se considera de uso libre; es decir su replica, uso, modificación y distribución no implican ningún delito o irregularidad.

La licencia BSD es una licencia que permite el uso y modificación del código publicado, hecho que permitió que en 1995 estudiantes de la universidad de Berkeley trabajaran con el proyecto principalmente incorporando un lenguaje de consultas de SQL, creando así el sistema Postgres95. Finalmente en 1996 se decidió reflejar en el motor la representación del estándar SQL y su nombre cambia a PostgreSQL.

Desarrolladores entusiastas y apasionados de tecnología de base de datos se unieron al proyecto y entre todos comenzaron a incorporar muchas características al motor, explotando su portabilidad, desarrollo y su esencia y espíritu de uso libre convirtiéndolo hoy en día en el motor de base de datos de uso libre a nivel mundial con mayor potencial y herramientas.

1.1 Definición PostgreSQL

PostgreSQL es un sistema de gestión de bases de datos objeto-relacionales (*ORDBMS)* que ha presentado varias versiones y nuevas características desde 1977, cuando fue su primer lanzamiento en la universidad de Berkeley en California Estados Unidos.

En 1996, debido a un nuevo esfuerzo sobre el uso de código abierto y a la incrementada funcionalidad del software, Postgres fue renombrado a PostgreSQL. El proyecto PostgreSQL sigue actualmente un activo proceso de desarrollo a nivel mundial gracias a un equipo de desarrolladores de código abierto.

2. Características de PostgreSQL

Está considerado como la base de datos de código abierto más avanzada del mundo. Proporcionando un gran número de características que normalmente sólo se encontraban en las bases de datos comerciales como ORACLE, o Sybase por mencionar solo algunas. A continuación se mencionan las características de mayor representación y funcionalidad respecto a otros manejadores de bases de datos.

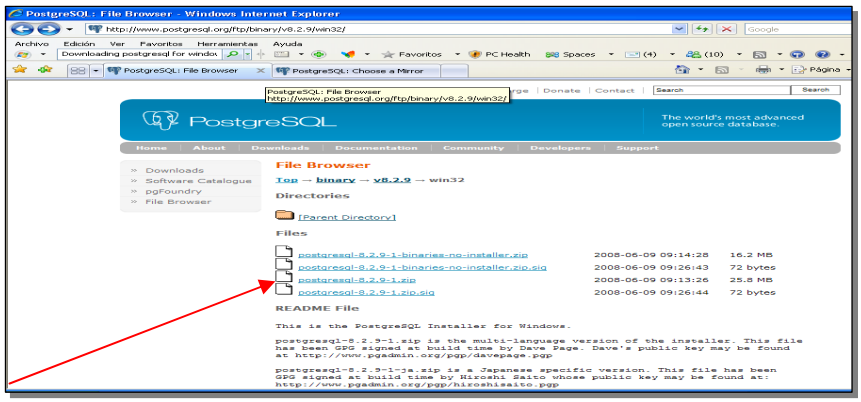
- Sistema administrador de base de datos (DBMS) objeto-relacional: PostgreSQL aproxima los datos a un modelo objeto-relacional, y es capaz de manejar complejas rutinas y reglas. Ejemplos de su avanzada funcionalidad son consultas SQL declarativas, control de concurrencia multi-versión, soporte multi-usuario, transacciones, optimización de consultas, herencia, y *arrays* (arreglos).

- Extensible: Soporta operadores, funciones, métodos de acceso y tipos de datos definidos por el usuario.
- Soporte SQL: Soporta la especificación SQL99 e incluye características avanzadas tales como las uniones (*joins*), SQL92.
- Integridad referencial: Manejo y soporte de integridad referencial, la cual es utilizada para garantizar la validez de los datos de la base de datos.
- Interfaz flexible: La flexibilidad del API de PostgreSQL ha permitido proporcionar soporte al desarrollo del *RDBMS*. Estas interfaces incluyen *Object Pascal, Python, Perl, PHP, ODBC, Java/JDBC, Ruby, TCL, C/C++, y Pike.*
- Lenguajes procedurales: Soporte para procedimientos internos, incluyendo un lenguaje nativo denominado *PL/pgSQL*.
- MVCC: Control de concurrencia multi-versión (*Multi-Version Concurrency Control*), es la tecnología que PostgreSQL usa para evitar bloqueos innecesarios. MVCC está considerado mejor que el bloqueo a nivel de fila porque un lector nunca es bloqueado por un escritor, en su lugar, PostgreSQL mantiene una ruta a todas las transacciones realizadas por los usuarios de la base de datos. PostgreSQL es capaz entonces de manejar los registros sin necesidad de que los usuarios tengan que esperar a que los registros estén disponibles.
- Cliente-servidor: PostgreSQL usa una arquitectura cliente-servidor. Esta es similar al método del Apache 1.3.x para manejar procesos. Hay un proceso maestro que se ramifica para proporcionar conexiones adicionales para cada cliente que intente conectar a PostgreSQL.
- *Write Ahead Logging* (WAL): La característica de PostgreSQL conocida como *Write Ahead Logging* incrementa la dependencia de la base de datos al registro de cambios antes de que estos sean escritos en la base de datos. Esto garantiza que en el hipotético caso de que la base de datos se caiga, existirá un registro de las transacciones a partir del cual podremos restaurar la base de datos, lo cual puede ser enormemente beneficioso en el caso de caída, ya que cualesquiera cambios que no fueron escritos en la base de datos pueden ser recuperados usando el dato que fue previamente registrado. Una vez que el sistema ha quedado restaurado

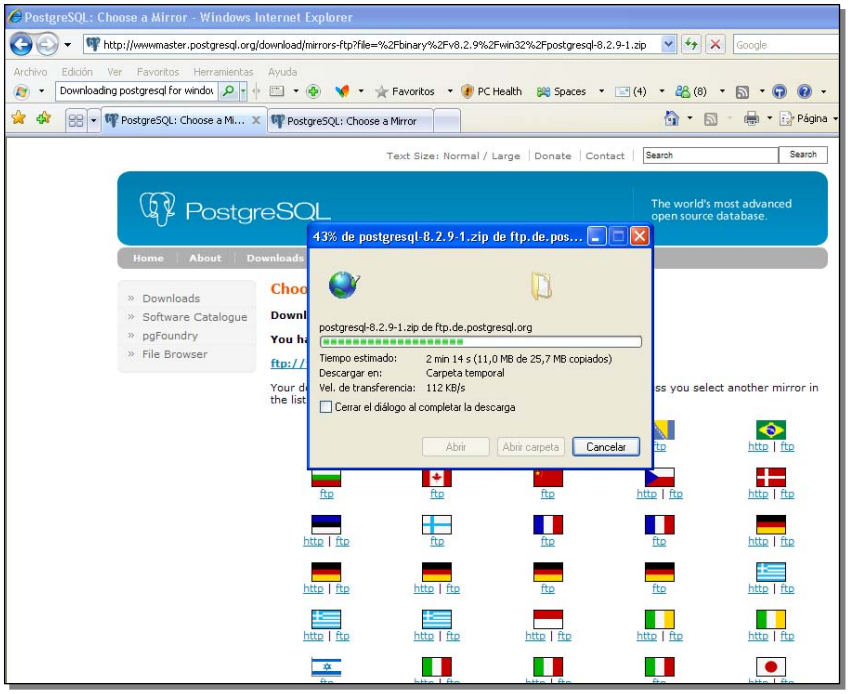
Cualquier usuario puede continuar trabajando desde el punto en que lo dejó cuando la base de datos dejo de estar disponible.

3. Instalación de PostgreSQL para Windows XP

Entrar a la siguiente dirección electrónica: http://www.postgresql.org/ftp/binary/v8.2.9/win32/



Se le da un click al archivo postgresql-8.2.9-1.zip y aparece la siguiente pantalla donde se debe seleccionar el país donde se encuentra el servidor de donde se bajara el archivo que permitirá instalar PostgresSQL, tal y como se muestra a continuación.

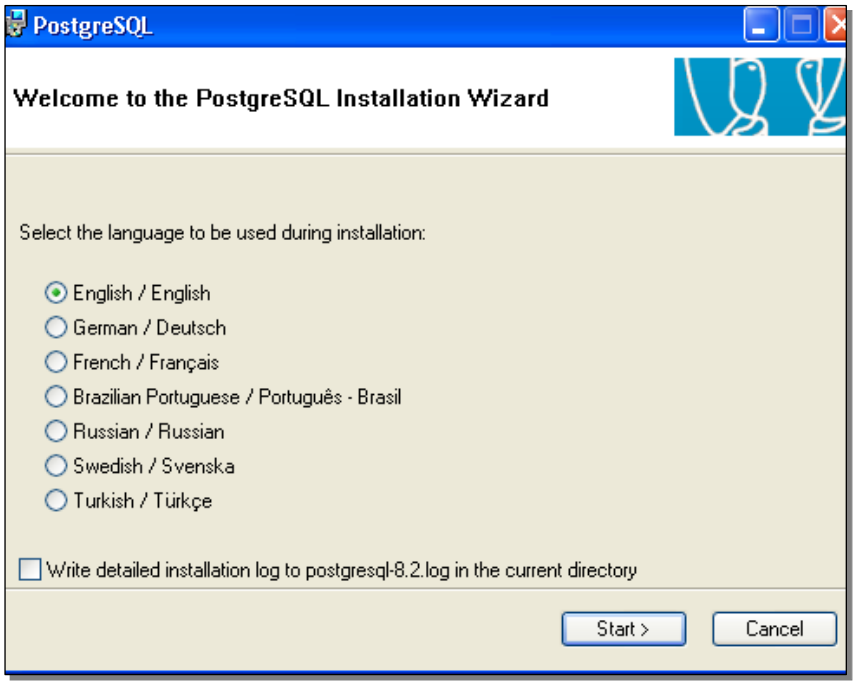


Una vez que se termina de descargar el archivo, hay que descomprimirlo y muestra los siguientes archivos:

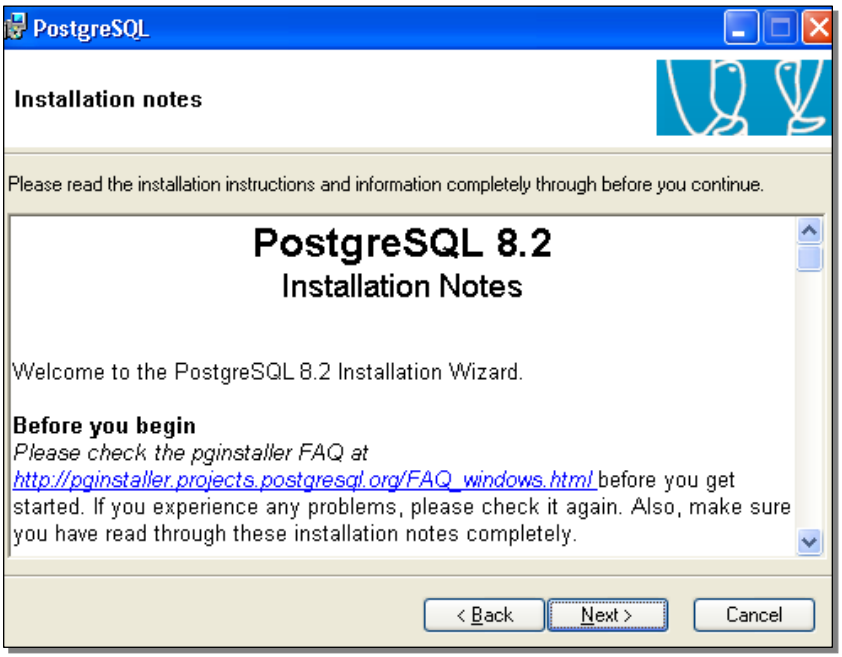
- postgresql-8.2-int (26,973 kb)
- postgresql-8.2 (134 kb)

Dar doble “click” sobre postgresql-8.2 y aparece la siguiente pantalla:

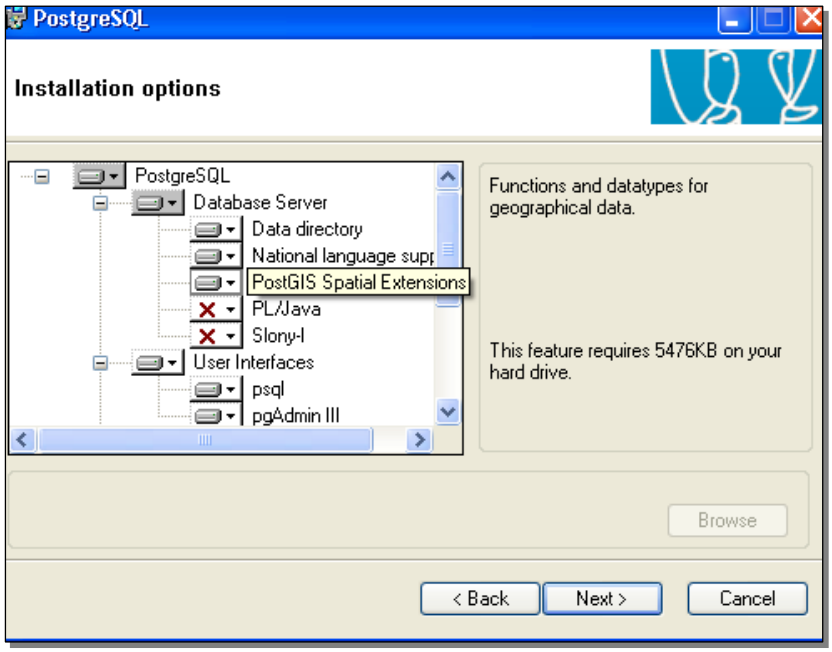
Indica en qué lenguaje se va instalar, se elige por *default* el idioma inglés y aparece la siguiente pantalla, que es el asistente de instalación de PostgreSQL. Indica que será instalado en su forma recomendada y se le da “*click” a Next*.



Aparece la siguiente ventana con las especificaciones de la instalación y la versión del PostgreSQL que será instalada:

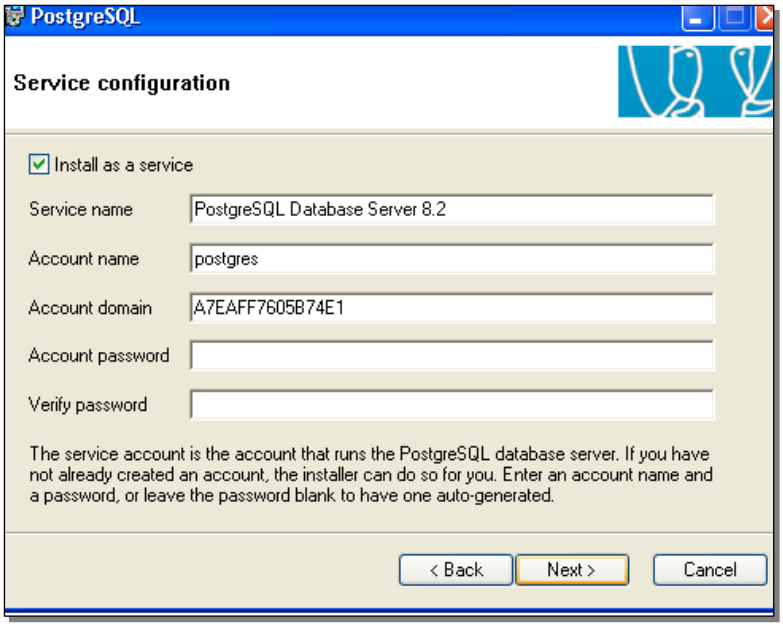


A continuación aparecen las opciones de instalación.

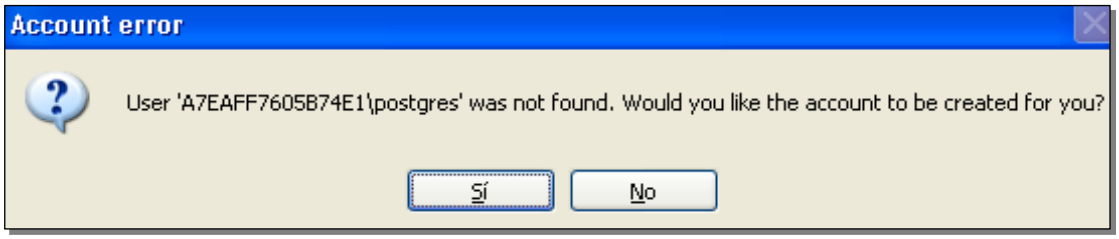


Presenta el espacio en disco con el que se debe de contar, y los módulos e interfaces que se instalarán, se dejan las opciones que vienen marcadas por "*default*",.se elije la opción *next* y despliega la siguiente pantalla, servicios de configuración, donde el PostgreSQL se instalará

Como un servicio, especificando el nombre del servicio: PostgreSQL *Database* Server 8.2, el nombre del usuario de la base de datos que en este caso es el *root* (superusuario) de la base de datos: PostgreSQL; el dominio actual encriptado: A7EAFF7605B74E1, y pregunta el *password* que se le va asignar a la base de datos: PostgreSQL. Se elige *next*:

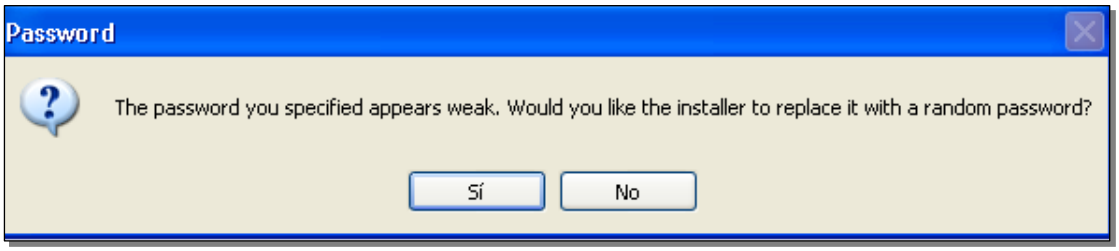


Aparece la siguiente ventana:

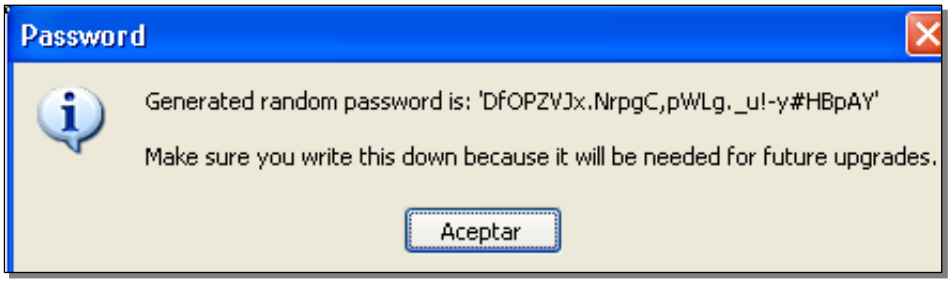


Indica que el usuario A7EAFF7605B74E1\postgres no se encuentra y será creado.

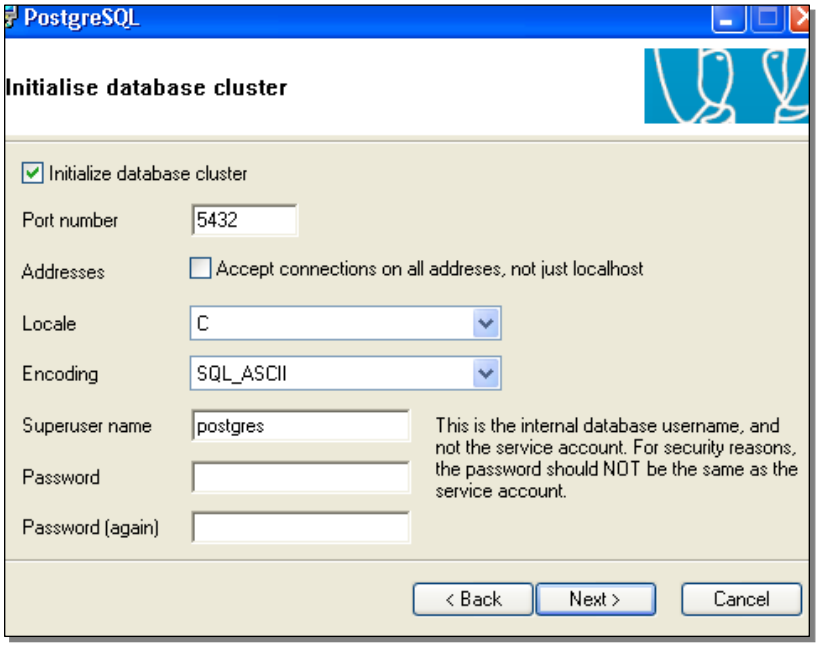
Habrá que elegir la opción Sí. A continuación aparece la ventana que indica que el *password* que se eligió es débil o que no es robusto (es decir, que es muy fácil de adivinar y poner en riesgo las bases de datos), porque es muy fácil de reconocer el *password* y entrar a las bases de datos, violando el contenido de ellas. Indica que si se desea que se reemplace por un *password* más aleatorio, habrá que responder la opción de Sí:



Se elije la opción de Sí y aparece la siguiente ventana con el nuevo *password* que se generó de forma aleatoria:

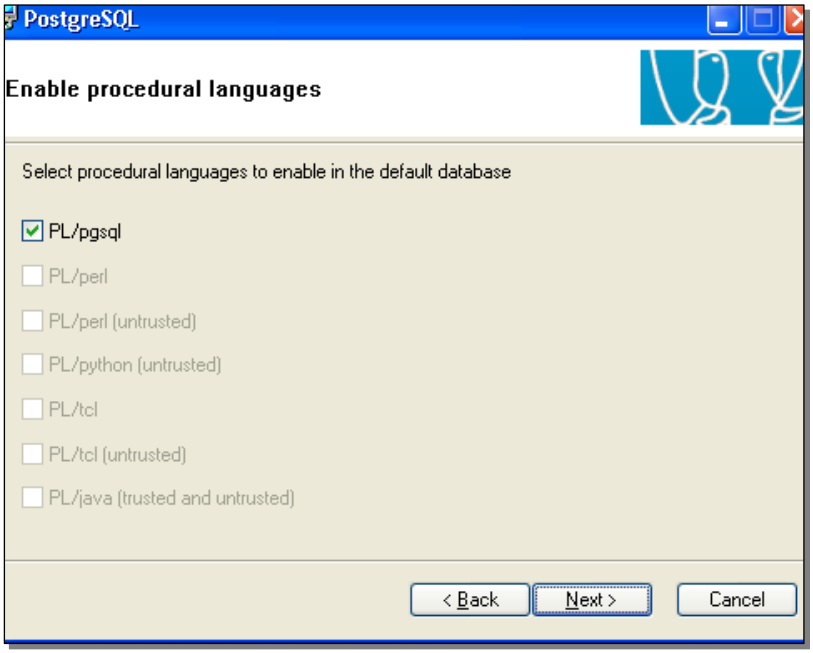


Se elige la opción *Next* para aceptar el nuevo *password* apareciéndonos la siguiente ventana:

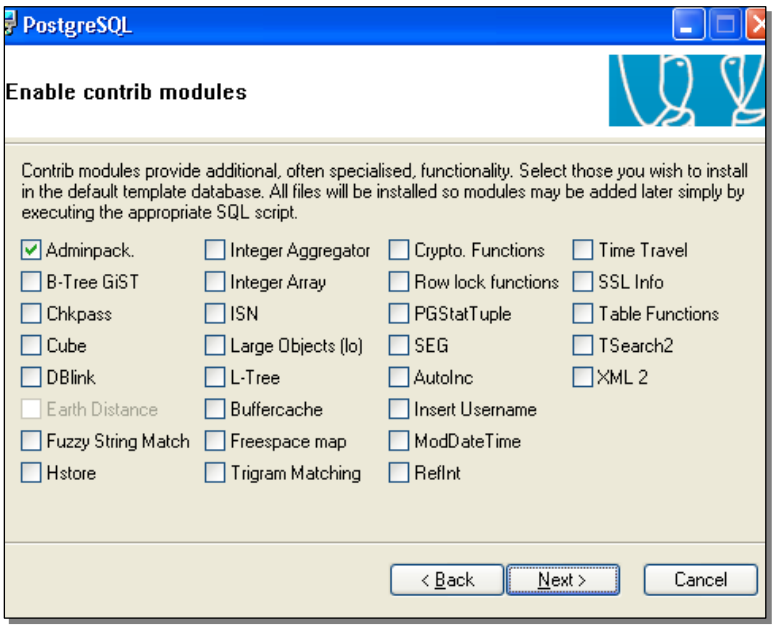


Ésta indica cuál será el puerto de comunicación de la base de datos que es 5432 por *default* (de uso común), y la localidad donde se instalará que es la unidad C, así mismo se indica la codificación *SQL_ASCII* y el nombre del superusuario que es postgres.

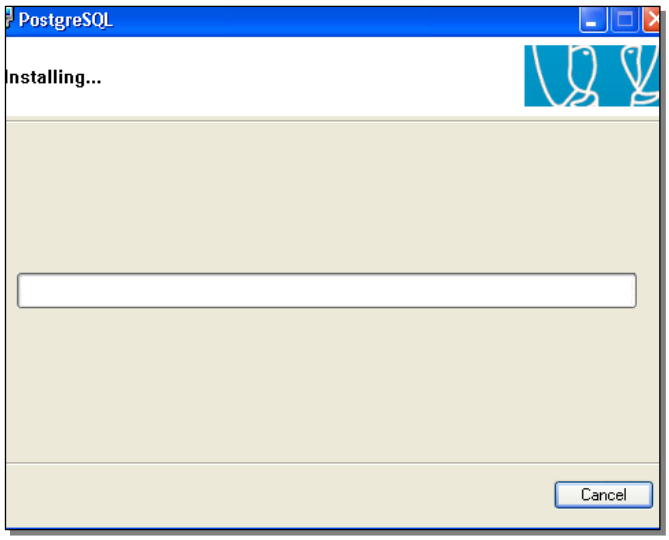
Al mismo tiempo pregunta también por la contraseña del superusuario, el cual será postgresql y cuando se teclea aparece con asteriscos, esto es por seguridad, a continuación se elige la opción *next*, aparece la siguiente ventana:



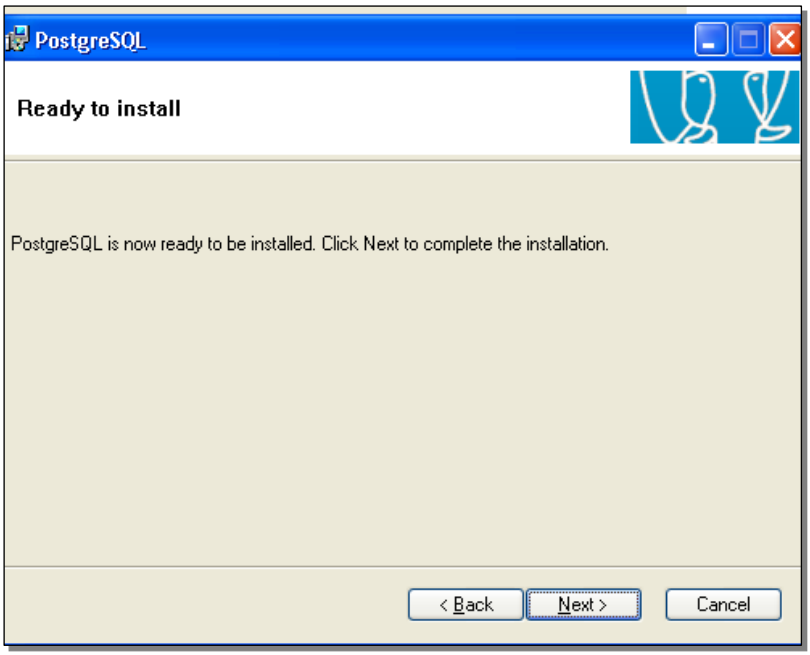
Por *default* se indica el lenguaje de procedimiento a utilizar en la base de datos: *PL/pgsql*, y se debe elegir la opción *Next*. Apareciendo la siguiente ventana, dónde también por *default* aparece la interfaz gráfica para administrar bases de datos *Adminpack*. Apareciendo la siguiente pantalla:



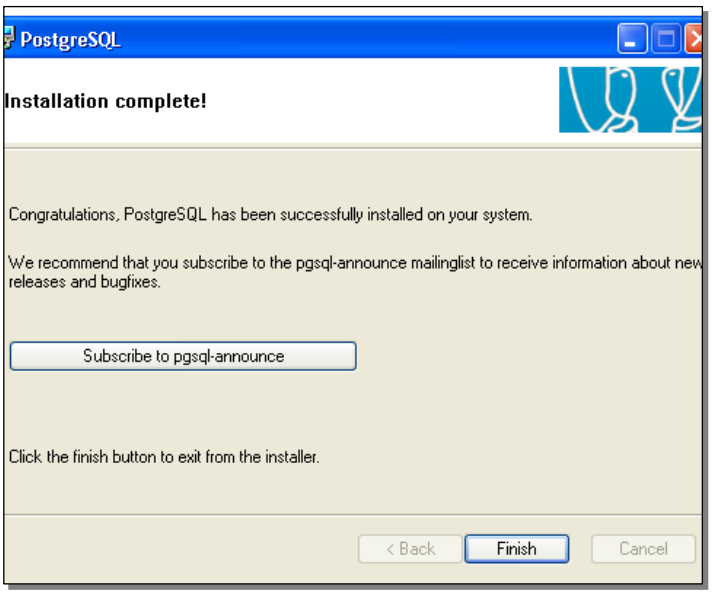
Se elije la opción *Next*, Y comienza la instalación:



Finalmente aparece la siguiente pantalla, la cual indica una instalación completa y se da “click” en finalizar:



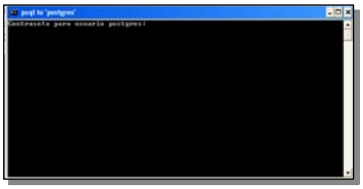
Finalmente se ha instalado PostgreSQL y está listo para usarse.



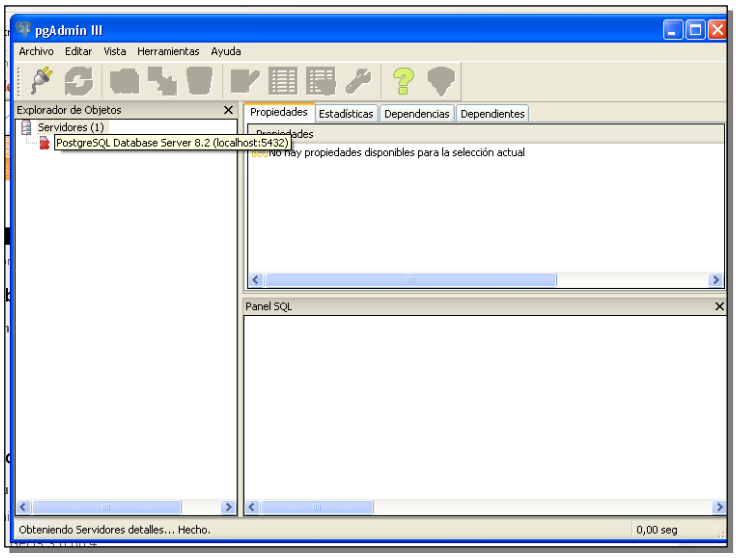
3.1 Inicio de PostgreSQL en Windows XP

Existen dos formas de iniciar PostgreSQL:

- La primera es utilizando la consola, la cual es una interfaz algo rudimentaria con la cual se mantiene comunicación con la base de datos PostgreSQL, se recomienda utilizar esta porque aquí se teclean los comandos básicos para la creación, actualización y gestión de las bases de datos y sus tablas así como sus respectivos atributos de la tablas, tales como los *constraint*, llaves primarias, etc., lo cual permitirá un mejor entendimiento de la creación de bases de datos, aunque sea rudimentaria esto no influye nada en su potencialidad.



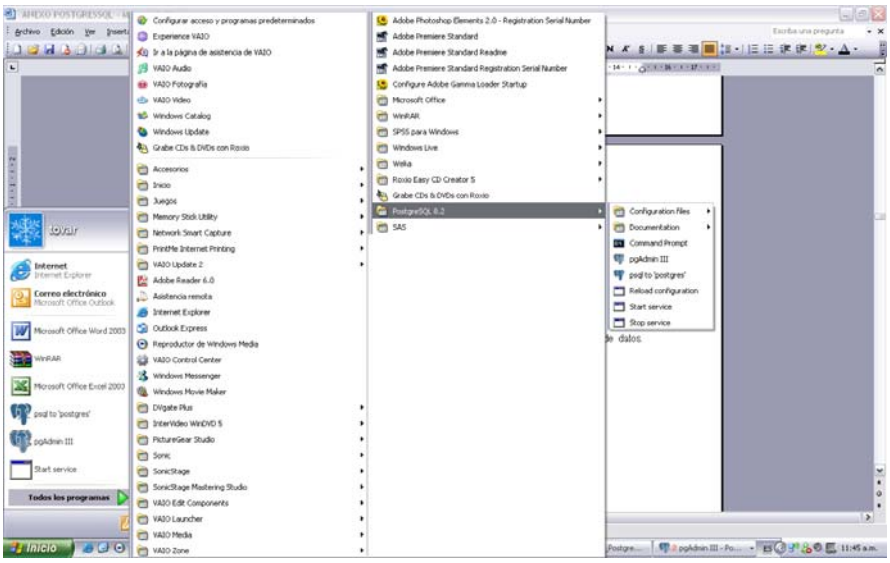
- La segunda forma es utilizar el PGADMIN III, la cual es una presentación mucho más amigable, dada la interfaz gráfica que muestra, pero es lo mismo con muchas más herramientas gráficas, es decir, más controles o funciones que permitirán una mejor gestión (creación, administración y manipulación de las bases de datos)



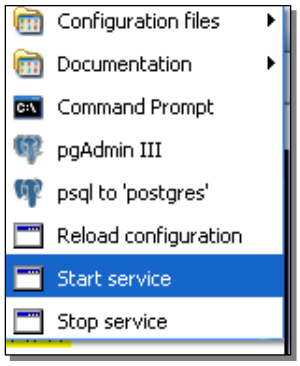
A continuación se revisara la primera opción de cómo accesar al PostgreSQL.

Primera Opción Consola: Se debe dar “click” en:

- Inicio
- Todos los programas
- PostgreSQL 8- 2

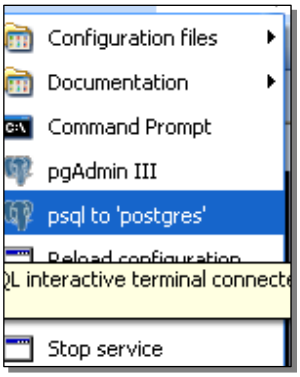


Despliega el siguiente menú:

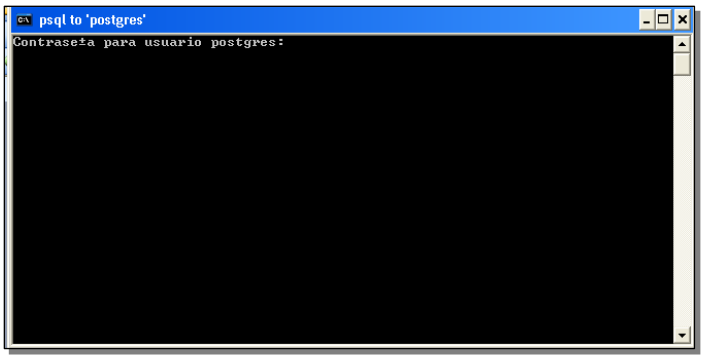


Deberá iniciar el servicio *Start Service*. Es decir se levanta el servidor de base de datos.

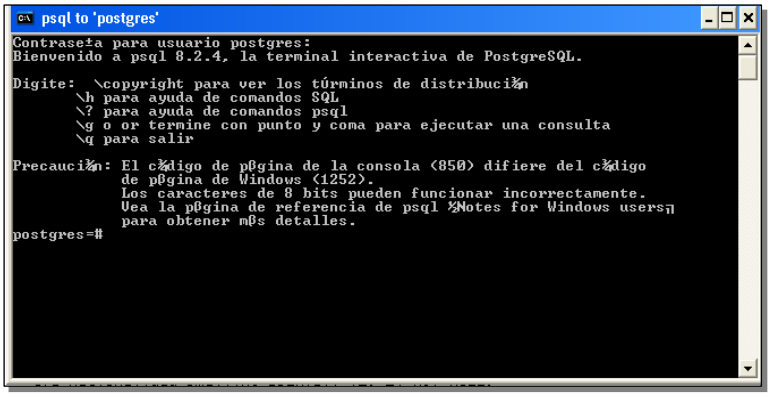
Después se debe elegir la opción *psql to 'postgres'*



Una vez hecho esto deberá elegirse *psql to postgres* apareciendo la siguiente pantalla para entrar a la consola que es la que permite interactuar con las bases de datos:



Se teclea la contraseña de postgres y aparece la siguiente pantalla:



Ya se puede comenzar a crear bases de datos y dentro de ellas las tablas necesarias. Cabe destacar que postgres=# es la línea de comandos (*Prompt*) de la consola, que es la que permite interactuar con el gestor de la base de datos.

Si se teclea en el *Prompt*, cualquiera de los siguientes comandos se obtendrá el resultado que se especifica en la descripción:

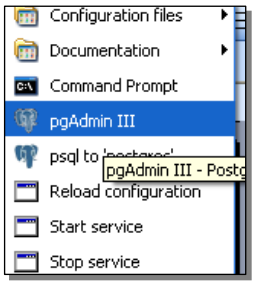
Comando (Prompt)	Descripción
postgres=# \copyright	Proporciona los derechos del PostgreSQL
postgres=# \h	Proporciona ayuda sobre los comandos de SQL
postgres=# \?	Proporciona ayuda sobre comandos de PSQL
postgres=# \g	Termina de ejecutar una consulta
postgres=# \q	Para salir el PostgreSQL

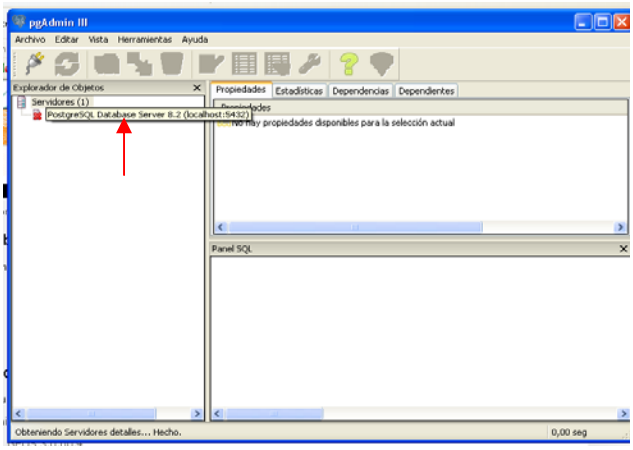
Segunda Opción *pgAdmin* III. Se sigue el mismo procedimiento anterior, se debe dar “click” en:

- Inicio
- Todos los programas
- PostgreSQL 8- 2

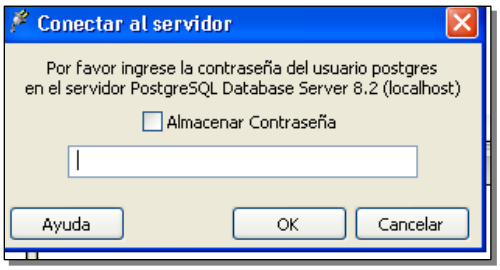
Ya no es necesario levantar el servicio para este caso. Aquí se debe de elegir la opción pgAdmin III.

Apareciendo la siguiente pantalla:

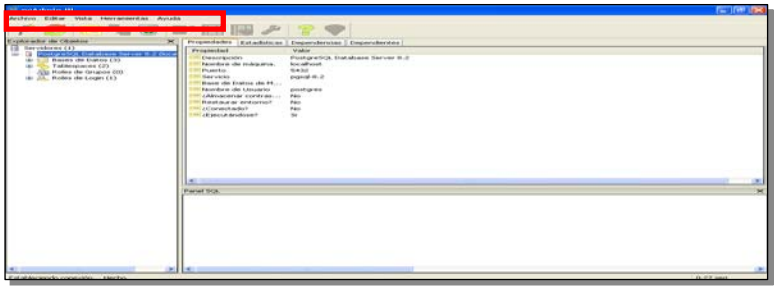




Con doble "click" a *PostgreSQL Database Server 8.2* (*localhost-5432*) y aparece una ventana solicitando que se ingrese el *password* del usuario PostgreSQL, el cual es postgresql:



De esta manera se puede acceder al administrador de bases de datos de PostgreSQL.



La barra de Menú de PGADMIN III, está compuesta por las siguientes opciones

Opción	Tareas
Archivo	Añadir servidor
	Cambiar contraseña
	Opciones
	Abrir Postgresql.conf
	Abrir pg_hba.conf
	Abrir pg_pass.conf
	Salir
Editar	Nuevo Objeto (Esquema, agregado, conversión, dominio, etc.,)
	Crear
	Borrar/eliminar
	Eliminar en Cascada
	Propiedades
Vista	Explorador de objetos
	Panel SQL Server
	Barra de herramientas
	Vistas por defecto
	Refrescar
	Contar
	Objetos del Sistema
Opción	Tareas
Herramientas	Replicación
	Conectar
	Desconectar

	Iniciar el servicio
	Detener el servicio
	Herramientas para consultas
	Scripts
	Ver datos
	Reportes
	Mantenimiento
	Resguardos
	Restaurar
	Asistente para permisos
	Configuración del servidor
	Estado del servidor
	Ayuda de PostgreSQL
	Contenidos de ayuda
	Sugerencias
Ayuda	FAQ PGADMIN III
	Sugerencia de día
	Reporte de error
	Acerca de…

La barra de Iconos (Barra de herramientas) de acceso rápido esta compuesta por:

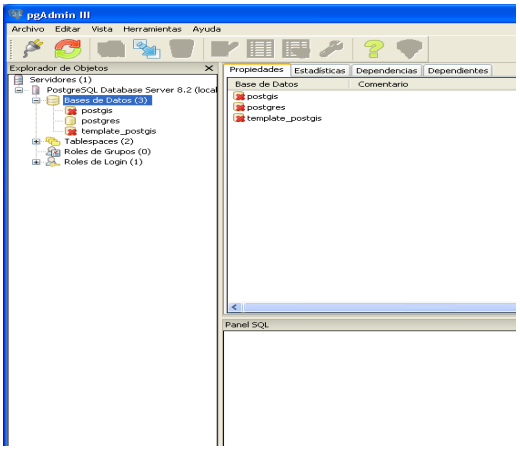
Dónde:

Icono (Tarea)	Descripción de tarea
	Añade una nueva conexión a un servidor
	Actualiza el objeto seleccionado
	Visualiza o edita el objeto seleccionado
	Crea un objeto del mismo tipo que se haya seleccionado
	Elimina el objeto seleccionado
	Ejecuta consultas SQL (Structure Query Lenguaje).
	Ve los datos del objeto seleccionado
	Aplica un filtro y ve los datos del objeto seleccionado
	Mantiene la base de datos o tabla actual
	Muestra sugerencias útiles acerca del objeto actual

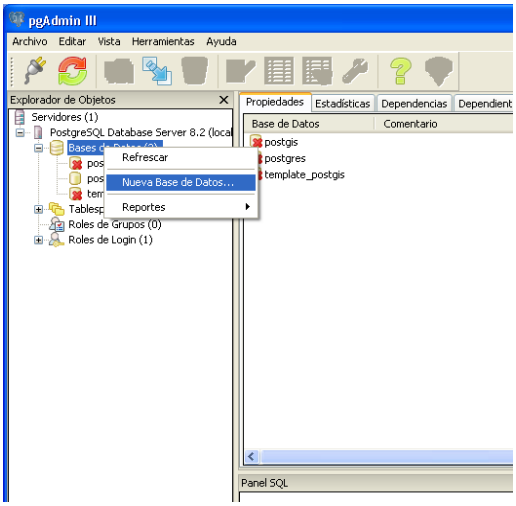
3.2 Creación de una base de datos

A continuación se especificara la manera de crear una base de datos en PostgreSQL, para ello se utilizará el lenguaje SQL el cual trabaja con una secuencia de instrucciones ya definidas para crear estos componentes llamados bases de datos, tablas, procesos almacenados *(store procedures*), etc., al cual se le conoce como sintaxis.

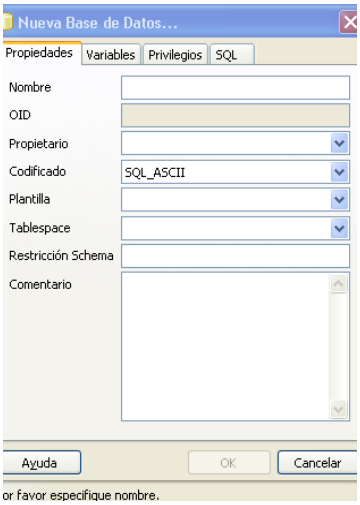
Se da un “click” en el icono de bases de datos (solicita el *password* del administrador de bases de datos)



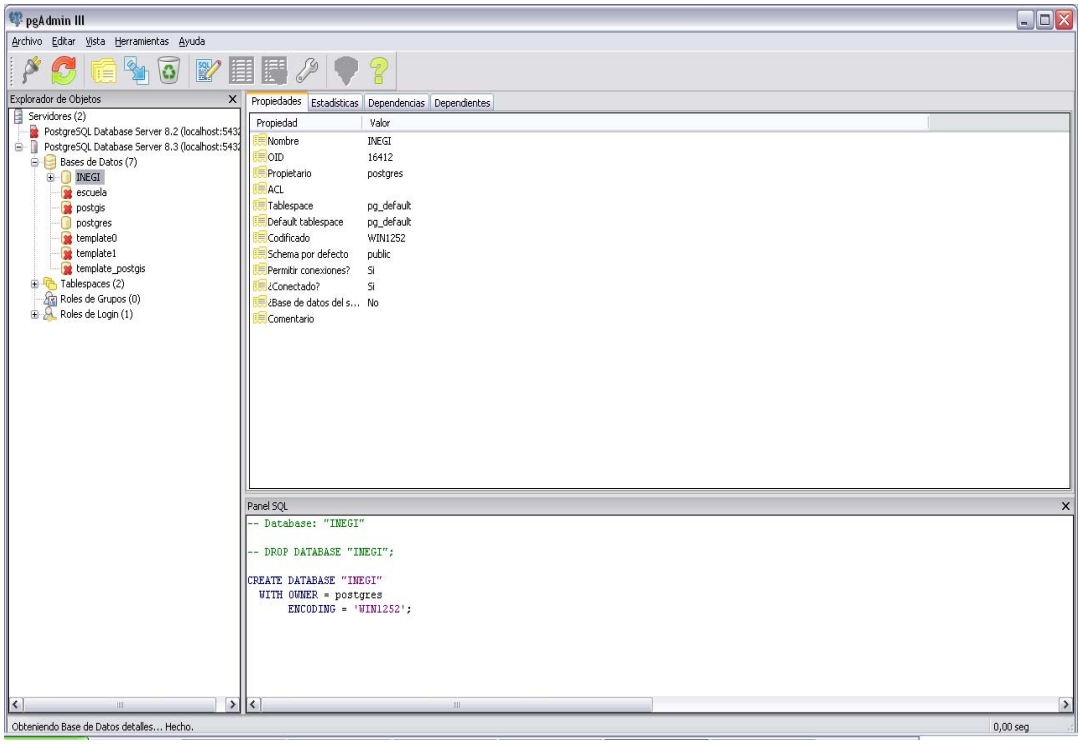
Se le da un “click” sobre el botón derecho del *mouse* apareciendo el menú contextual siguiente y se le da “click” a nueva base de datos



Aparece la siguiente pantalla:



En “Nombre” se debe de colocar el nombre de la base de datos en este caso “INEGI”, y como se puede observar en la siguiente imagen ya se tiene la base de datos.



3.2.1 Creación de tablas

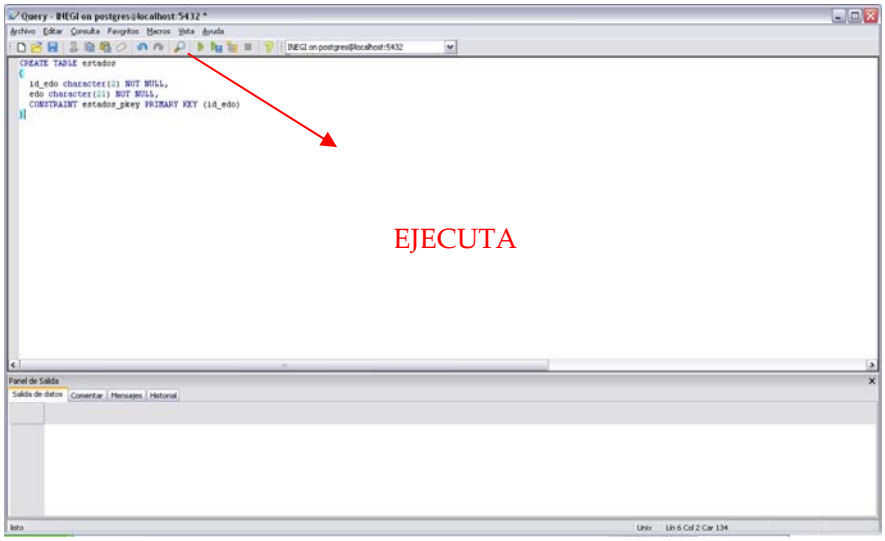
La sintaxis más simple de para la creación de una tabla es:

CREATE TABLE <nombre_tabla>
(
<nombre_campo1> <tipo_de_dato> [NULL | NOT NULL] [DEFAULT <val_predeterminado>],
<nombre_campo2> <tipo_de_dato> [NULL | NOT NULL] [DEFAULT <val_predeterminado>],
<nombre_campoN> <tipo_de_dato> [NULL | NOT NULL] [DEFAULT <val_predeterminado>]
)

Se da "click" al icono  ,

Y aparece la siguiente ventana donde se puede escribir el lenguaje estándar SQL

Donde primero se debe elegir la base de datos INEGI, donde se creara la siguiente tabla con las siguientes instrucciones SQL:



Se ejecuta con el botón "*PLAY*" y si no existen errores de sintaxis la tabla se creara bajo las condiciones descritas en las consultas (*query*), como tipo de dato, llave primaria (*CONSTRAINT – PRIMAY KEY*), si el campo acepta valores nulos *(NOT NULL).*

De esta manera se construyo la base de datos “INEGI”, que cuenta con las siguientes tablas:

TABLA	CAMPO
ESTADOS	ID_EDO
	EDO
CLAS_EDO	ID_EDO
	EDO
	POB
CONAPO	ANIO
	POB
	IND_01
	IND_02
	IND_03
	IND_04
	IND_05
	IND_06
	IND_07
	IND_08
	IND_09
	IND_10
	IND_11
	IND_12
	IND_13
	IND_14
	IND_15
	IND_16

	IND_17
	IND_18
	IND_19
	IND_20
	IND_21
	IND_22
	IND_23
	IND_24
	EDO
TABLA	CAMPO
IND_30	ID_EDO
	EDO
	IND_01
	IND_02
	IND_03
	IND_04
	IND_05
	IND_06
	IND_07
	IND_08
	IND_09
	IND_10
	IND_11
	IND_12
	IND_13
	IND_14

	IND_15
	IND_16
	IND_17
	IND_18
	IND_19
	IND_20
	IND_21
	IND_22
	IND_23
	IND_24
	IND_25
	IND_26
	IND_27
	IND_28
	IND_29
	IND_30
	IND_31
TABLA	CAMPO
IND_MARG_2005	ID_EDO
	EDO
	POB
	IND_01
	IND_02
	IND_03
	IND_04
	IND_05

	IND_06 IND_07 IND_08 IND_09 IND_10 IND_11

El nombre de cada campo se abrevió para que no ocupara espacio innecesario, además de que resulta una buena práctica para cuando se selecciona un campo, el cual entre más breve evita cualquier cantidad de errores, por lo cual se generaron catálogos de la base de datos de cada tabla con sus respectivos campos.

TABLA: ESTADOS			
CAMPO	DESCRIPCIÓN	TIPO	TABLA
ID_EDO	CLAVE DE LA ENTIDAD FEDERATIVA (2 DÍGITOS)	CHAR (2)	ESTADOS
EDO	ESTADO DE LA REPÚBLICA MEXICANA	CHAR (20)	ESTADOS

TABLA: CLAS_EDO			
CAMPO	DESCRIPCIÓN	TIPO	TABLA
ID_EDO	CLAVE DE LA ENTIDAD FEDERATIVA (2 DÍGITOS)	CHAR (2)	CLAS_EDO
EDO	ESTADO DE LA REPÚBLICA MEXICANA	CHAR (20)	CLAS_EDO
POB	POBLACIÓN TOTAL POR ESTADO	INT 8	CLAS_EDO

TABLA: IND_MARG_2005			
CAMPO	DESCRIPCIÓN	TIPO	TABLA
ID_EDO	CLAVE DE LA ENTIDAD FEDERATIVA (2 DÍGITOS)	CHAR (2)	IND_MARG_2005
EDO	EDO	CHAR (21)	IND_MARG_2005
POB	POBLACIÓN TOTAL	INT8	IND_MARG_2005
IND_01	% POBLACIÓN ANALFABETA DE 15 AÑOS O MÁS	FLOAT4	IND_MARG_2005
IND_02	% POBLACIÓN SIN PRIMARIA COMPLETA DE 15 AÑOS O MÁS	FLOAT4	IND_MARG_2005
IND_03	% OCUPANTES EN VIVIENDAS SIN DRENAJE NI SERVICIO SANITARIO	FLOAT4	IND_MARG_2005
IND_04	% OCUPANTES EN VIVIENDAS SIN ENERGÍA ELÉCTRICA	FLOAT4	IND_MARG_2005
IND_05	% OCUPANTES EN VIVIENDAS SIN AGUA ENTUBADA	FLOAT4	IND_MARG_2005
IND_06	% VIVIENDAS CON ALGÚN NIVEL DE HACINAMIENTO	FLOAT4	IND_MARG_2005
IND_07	% OCUPANTES EN VIVIENDAS CON PISO DE TIERRA	FLOAT4	IND_MARG_2005
IND_08	% POBLACIÓN EN LOCALIDADES CON MENOS DE 5 000 HABITANTES	FLOAT4	IND_MARG_2005
IND_09	% POBLACIÓN OCUPADA CON INGRESO DE HASTA 2 SALARIOS MÍNIMOS	FLOAT4	IND_MARG_2005
IND_10	ÍNDICE DE MARGINACIÓN	FLOAT4	IND_MARG_2005
IND_11	GRADO DE MARGINACIÓN	CHAR (8)	IND_MARG_2005
IND_12	LUGAR QUE OCUPA EN EL CONTEXTO NACIONAL	INT2	IND_MARG_2005

TABLA: IND_30			
CAMPO	DESCRIPCION	TIPO	TABLA
ID_EDO	CLAVE DE LA ENTIDAD FEDERATIVA (2 DÍGITOS)	CHAR (2)	IND_30
IND_01	PORCENTAJE DE POBLACIÓN EN VIVIENDAS CON AGUA ENTUBADA EN EL ÁMBITO DE LA VIVIENDA	FLOAT4	IND_30
IND_02	PORCENTAJE DE POBLACIÓN EN VIVIENDAS CON ENERGÍA ELÉCTRICA	FLOAT4	IND_30
IND_03	PORCENTAJE DE POBLACIÓN EN VIVIENDAS CON DRENAJE	FLOAT4	IND_30
IND_04	PORCENTAJE DE POBLACIÓN EN VIVIENDAS CON PISO DIFERENTE DE TIERRA	FLOAT4	IND_30
IND_05	PORCENTAJE DE POBLACIÓN EN VIVIENDAS CON PAREDES DE MATERIALES DURABLES	FLOAT4	IND_30
IND_06	PORCENTAJE DE POBLACIÓN EN VIVIENDAS CON TECHOS DE MATERIALES DURABLES	FLOAT4	IND_30
IND_07	PORCENTAJE DE POBLACIÓN EN VIVIENDAS SIN HACINAMIENTO	FLOAT4	IND_30
IND_08	PORCENTAJE DE POBLACIÓN EN VIVIENDAS CON SERVICIO SANITARIO EXCLUSIVO	FLOAT4	IND_30
IND_09	PORCENTAJE DE POBLACIÓN EN VIVIENDAS QUE USAN GAS O ELECTRICIDAD PARA COCINAR	FLOAT4	IND_30
IND_10	PORCENTAJE DE POBLACIÓN EN VIVIENDAS CON REFRIGERADOR	FLOAT4	IND_30
IND_11	PORCENTAJE DE POBLACIÓN EN VIVIENDAS CON RADIO, RADIOGRABADORA O TELEVISIÓN	FLOAT4	IND_30
IND_12	PORCENTAJE DE POBLACIÓN EN VIVIENDAS CON TELÉFONO	FLOAT4	IND_30
IND_13	PORCENTAJE DE POBLACIÓN EN VIVIENDAS CON AUTOMÓVIL O CAMIONETA PROPIOS	FLOAT4	IND_30
IND_14	PORCENTAJE DE POBLACIÓN CON DERECHOHABIENCIA A SERVICIOS DE SALUD	FLOAT4	IND_30
IND_15	PORCENTAJE DE POBLACIÓN DE 15 AÑOS Y MÁS ALFABETA	FLOAT4	IND_30
IND_16	PORCENTAJE DE NIÑOS DE 6 A 14 AÑOS QUE ASISTEN A LA ESCUELA	FLOAT4	IND_30
IND_17	PORCENTAJE DE ADOLESCENTES DE 12 A 17 AÑOS QUE ASISTEN A LA ESCUELA	FLOAT4	IND_30
IND_18	PORCENTAJE DE POBLACIÓN DE 15 AÑOS Y MÁS CON INSTRUCCIÓN POSTPRIMARIA	FLOAT4	IND_30
IND_19	PORCENTAJE DE POBLACIÓN OCUPADA FEMENINA	FLOAT4	IND_30
IND_20	PORCENTAJE DE POBLACIÓN ECONÓMICAMENTE ACTIVA ENTRE 20 Y 49 AÑOS	FLOAT4	IND_30
IND_21	PERCEPTORES POR CADA 100 PERSONAS	FLOAT4	IND_30
IND_22	PORCENTAJE DE POBLACIÓN OCUPADA QUE PERCIBE MÁS DE DOS Y MEDIO SALARIOS MÍNIMOS	FLOAT4	IND_30
IND_23	PORCENTAJE DE POBLACIÓN OCUPADA QUE PERCIBE MÁS DE CINCO SALARIOS MÍNIMOS	FLOAT4	IND_30
IND_24	PORCENTAJE DE POBLACIÓN EN HOGARES QUE PERCIBEN MÁS DE $10.42 DIARIOS POR PERSONA.	FLOAT4	IND_30
IND_25	PORCENTAJE DE POBLACIÓN OCUPADA QUE SON TRABAJADORES FAMILIARES SIN PAGO	FLOAT4	IND_30
IND_26	PORCENTAJE DE POBLACIÓN OCUPADA EN EL SECTOR TERCIARIO FORMAL	FLOAT4	IND_30
IND_27	PORCENTAJE DE POBLACIÓN OCUPADA QUE SON PROFESIONISTAS O TÉCNICOS	FLOAT4	IND_30
IND_28	PORCENTAJE DE HIJOS SOBREVIVIENTES DE MUJERES DE 20 A 34 AÑOS DE EDAD	FLOAT4	IND_30
IND_29	SEGREGACIÓN DE GÉNERO EN TÉRMINOS DE ALFABETISMO	FLOAT4	IND_30
IND_30	PORCENTAJE DE POBLACIÓN ECONÓMICAMENTE INACTIVA DE 65 AÑOS Y MÁS QUE ES JUBILADA O PENSIONADA	FLOAT4	IND_30

TABLA CONAPO			
CAMPO	DESCRIPCION	TIPO	TABLA
ANIO	ANIO	CHAR(4)	CONAPO
POB	POBLACI_N_A_MITAD_DE_A_O	INT8	CONAPO
IND_01	HOMBRES	INT8	CONAPO
IND_02	MUJERES	INT8	CONAPO
IND_03	NACIMIENTOS	INT8	CONAPO
IND_04	DEFUNCIONES	INT8	CONAPO
IND_05	CRECIMIENTO_NATURAL	INT8	CONAPO
IND_06	INMIGRANTES_INTERESTATALES	INT8	CONAPO
IND_07	EMIGRANTES_INTERESTATALES	INT8	CONAPO
IND_08	MIGRACI_N_NETA_INTERESTATAL	INT8	CONAPO
IND_09	MIGRACI_N_NETA_INTERNACIONAL	INT8	CONAPO
IND_10	CRECIMIENTO_SOCIAL_TOTAL	INT8	CONAPO
IND_11	CRECIMIENTO_TOTAL	INT8	CONAPO
IND_12	TASA_BRUTA_DE_NATALIDAD_	FLOAT4	CONAPO
IND_13	TASA_BRUTA_DE_MORTALIDAD_	FLOAT4	CONAPO
IND_14	TASA_DE_CRECIMIENTO_NATURAL__	FLOAT4	CONAPO
IND_15	TASA_DE_INMIGRACI_N_INTERESTATAL	FLOAT4	CONAPO
IND_16	TASA_DE_EMIGRACI_N_INTERESTATAL_	FLOAT4	CONAPO
IND_17	TASA_DE_MIGRACI_N_NETA_INTERESTAL	FLOAT4	CONAPO
IND_18	TASA_DE_MIGRACI_N_NETA_INTERNACIONAL	FLOAT4	CONAPO
IND_19	TASA_DE_CRECIMIENTO_SOCIAL_TOTAL	FLOAT4	CONAPO
IND_20	TASA_DE_CRECIMIENTO_TOTAL__	FLOAT4	CONAPO
IND_21	TASA_GLOBAL_DE_FECUNDIDAD	FLOAT4	CONAPO
IND_22	ESPERANZA_DE_VIDA_TOTAL	FLOAT4	CONAPO
IND_23	ESPERANZA_DE_VIDA_HOMBRES	FLOAT4	CONAPO
IND_24	ESPERANZA_DE_VIDA_MUJERES	FLOAT4	CONAPO
EDO	EDO	CHAR(21)	CONAPO