UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

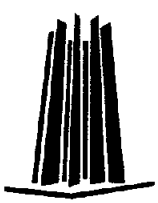
FACULTAD DE ESTUDIOS SUPERIORES
ARAGÓN

"LA MINERÍA DE DATOS Y SU IMPLEMENTACIÓN EN LA ADMINISTRACIÓN ESCOLAR DEL COLEGIO DE BACHILLERES"

T R A B A J O E S C R I T O
EN LA MODALIDAD DE TESIS
QUE PARA OBTENER EL TÍTULO DE:
INGENIERO EN COMPUTACIÓN
P R E S E N T A :

L U I S U B A L D O
G O D Í N E Z F L O R E S

ASESOR: ING. RODOLFO VÁZQUEZ MORALES



MÉXICO, 2005.

A mis padres **Arturo** *y* **Hortensia** *por cariño y apoyo a lo largo de todos es años, a* **Lenia** *para que se motive y term su carrera, a mi jechu* **Aurora** (†) *por atenciones que me brindó en su vida. todos los que me han acompañado en camino que sin duda lleva al éxito, el de perseverancia. Por último, y no por menos importante, quiero expresar todo agradecimiento al Ingeniero* **Rodo Vázquez Morales** *por su disposición invaluable consejo y asesoría, ya que sin este trabajo no hubiera sido posible.*

INDICE

OBJETIVOS.

OBJETIVO GENERAL.

Desarrollar una investigación sobre la Minería de Datos, sus tendencias y alcances a fin de que sirva como una referencia básica sobre el tema, consultando textos especializados y realizando una aplicación real para el Plantel 11 "Nueva Atzacoalco" del Colegio de Bachilleres.

OBJETIVOS PARTICULARES.

- Conocer los avances en el campo de la Minería de Datos.
- Desarrollar una herramienta de análisis para los datos académicos que se generan en el Colegio.
- Generar modelos de Minería de Datos para realizar predicciones de indicadores como Egreso, Regularidad, Permanencia y Superación de la matrícula del Plantel.
- Centralizar datos administrativos como: horarios, claves de materia, categorías y tabulador de sueldos y salarios.
- Centralizar datos académicos, como: cursos, situación escolar de los profesores y desarrollo profesional.
- Implementar instrumentos de visualización, para generar reportes acordes a las necesidades de información del personal directivo.
- Automatizar procesos y trámites escolares en el Colegio de Bachilleres.
- Mostrar un panorama histórico del comportamiento académico del alumnado del Plantel.

"-Adso –dijo Guillermo - , resolver un misterio no es como deducir a partir de primeros principios. Y tampoco es como recoger un montón de datos particulares para inferir después una ley general. Equivale más bien a encontrarse con uno, dos o tres datos particulares que al parecer no tienen nada en común, y tratar de imaginar si pueden ser otros tantos casos de una ley general que todavía no se conoce, y que quizá nunca ha sido enunciada."

Umberto Eco
El nombre de la rosa, p.434

Introducción.

El presente trabajo es una investigación sobre la Minería de Datos, sus alcances y sus perspectivas.

Busca ser una referencia para introducir a los jóvenes investigadores a este nuevo campo del conocimiento computacional. Debo mencionar, en este sentido, la dificultad de encontrar material impreso en el idioma español no así en el inglés donde las publicaciones son muy bastas tal vez por la importancia y el valor que tiene para las organizaciones la información significativa.

Aunque solo se abordan dos técnicas de minería de datos, los árboles de decisión y clustering, por ser los utilizados en los Analysis Services de Microsoft SQL Server 2000, se debe mencionar que existen técnicas que incorporan los nuevos avances en Aprendizaje Automático como son las Redes Neuronales y la Lógica Difusa, de las cuales se encuentra un basto catálogo de documentos en Internet.

También se incorporo el proceso que se siguió para desarrollar una aplicación de minería de datos para el Colegio de Bachilleres la cual busca, por una parte, demostrar que es posible la inclusión de tecnología que es clasificada como costosa en el ámbito público y por otro lado su aplicación en la toma de decisiones en el sector educativo.

Se abordo el tema de la Minería de Datos como un proceso cíclico que incluye el almacenamiento de datos (Data Warehousing), el análisis y búsqueda de conocimiento (Aprendizaje Automático) y toma de decisiones (I Bussines), y no solo la parte de análisis y aprendizaje automático.

La tesis está estructurada de la siguiente manera: el capitulo uno aborda de manera breve el concepto de Minería de Datos y parte de sus historia y tendencias de desarrollo, el capitulo dos hace una revisión de lo que es un Data Warehousing y su importancia dentro de una organización. El capitulo tres describe los fundamentos estadísticos y matemáticos de los árboles de decisión y de las técnicas de clustering o clasificación así como una introducción al aprendizaje automático. En el capitulo cuatro de aborda el proceso completo que conlleva emprender un proyecto de minería de datos así como también los retos a vencer para el éxito del mismo.

Para el capitulo cinco se recopilan algunas empresas, comunidades y productos dedicados al desarrollo de aplicaciones de minería de datos.

El capitulo seis contiene el desarrollo de una aplicación real de minado de datos donde se incluye la estructuración de un Data Warehouse y el proceso de creación de un modelo de minería que predice las posibilidades de éxito de egreso de un estudiante en base a los datos proporcionados cuando ingresa a la institución.

Finalmente en el capitulo siete se colocaron las conclusiones sobre el proyecto y su éxito.

1. ¿Qué es la Minería de Datos?

1.1. Concepto.

La necesidad en las empresas de tener a la mano aplicaciones que les permitan obtener información rápida y confiable para la toma de decisiones ha llevado al desarrollo del concepto de Minería de Datos.

Podemos definir a la MD de la siguiente forma:

"Una actividad de extracción cuyo objetivo es el de descubrir hechos contenidos en las bases de datos" [1]

La MD es un conjunto de técnicas desarrolladas en el campo de la Inteligencia Artificial, en particular del área de Reconocimiento de Patrones, que permiten encontrar patrones de comportamiento en los datos, contenidos en grandes almacenes de información. Estos patrones y tendencias son difíciles de detectar debido a los grandes volúmenes de información. Se pueden utilizar varios algoritmos para minar los datos.

"Al Descubrimiento de Conocimiento de Bases de Datos (KDD)[2] a veces también se le conoce como minería de datos (*Data Mining*).

Sin embargo, muchos autores se refieren al proceso de minería de datos como el de la aplicación de un algoritmo para extraer patrones de datos y a KDD al proceso completo (pre-procesamiento, minería, post-procesamiento) ".[3]

Se calcula aproximadamente que el tamaño de la información contenida en las macro empresas se duplica cada 20 meses, sin embargo la velocidad de

[1] http://answermath.com/data-mining/mineria-de-datos-2-concepto.htm, consultada el 20 de marzo de 2004
[2] Iniciales en inglés para descubrimiento del conocimiento en datos (Knowledge Data Discovering).
[3] F. Morales Eduardo, http://dns1.mor.itesm.mx/~emorales/Cursos/KDD01/principal.html, ITESM, consultada el 20 de marzo de 2004

procesamiento y análisis no aumentan a ese ritmo. Por otra parte las técnicas tradicionales tienen ya mucho tiempo funcionando y no responden en muchos casos a las necesidades de las empresas.

El proceso de extracción del conocimiento se caracteriza por identificar patrones y tendencias que sean validos, novedosos, útiles y entendibles. Estas medidas de utilidad y validez se establecen por los usuarios y los diseñadores de la aplicación.

Las metas de la MD son procesar grandes cantidades de información, identificar los patrones más significativos y relevantes para presentarlos como conocimiento apropiado para los fines del usuario.

El proceso de KDD involucra varias disciplinas entre las que se encuentran las siguientes:

- Tecnología de base de datos y data warehousing.
- Aprendizaje automático (IA).
- Reconocimiento de patrones.
- Visualización.
- Cómputo de alto desempeño.
- Análisis estadístico.

Los sistemas de minado de datos deben proporcionar de manera automática hipótesis, después de un análisis exhaustivo de los datos, sobre el comportamiento de la empresa u organización en la cual se esté implementando.

1.2. Componentes del KDD

El Descubrimiento de Conocimiento de Bases de Datos (KDD) incluye ciertos componentes que se encuentran integrados en la figura 1.

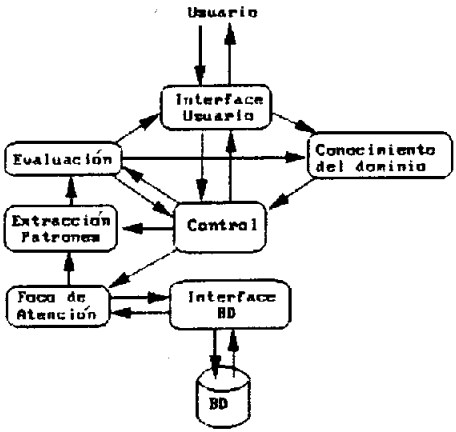


Figura 1. Componentes de KDD

Dentro de las partes que componen al proceso del Descubrimiento de Conocimiento en Bases de Datos tenemos las siguientes:

- **Conocimiento del dominio y preferencias del usuario:** Establece las necesidades del usuario y el conocimiento previo de la problemática y manera como funciona la organización. Busca orientar y ayudar en la búsqueda de patrones interesantes. Debe realizarse un balance entre eficiencia y que tan completo es el del conocimiento que se tiene.

- **Control del descubrimiento:** Se refiere a que hacer con el conocimiento obtenido y está a cargo del usuario.

- **Interfaces:** Entre la base de datos y el usuario.

- **Foco de atención:** Especifica qué tablas, campos y registros acceder. Tiene mecanismos de selección aleatoria de registros tomando muestras estadísticas significativas. Algunas técnicas para enfocar la atención incluyen: agregación (juntar valores), partición de datos y proyección.

- **Extracción de patrones:** Un patrón se refiere a cualquier relación entre los elementos de la base de datos. En esta fase se aplican los algoritmos de extracción.

- **Evaluación:** Es importante validar el conocimiento obtenido en base a los criterios del usuario. Los algoritmos de extracción están basados en significancia estadística, que no necesariamente puede representar información importante o interesante para la empresa.

1.3. Fases y características de un proyecto de Minería de Datos.

Las fases que integran un proyecto de minería de datos se encuentran integradas por selección, procesado, selección de características, extracción de conocimiento y evaluación hasta llegar al modelo clasificador del conocimiento. Fig. 2.

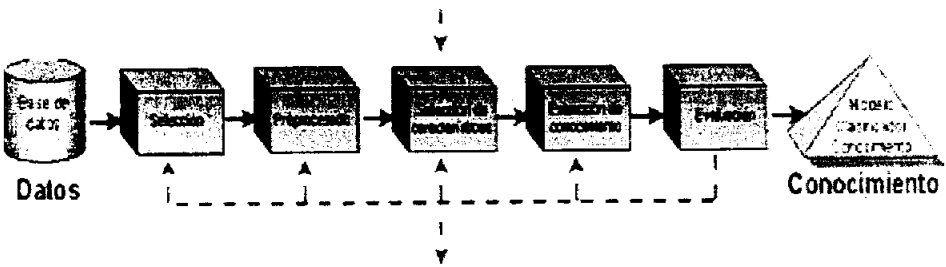


Figura 2. Fases de un proyecto de Minería de Datos.

Un estándar industrial utilizado por más de 160 empresas e instituciones, dedicadas a proyectos de extracción del conocimiento de todo el mundo llamado **CRISP-DM** (*CRoss Industry Standard Process for Data Mining*) establece que los pasos a seguir en un proyecto de MD son[4]:

[4] http://www.crisp-dm.org , consultada el 23 de marzo de 2004

1. **Filtrado de Datos**: En esta etapa se limpian los datos obtenidos de una base o un data warehouse eliminando valores incorrectos, no validos, se obtienen muestras significativas o se reduce el número de valores posibles.

2. **Selección de características**: La selección de características reduce el tamaño de los datos eligiendo las variables más influyentes en el problema, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería.

Los métodos para la selección de características son dos:

- Los que están basados en la elección de los mejores atributos del problema.
- Los que buscan variables independientes mediante test de sensibilidad, algoritmos de distancia o heurísticos.

3. **Extracción del conocimiento:** Mediante algún algoritmo se extrae un modelo de conocimiento que representa patrones de comportamiento y tendencias en las variables del problema. Dependiendo del algoritmo para la extracción del conocimiento se debe usar un preprocesado diferente.

4. **Interpretación y evaluación:** Después de generar el modelo del conocimiento se debe proceder a interpretar y evaluar los resultados, si estos no fueran los deseados o no proporcionan la información requerida se debe proceder a modificar alguno de los pasos anteriores.

Una aplicación que implemente Minería de Datos tiene las siguientes características en su diseño:

- **Rapidez:** La información requerida debe ser entregada al usuario final en un periodo máximo de cinco segundos.

- **Análisis:** El sistema debe ser capaz de entregar análisis numérico y estadístico de los datos.

- **Compartido:** La aplicación debe implementar las suficientes medidas de seguridad para garantizar la confidencialidad de los datos a través de una larga base de usuarios.

- **Multidimensional:** Los datos no deben ser considerados solo como un conjunto de filas y columnas sino que además deben tener otras dimensiones como períodos de tiempo. Un data warehouse podría soportar diferentes niveles y jerarquías exactamente como la empresa.

- **Información:** Todos los datos requeridos deben de estar al alcance de la aplicación, sin considerar donde se encuentren y sin restricción en el volumen de la información.

En los capítulos siguientes se tratarán con más detalle cada una de las etapas del proyecto de MD.

1.4. Objetivos de la Minería de Datos.

El objetivo general al implementar proyectos de Minería de Datos son: encontrar patrones de comportamiento que permitan a los trabajadores del conocimiento fundamentar las decisiones que se toman en las empresas.

Para esto se busca encontrar ***reglas de asociación***, las cuales se dividen en dos: ***reglas de correlación*** (ejemplo: Si un cliente compra un video hoy, es muy probable que en los próximos días compre películas de video) y ***reglas de***

clasificación (ejemplo: Si un cliente tiene ingresos superiores a $75,000 es un buen sujeto para concederle un crédito).

Otro objetivo es el de realizar tareas de descripción y predicción de las operaciones que se realizan dentro de las organizaciones. La descripción se usa para dar un análisis preliminar de los datos.

En el caso de la predicción se tienen dos tipos: La ***clasificación*** que es el proceso de inducir un modelo para poder predecir una clase, considerando los valores de los atributos. Utiliza árboles de decisión, reglas y análisis de discriminantes. Por otro lado se tiene a la ***estimación o regresión*** que busca establecer un modelo para poder predecir el valor de la clase dados los valores de los atributos. En ésta se utilizan árboles de regresión, regresión lineal, redes neuronales, kNN,etc.

1.5 Historia de los Sistemas de MD

Algunos de los sistemas que más han influido en el desarrollo actual de la minería de datos son los siguientes:

AM (Lenat '79): Uno de los más famosos, simulaba el proceso que hace un matemático para encontrar nuevos conceptos y relaciones entre ellos usando heurísticas.

Eurisko (Lenat): Se basa en que se puede conseguir nuevo conocimiento usando heurísticas. Al surgir nuevo conocimiento se requieren nuevas heurísticas. Nuevas heurísticas se pueden obtener de heurísticas. Al surgir nuevo conocimiento se requieren nuevas representaciones. Nuevas representaciones se pueden obtener con heurísticas.

BACON (Langley '81 -83): Tenia como entrada un conjunto de variables dependientes e independientes + datos y entregaba como salida leyes cuantitativas que relacionan las variables.

"Glauber: Descubre leyes cualitativas cambiando las substancias específicas por clases abstractas (e.g., acido o alcalino) y determinando el nivel de generalidad de cada ley mediante cuantificadores universales y existenciales.

Stahl: Determina componentes de varias substancias. Por ejemplo (componentes de {agua} son {hidrógeno oxígeno}).

Utiliza varios operadores (heurísticas) para ello. Sigue una búsqueda depth-first sin backtraking.

Dalton: A partir de reacciones y componentes de las substancias involucradas infiere un modelo de cada reacción especificando el número de moléculas y el número de partículas en cada compuesto."[5]

El principal fin de estos sistemas era simular el proceso de descubrimiento del conocimiento en los humanos por medio de heuristicas.

1.6 Retos y tendencias en el desarrollo de aplicaciones de Minería de Datos.

El incremento en la capacidad de procesamiento de la información así como el aumento en el volumen de datos que manejan las organizaciones hace necesario implementar nuevas técnicas que permitan automatizar las tareas de análisis y de toma de decisiones.

[5] F. Morales Eduardo.: *ibid.*

De esta forma los sistemas basados en MD han encontrado gran aceptación en el área de ventas ayudando a predecir comportamientos de mercado, en el área médica y en la Internet con el minado de texto.

Actualmente se cuenta en el mercado con herramientas genéricas y empresas dedicadas a la extracción del conocimiento.

El surgimiento del comercio electrónico abre la posibilidad a las empresas de cubrir un mercado más amplio y la necesidad de conocer con más detalle las preferencias y gustos de sus clientes. El crecimiento de las bases de datos que contienen información relevante de los clientes a las que constantemente se agregan más registros hace necesaria la implementación de herramientas automatizadas para explorar su contenido.

Sin embargo a veces el llenado incompleto de los registro dificulta el proceso de la minería de datos. Por otro lado existen bases de datos no relacionales y mal diseñadas que también complican el proceso de implementar algoritmos para la extracción de patrones relevantes.

"... los retos incluyen tener datos incompletos e inexactos, insuficientes herramientas y recursos, administradores no comprometidos y un continuo cambio en los datos. Además se deben minar bases de datos distribuidas, minado de bases de datos de sitios web, seguridad y privacidad del minado de datos."[6]

[6] THURAISINGHAM, Bhavani. : *Data Mining Technologies, techniques, tool and trends* CRC Press, USA, 1999, p. 219.

1.7. Ventajas y desventajas de la Minería de Datos.

"Tomar decisiones es el trabajo más importante de un funcionario, también es el más duro y el más riesgoso. Una mala decisión puede echar a perder una organización o una carrera."[7]

En efecto la toma de decisiones es de las tareas mas complejas que tiene un ejecutivo o funcionario de alto nivel, la minería de datos plantea proveer de las herramientas necesarias para que los analistas se conviertan de generadores de reportes en knowledge workers, "estas personas serán capaces de analizar y entender los diferentes aspectos de los factores que inciden en el desempeño de la organización"[8].

"Los beneficios de un Data Warehouse son concentrar datos dispersos y generar información integrada; mejorar la oportunidad con la que se genera y entrega información; elevar la calidad de la información y reducir la incertidumbre respecto a la confiabilidad."[9]

Podemos apreciar que la minería de datos se empieza a utilizar en instituciones gubernamentales ya que a raíz de la modernización tecnológica implementada en este gobierno, todas o casi todas las secretarías de gobierno cuentan ya con una página web y a la vez se permite realizar más trámites por este medio electrónico. Sin embargo la cantidad de datos crece constantemente y es necesario revisar las tendencias que toma la información, para que las decisiones sean más oportunas.

Implementar un data warehouse consiste en: analizar los requerimientos de información de las áreas usuarias, documentar las reglas de negocio y las fuentes de datos; diseñar las bases de datos que almacenarán la información requerida

[7] INFINITA CONSULTORES. http://www.infinitax.com/notas/notas.htm, consultada el 28 de diciembre de 2004

[8] SARABIA Ramírez Luis G. y BOLAÑOS Usla Miguel R.: *El Data Warehouse de Bancomext* En: Política Digital. NEXOS. México. Número 15, febrero 2004 p. 62

[9] SARABIA Ramírez Luis G. y BOLAÑOS Usla Miguel R. *ibid.*

por los usuarios y construir los procesos de extracción – transformación – carga, necesarios para recuperar los datos fuente y colocarlos dentro del data warehouse, desarrollar las herramientas para analizar, explotar y reportar la información almacenada.

Sin embargo la decisión de implementar un sistema de minería de datos y un data warehouse debe estar justificada siempre en base al volumen de la información que se maneja dentro de la empresa. Ya que de no contar con una cantidad significativa de datos se corre el riesgo de dar resultados imprecisos. Además la implementación de estos sistemas es cara y requiere de tiempo para concluirse y se puede ocasionar que la empresa genere un gasto sin que esté plenamente justificado.

La minería de datos es útil cuando:

- El sistema es parcialmente desconocido y existe una naturaleza aleatoria de los datos.
- Existe una cantidad enorme de datos
- Se cuenta con la infraestructura de hardware y software adecuado, como servidores que manejen grandes volúmenes de información (más de 1,000,000 de registros) y manejadores de bases de datos diseñados para tratar éesta cantidad de datos (como Oracle o SQL Server 2000).

Algunas áreas donde la Minería de datos ha sido exitosa son:

- Detección de fraudes
- Análisis de riesgo en créditos
- Clasificación de cuerpos celestes
- Minería de textos

En el caso de la detección de fraudes financieros, las instituciones bancarias en México están comenzando a utilizar aplicaciones de minado de datos al llevar un registro del comportamiento de las transacciones de sus clientes.

2. Preparando los datos: DATA WAREHOUSING.

2.1.- ¿Qué es un Data Warehouse?

"Un data warehouse es una colección de datos en la cual se encuentra integrada la información de la institución y que se usa como soporte para el proceso de la toma de decisiones gerenciales"[10]

Todo proyecto de minería de datos requiere de una fuente de datos bien estructurada que permita su posterior análisis; básicamente ésta es la sustancia de un data warehouse. Así que describiremos algunas de sus características operacionales:

- El objetivo de un sistema de MD y data warehouse es apoyar a las organizaciones en la toma de decisiones, esto es no se piensa en reemplazar los sistemas operacionales de la organización.

- Un data warehouse es una colección de datos orientados a un tema, integrado, no volátil, de tiempo variante, que se usa para el soporte del proceso de toma de decisiones gerenciales.

- "Un data warehouse es un conjunto de tecnologías encaminadas a apoyar al gestor del conocimiento (ejecutivos, gerentes, analistas) a tomar mas rápidas y mejores decisiones" [11]

En la actualidad las empresas cuentan con una serie de modelos que les permiten administrar su información. Sin embargo esta información se encuentra desagrupada; por ejemplo: el sistema de información del departamento de control

[10] CASARES, Claudio.: *Data Warehousing*, www.programacion.com/bbdd/tutorial/warehouse/1/ consultada 24 de marzo de 2004

[11] JARKE, Matthias et al. : *Fundamental of Data Warehouses* Springer –Verlag, Berlin Heidelberg, Alemania, 2000, p. 1

escolar de un colegio contiene datos sobre el desempeño académico de alumnos y maestros (índice de reprobación, permanencia, aprovechamiento), por otro lado el departamento de recursos humanos posee información relacionada al desempeño laboral de un profesor (faltas, retardos, días económicos), además puede ocurrir que las bases de datos de cada departamento pertenecieran a programas diferentes y basadas en modelos distintos (relacionales, de red, lógicos). En ese sentido la información contenida en cada uno de los sistemas puede ser útil para un director o un subdirector que necesite tomar decisiones.

Los sistemas OLTP (online transaction processing) no brindan el soporte suficiente para la toma de decisiones, ya que están diseñados para dar información inmediata, cuando mucho un año de antigüedad, se actualizan en la mayoría de los casos diariamente o en periodos de tiempo muy cortos, contienen información local o de un solo departamento.

Los sistemas de data warehousing implementan herramientas OLAP (online analytic processing), estos sistemas almacenan información antigua de 5 años o mas, esta información es constante (no sufre cambios) ya que se trata de un archivo histórico de la empresa y además incluye información de todos los departamentos que conforman la organización.

El concepto de MD está ligado al de data warehouse a diferencia de una base de datos de un sistema operacional, que solo contiene información del periodo actual de una empresa, el data warehouse es un conjunto de instantáneas del comportamiento histórico de una organización.

Así el data warehouse es el lugar donde se encuentran los datos que posteriormente serán minados.

2.2.- Características.

Un almacén de datos o Data Warehouse debe cumplir con las siguientes características:

- **Orientado al tema:** Una primera característica es que la información se clasifica en base a los aspectos que son de interés para la empresa. En el ambiente data warehousing se organiza alrededor de sujetos tales como cliente, vendedor, producto y actividad. Figura 3.

Lo anterior se desprende del hecho de que un data warehouse brinda soporte a la toma de decisiones. Así para un gerente, más que saber cuantas ventas se realizaron, sería más interesante saber que tienda las realizó, las características de los compradores y vendedores y atributos de los productos vendidos.

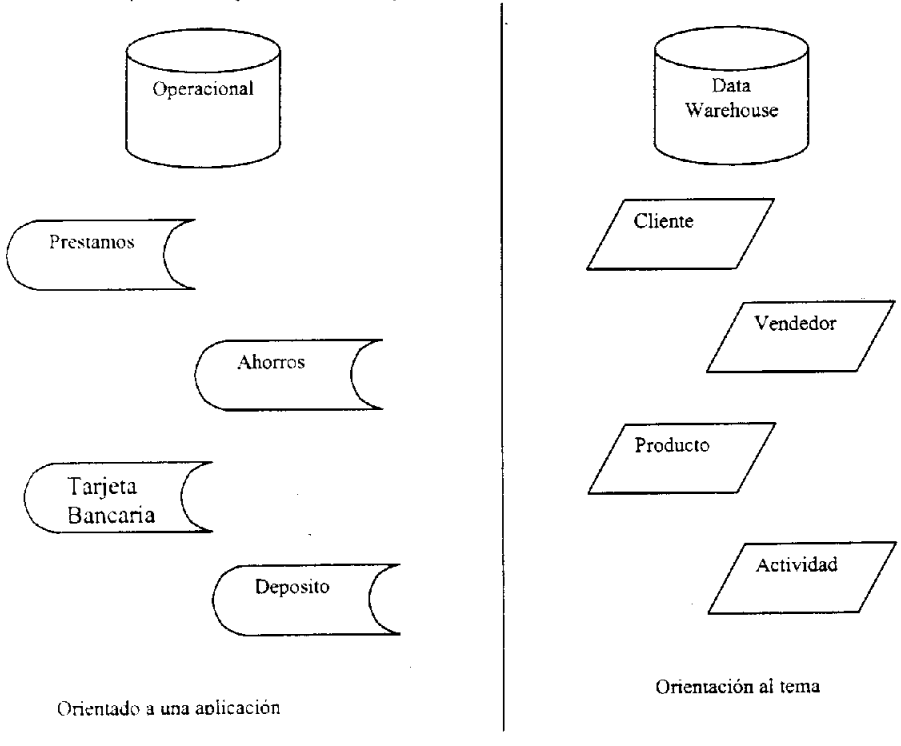


Figura 3. El data warehouse tiene una fuerte orientación al tema.

Integrado: "El aspecto más importante del ambiente data warehousing es que la información encontrada al interior esté siempre integrada."[12]

El contenido de un data warehouse proviene de bases de datos heterogéneas donde la referencia a un atributo puede tener varias representaciones, por ejemplo la estatura de un individuo podría estar dada en centímetros en una base mientras en otra talvez esté en pulgadas.

"Cualquiera que sea la forma del diseño, el resultado es el mismo – la información necesita ser almacenada en el data warehouse en un modelo globalmente aceptable y singular, aún cuando los sistemas operacionales subyacentes almacenen los datos de manera diferente."[13]

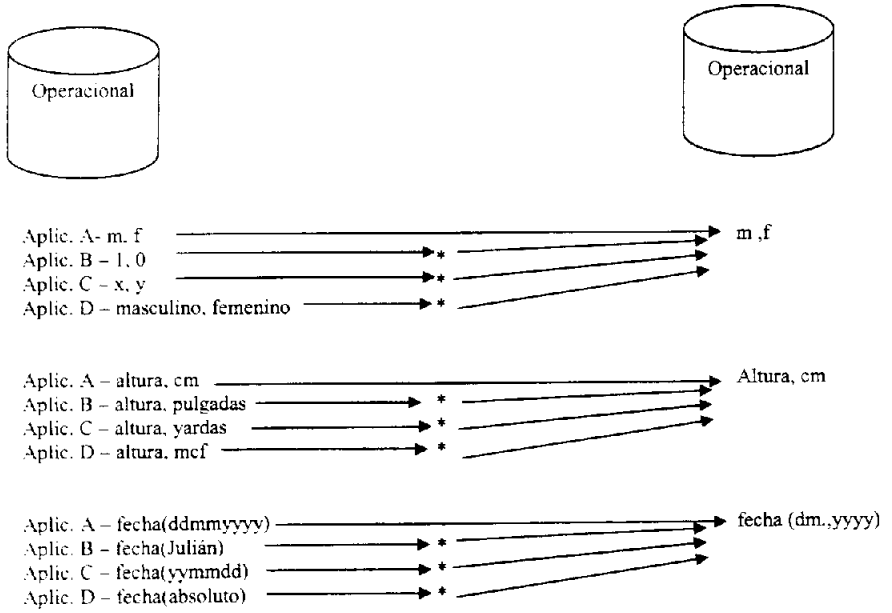


Figura 4. Cuando los datos se mueven al data warehouse desde las aplicaciones orientadas al ambiente operacional, los datos se integran antes de entrar al deposito.

[12] CASARES, Claudio.: *Data Warehousing*, www.programacion.com/bbdd/tutorial/warehouse/4/ consultada 24 de marzo de 2004
[13] CASARES, Claudio.: *Data Warehousing*, *ibid*

- **De tiempo variante:** Toda la información del data warehouse es requerida en algún momento. Como la información en el data warehouse es solicitada en cualquier momento (es decir, no "ahora mismo"), los datos encontrados en el deposito se llaman de tiempo variante.

  Los datos históricos son de poco uso en el procesamiento operacional. La información del depósito debe incluir los datos históricos para usarse en la identificación y evaluación de tendencias.

  De esta forma un data warehouse incluye datos históricos y datos actuales. Podemos imaginar el data warehouse como un álbum de fotografías del comportamiento de la empresa a lo largo de varios años.

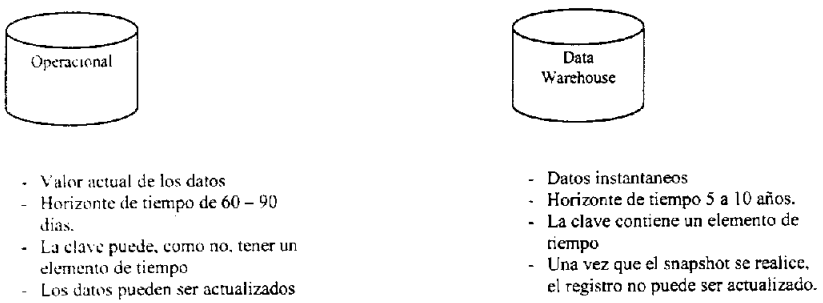


Figura 5. Las bases de datos operacionales guardan datos actuales, en cambio el data warehouse contiene un histórico del comportamiento de la empresa.

- **No volátil:** "La información es útil solo cuando es estable. La perspectiva más grande, esencial para el análisis y la toma de decisiones, requiere una base de datos estable."[14]

[14] CASARES, Claudio. : *ibid*

Toda vez que el data warehouse contiene un registro histórico de la empresa damos por hecho que los datos han sido registrados de manera correcta, por tal motivo podemos con seguridad basar en ellos nuestras decisiones.

Desde luego que puede haber correcciones en los datos almacenados, pero éstas serán mínimas.

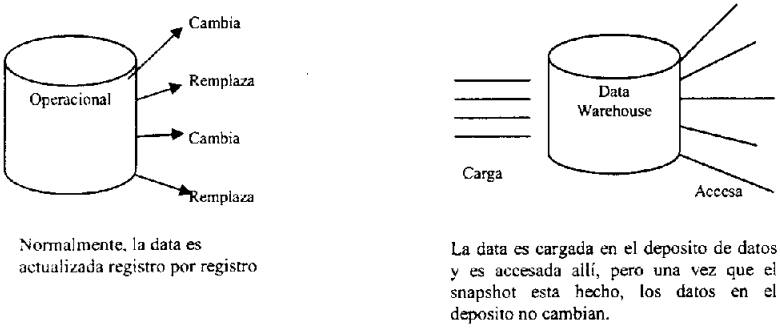


Figura 6. Una vez cargados los datos en el data warehouse no pueden ser modificados.

2.3. Componentes

La arquitectura de un data warehouse se compone de varias capas, cada una conformada de datos de una capa anterior, las partes que componen un almacén de datos se muestran en la Figura 7.

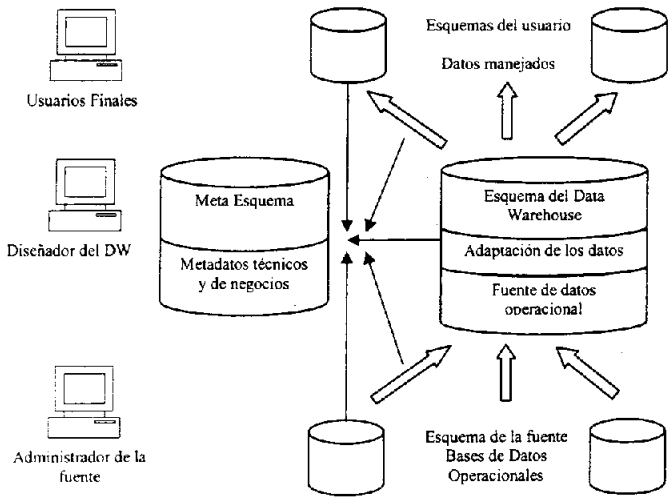


Figura 7. Arquitectura genérica de un Data Warehouse,

Como se pudo observar el almacén de datos se compone de tres capas Fuente de Datos, la más baja; el Data Warehouse Global, es la capa intermedia y final mente la capa de datos que se entrega a los usuarios finales conocida como Local Data Warehouse, considerando mas detalladamente los siguientes puntos:

Origen de datos o bases de datos operacionales: Es el nivel más bajo. "Consisten de datos organizados en sistemas de bases de datos abiertos y sistemas heredados, o de datos no estructurados o semi estructurados almacenados en archivos... son usualmente heterogeneos"[15]. Se refiere a los sistemas que ya existen en las organizaciones y del se alimenta el data warehouse, estos sistemas pueden contener archivos de diversa naturaleza, por ejemplo de ACCESS y de DBASE.

Los datos contenidos en este nivel deben ser limpiados e integrados para que sean parte del nivel superior.

[15] JARKE, Matthias et al. :, *obr. cit.*, p. 2

Data Warehouse Primario o Global: "Es una colección de bases de datos integradas, no volátiles, orientados a sujetos, diseñados para el soporte de la toma de decisiones, donde cada unidad de información será relevante en algún momento de tiempo"[16]

En este nivel los datos son limpiados y colocados en un formato estándar. Es la parte central del bodegón de datos y a la que más trabajo cuesta llegar. Podemos ver a este componente como un conjunto de fotografías que describen el comportamiento de una empresa. Mantiene un registro histórico de los datos.

Local Warehouses: "Contienen conjuntos de datos derivados del warehouse global, seleccionados para soportar actividades como procesamiento de información, toma de decisiones, análisis histórico, análisis de tendencias o análisis de integración".[17]

Los data marts son un tipo especial de warehouse local, un data mart es un pequeño almacén de datos, el cual contiene un subconjunto de datos del data warehouse global. Puede ser usado en un solo departamento y contiene solo información relevante sobre el área que lo utiliza.

2.4 Estructura física de un DW.

Tenemos tres posibles arquitecturas para diseñar un data warehouse:

- Centralizada
- Descentralizada
- Departamental

[16] JARKE, Matthias et al. : *ibid*
[17] JARKE, Matthias et al. : *ibid*

En la arquitectura centralizada (figura 8) existe un solo data warehouse que almacena los datos necesarios para el análisis del negocio. La desventaja de esta arquitectura es la pérdida de performance ya que todas las consultas y actualizaciones deben ser procesadas en un solo sistema.

Por otro lado, el acceso a los datos es fácil y sin complicaciones porque solo un modelo de datos es relevante. La construcción y el mantenimiento es más fácil que en un entorno distribuido. Ésta arquitectura es útil en compañías donde la administración de información también es centralizada.

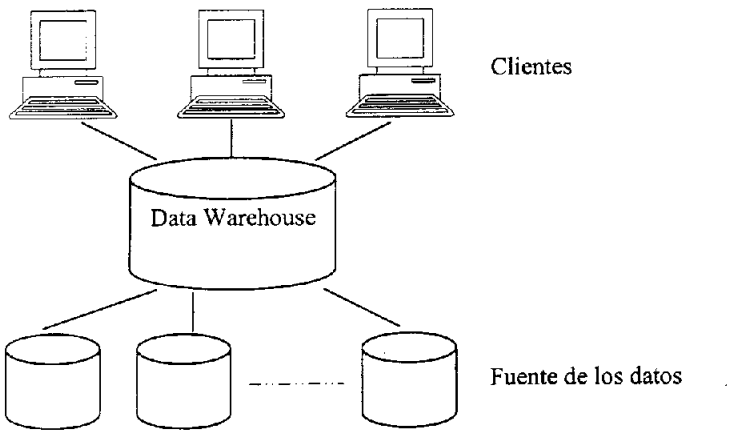


Figura 8. Arquitectura de un data warehouse centralizado.

En el caso de la arquitectura descentralizada, figura 9., el data warehouse global solo es lógico a diferencia del warehouse departamental donde es físico. Las ventajas de estas arquitecturas, son respuesta más rápida con respecto al tiempo, al momento de usar muchas máquinas se reducen los costos en equipo voluminoso ya que el espacio de almacenamiento está repartido en varios lugares.

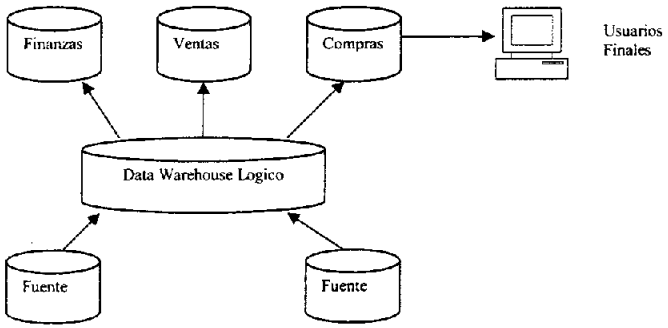


Figura 9. Arquitectura de un Data Warehouse

Un data warehouse no es estático evoluciona y crece con el tiempo. Por esto la arquitectura elegida para construir el data warehouse debe ser fácil de expandir y reestructurar.

2.5. Modelo de datos.

Debemos tomar en consideración que la función principal de un data warehouse es apoyar a los administradores en la toma de decisiones. Los sistemas basados en el modelo relacional utilizan tablas o relaciones normalizadas (al menos hasta la tercera forma normal), que evitan redundancia de datos y optimizan las consultas. Sin embargo lo que es bueno para los sistemas de transacciones o OLTP, no lo es para los sistemas de análisis o OLAP.

Los sistemas OLAP se nutren de datos adquiridos de los sistemas OLTP por lo tanto las tablas fuente están normalizadas. Para que los datos sean funcionales en OLAP las bases de datos sufren un proceso de transformación en cubos (modelo Multidiensional) o en esquemas estrella o de copo de nieve (modelo súper - relacional).

2.5.1. Cubos de información.

"Los cubos son elementos claves en OLAP, una tecnología que provee rápido acceso a datos en un almacén de datos (data warehouse)"

"Los cubos son subconjuntos de datos de un almacén de datos, organizados, sumarizados dentro de una estructura multidimensional."[18]

2.5.2. Multidimensión.

Las consultas ejecutadas en bases de datos relacionales nos dan un vistaso inmediato de las transacciones que realiza una empresa como por ejemplo cuantos artículos X se vendieron en el día o la venta total de una tienda en el mes.

El análisis en línea involucra consultas más complejas tales como: A lo largo de 5 años ¿cuál tienda ha vendido más en determinada región y su producto más vendido? o ¿cuál es el promedio de venta de cada uno de los trabajadores a través de los últimos 2 años?

"... existen muchos cruces de información que en muchas ocasiones no nos damos cuenta, generalmente vemos todo en dos dimensiones, pero estos cruces a los que se les llama la múltiple dimensión, nos dan información muy rica, para poder visualizar situaciones de negocios, muchas veces no imaginadas"[19].

Supongamos el caso de alguna escuela que cuente con la siguiente información base:

1.- 120 profesores
2.- 2500 alumnos

[18] http: www.prado.com.mx Cubos cubos-definiciones-y-conceptos-4.htm
Consultada el 20 de marzo de 2004
[19] http: www.prado.com.mx Cubos Cubos_Index.htm , consultada el 20 de marzo de 2004

3.- 60 materias

4.- Mide las siguientes variables: aprobación, permanencia, aprovechamiento, egreso.

5.- Cuenta con 10 semestres de información.

El número de combinaciones que la información puede tener es:

120 x 2500 x 60 x 4 x 10 = 720,000,000

¡Setecientos veinte millones de combinaciones!

"Generalmente no conocemos ni la centésima parte de los que ocurre en nuestra empresa, con dos o tres reportes en dos dimensiones, estamos ciegos..."[20] .

"Conceptualmente una base de datos multidimensional usa la idea de un cubo para representar las dimensiones disponibles para un usuario. Adicionalmente dentro de una dimensión puede haber jerarquías y niveles."[21]

Una jerarquía ordena por categorías y niveles de detalle a los datos que tienen un mismo significado para la empresa.

Un ejemplo sería:

Empleado: Maestros: Materia que imparten:

En el modelo multidimensional visualizamos a la base de datos como un espacio de 3 o más dimensiones, cada dimensión representada por una tabla. Figura 10.

---

[20] *ibid*

[21] JIMENEZ, Claudia.: *Bases de Datos Multidimensionales*, www.inf.udec.cl/~basedato/trabajos/multidimensionales.pdf, Departamento de Ingeniería Informática y Ciencias de la Computación Universidad de Concepción, Agosto 2002 p. 4, Consultada el 27 de marzo de 2004

La necesidad de construir una base de datos multidimensional parte de construir una estructura de datos donde se puedan aplicar herramientas OLAP.

OLAP o On line Analytical Processing puede ser definido como "el proceso interactivo de crear, mantener, analizar, y elaborar informes sobre los datos", a demás se añade que los en si son percibidos y manejados como si estuvieran almacenados en un "arreglo multidimensional"

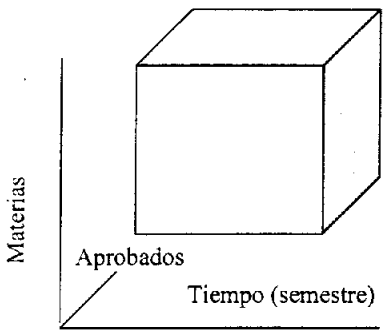


Figura 10. Ejemplo de un cubo de información donde las dimensiones son el tiempo, los alumnos aprobados y las materias.

"Los sistemas OLAP deben:

- Soportar requerimientos complejos de análisis.
- Analizar datos desde diferentes perspectivas.
- Soportar análisis complejos contra un volumen ingente de datos".[22]

[22] www.csi.map.es/csi/silice/DW2251.html

Una herramienta OLAP permite la navegación multidimensional, entre otras las siguientes acciones:

Rotar (swap): Alterar las filas por columnas.
Bajar (Down): Bajar el nivel de visualización.
Detallar (Drilldown): informar para una fila en concreto, de datos a un nivel inferior.
Expandir (expand) : Igual que el punto anterior pero sin perder la información a un nivel superior para éste y el resto de los valores.
Colapsar (Collapse): Operación inversa a la anterior.

Existen dos enfoques para la construcción de un sistema OLAP.

OLAP Multidimensional o MOLAP y
OLAP Relacional o ROLAP

2.5.3 MOLAP

Las características de un sistema con un enfoque basado en MOLAP son:

Los datos se almacenan físicamente en una estructura matricial especial y deben ser obtenidos de tablas sin normalizar.

Está basado en el paradigma propuesto por Ralph Kimball que propone que el data warehouse es un conglomerado de datos dentro de una empresa y debe ser almacenada siguiendo el modelo multidimensional.

Fue la primera propuesta cuando se vislumbro la idea un data warehouse y un sistema OLAP; sin embargo presenta problemas con una cantidad muy grande de datos y el hecho de tener que desnormalizar o transformar tablas para alimentar al sistema es un proceso laborioso.

Escapa del objetivo de esta tesis mostrar con más detalle este punto, solamente diremos que los sistemas MOLAP van en retirada del mercado.

2.5.4 ROLAP

Un sistema ROLAP soporta bases de datos relacionales esto es que están al menos en la tercera forma normal (3FN).

Consta de 3 niveles:

- Nivel de base de datos relacional.
- Nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios.
- El motor ROLAP se integra con niveles de presentación, a través de los cuales los usuarios realizan el análisis OLAP.

No hay una idea correcta o incorrecta de implementación ya que los paradigmas ROLAP y MOLAP representan diferentes filosofías para implementar un data warehouse.

2.6. Diseño de un data warehose

Comenzaremos con una introducción al modelo relacional ya que la mayoría de los sistemas de donde se alimentara nuestro data warehouse están basados en él.

Para que se considere que una base de datos cumple con el modelo relacional propuesto por Codd, las tablas o entidades de la base deben estar en la 3 forma normal; esto es:

"Una relación está en 3FN si y solo si los atributos no clave (si los hay) son:

a) Mutuamente independientes, y
b) Dependientes de la clave primaria. "[23]

"Una relación en tercera forma normal (3FN) si y sólo, en todo momento, cada tupla está formada por un valor de clave primaria que identifica a alguna identidad, junto con un conjunto de cero o más valores de atributos independientes entre sí, los cuales describen de alguna manera esa entidad."[24]

Ahora bien el proceso de normalización genera un número muy grande de tablas lo cual es la antitesis de los que se necesita para un almacén de datos; las tablas que conforman un bodegón de datos son en algunos casos consultas de dos o más tablas por lo que para lograr resultados de ejecución satisfactorios, las relaciones deben ser tratadas con un proceso de desnormalización.

Para diseñar tablas fáciles de navegar y consultar se utiliza una técnica conocida como modelo dimensional o esquema estrella.

2.6.1. El Modelo Dimensional o Esquema Estrella.

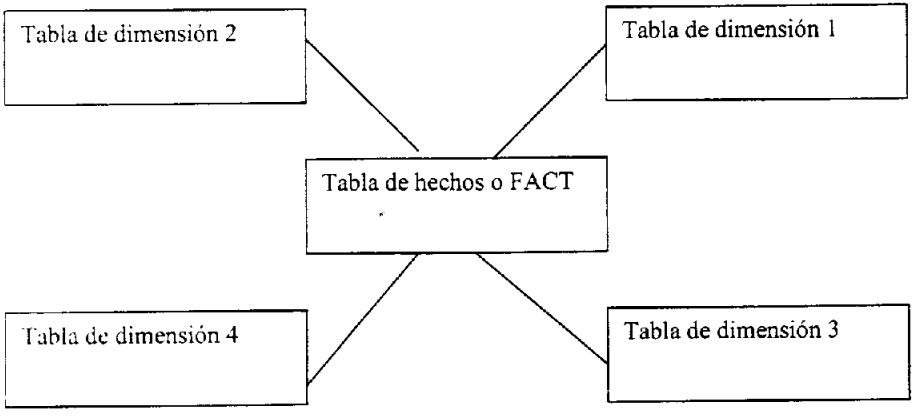


Figura 11. Esquema del modelo multidimensional de estrella o star schem

[23] DATE, C J.: *Introduccion a los sistemas de bases de datos.* Ed. Addison – Wesley Iberoamericana, V. 1, 5ª E, 1993, p. 522

[24] DATE, C J: *ibid*, p. 523

El esquema estrella también conocido como esquema multidimensional, cubo de datos y star schema, consiste de una tabla central o tabla de hechos rodeada de varias tablas de dimensión.

"Las relaciones entre las tablas de hechos y las tablas de dimensión deben ser simples y claras, tal que hay una sola posibilidad de unir dos tablas cualquiera y que el significado de la unión es obvio y bien entendido."[25]

El esquema estrella está pensado para proveer a los trabajadores del conocimiento[26] de herramientas que les permitan y faciliten consultar la información contenida en el almacén de datos.

Para realizar el esquema estrella de una base de datos en particular lo que se debe hacer es identificar los hechos y sus dimensiones. Los hechos son las actividades claves en una empresa e impactan a toda la empresa o un sector. Las dimensiones son los elementos que pueden ejercer alguna influencia sobre la tendencia que muestran los hechos.

La dificultad radica en que a veces las relaciones entre los hechos y las dimensiones no son tan claras, así que es deber del gestor del conocimiento hacer un análisis muy detallado de cada una de las necesidades de los administradores y gerentes. De igual forma se deben revisar todos los reportes generados por los sistemas tipo OLTP, para de esta forma tener una idea más clara de cuales son las necesidades de información.

2.6.2. Ejemplo de Modelo Estrella.

Como se pude observar en la figura 12 muestra un ejemplo de esquema estrella, la tabla central de Ventas al centro es la tabla de hechos y contiene

[25] ENSOR, Dave and STEVENSON, Ian.: *Oracle Design*. O'REILLY. EUA. 1997. p. 348

[26] Directores, gerentes, administradores y todo aquel que requiera información para la toma de decisiones.

claves externas Código_Área, Código Tiempo, Código_Cuenta y Código de Producto, que corresponden a cada dimensión del hipercubo. La estructura jerárquica de cada dimensión y de la estrella se utiliza para realizar operaciones de Drilldown (recorrer la estructura jerárquica, yendo de lo general a lo particular y obteniendo una vista más a detalle de los datos) y de Roll – Up (Intercambiar los ejes del hipercubo).

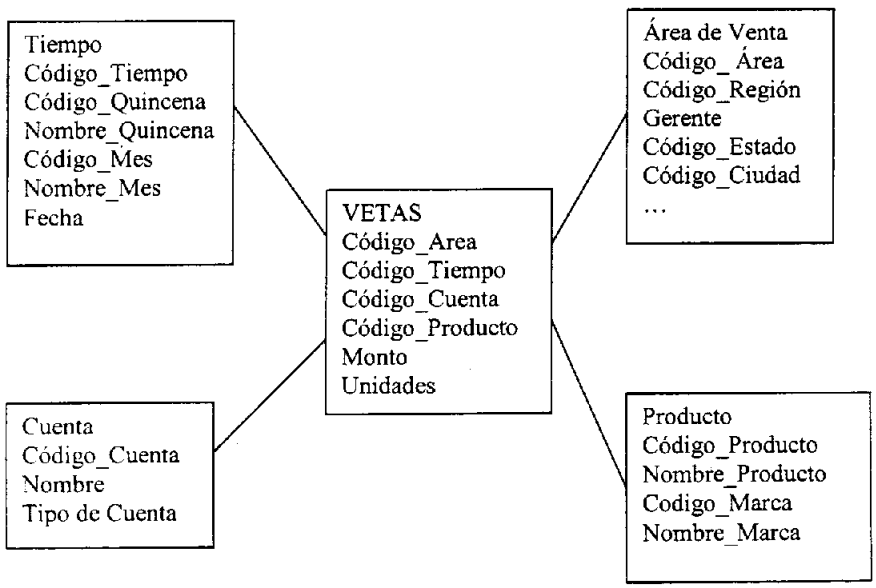


Figura 12. Esquema estrella para el departamento de ventas.

Las tablas de dimensión son estructuras no normalizadas. El modelo Copo de Nieve o Snowkflake schema, es una extensión del esquema estrella y consiste en normalizar cada tabla de dimensión del esquema estrella figura 13.

El tiempo tiene características y propiedades interesantes dentro de un cubo de datos. Es poco común que se consulte al data warehouse para solicitar el detalle de cada venta diariamente, por el contrario se piden datos sobre las ventas de un producto en un mes para conocer tendencias de mercado; de esta forma "la

dimensión tiempo es a menudo la base sobre la cual la tabla de hechos se particiona"[27].

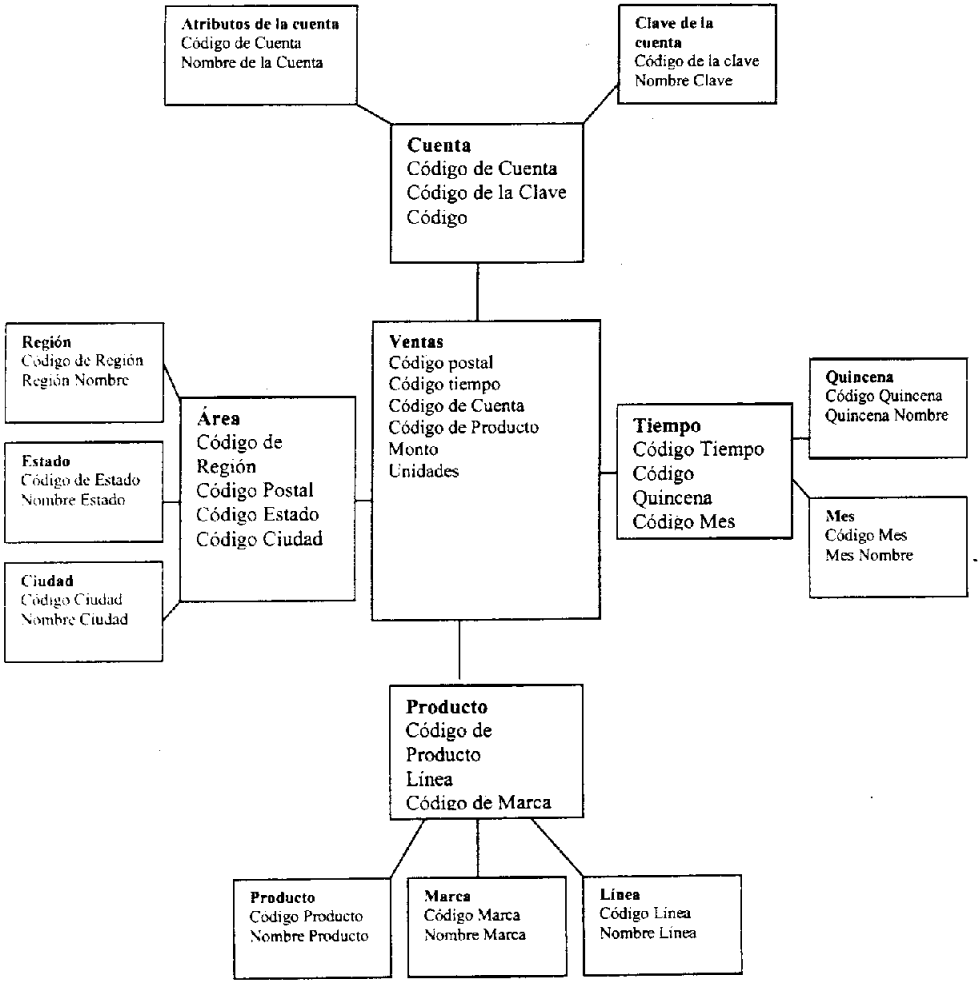


Figura 13. Esquema de copo de nieve para el departamento de ventas.

[27] ENSOR, Dave and STEVENSON, Ian:, Obr. cit., p. 350

La elección de la medida de tiempo (día, semana, mes, etc.), que se va a utilizar en el bodegón de datos, es muy importante ya que determina el detalle con que se mostrara la información. De esta manera si elegimos periodos de un mes nuestro sistema OLAP podrá entregar informes por trimestre, año, semestre, sin embargo no se puede llegar a un nivel más bajo como el de semana.

El diseño del almacén de datos de cualquier proyecto de minería es una tarea compleja que requiere de mucho tiempo y sobre todo de un conocimiento exacto de las necesidades de información de la empresa u organización para la cual se va a elaborar.

Un data warehouse mal planteado podría llevarnos a generar una aplicación que no responda a los requerimientos de los gerentes y podría convertirse en un lastre para las organizaciones.

3. Extracción y análisis de datos

Un data warehouse almacena física y lógicamente un conjunto de datos que describen el comportamiento de una empresa u organización a través del tiempo.

Una vez que se ha construido este componente se procede a explotar los datos que contiene para encontrar modelos que después de interpretados nos generen un conocimiento.

Mencionábamos que dentro de los principales objetivos de un data warehouse se encontraba el de proveer de información suficiente a los administradores de tal forma que se permita evaluar resultados, para una toma de decisiones adecuada y oportuna. La estadística y la probabilidad son herramientas formales que nos permite analizar y comparar datos con el fin entregar información útil y significativa a los gerentes y ejecutivos de cualquier empresa.

Además el pensamiento estadístico ha influido enormemente en el desarrollo del área de aprendizaje automático o machine learning.

El motivo para manejar conceptos de estadística y probabilidad en el área de minería de datos se basa en dos hechos, el primero es que como trabajadores del conocimiento necesitamos de los elementos teóricos para interpretar correctamente los datos y predecir tendencias, el segundo es el hecho de que los algoritmos de aprendizaje (clustering y decisión trees) basan su funcionamiento teórico en éstas dos disciplinas.

Los algoritmos que generan árboles de decisiones utilizan conceptos de probabilidad principalmente el Teorema de Bayes y el concepto de entropía. En el caso de los algoritmos de clasificación o clustering se necesitan los conceptos de desviación estándar, media, moda, mediana.

Dentro de los conceptos de estadística que se utilizan en esta disciplina están los de variable aleatoria, distribución de probabilidad, distribución estándar y varianza. Conceptos que nos ayudaran a entender varias tareas de la minería de datos tales como *predicción, clasificación, estimación y muestreo.*

Solo dos técnicas de clasificación, árboles de decisión y clustering, se utilizan en la herramienta comercial llamada **SQL Server** de Microsoft.

3.1. Razonamiento estadístico y probabilístico.

Para entender adecuadamente los conceptos mostrados en los algoritmos siguientes, es necesario dar un repaso de los conceptos básicos de la estadística y la probabilidad.

3.1.1. Probabilidad.

"La probabilidad es el estudio de experimentos aleatorios o no deterministicos."[28] Al evento de lanzar un dado y de que al caer salga un dos, se le llama la probabilidad de que el evento tal sea exitoso. La probabilidad de un evento está dada por la relación:

$$P(A) = \frac{A}{S}$$

que se lee como la probabilidad de que se cumpla el evento A es igual al número de maneras como puede ocurrir el evento A sobre el número de maneras como puede ocurrir el espacio muestral S.

[28] LIPSCHUTZ, Seymour. Matemáticas para computación, Ed. Mc. Graw – Hill,1992, p. 276.

La probabilidad de que salga un dos al tirar un dado sería igual a:

$$P(2)=1/6$$

donde 1 es el número de lados que tienen un dos y 6 es el número de caras del dado.

Un espacio finito de probabilidad consiste de un conjunto finito S, junto con una función de valor real P( ) , que satisface las siguientes propiedades:

1.- Para cada evento A, P(A)>=0

2.-P(S) = 1

3.- Si los eventos A y B son mutuamente excluyentes,
entonces P(A U B)=P(A)+P(B)

3.1.2. Probabilidad condicional y Teorema de Bayes.

La probabilidad de que ocurra un evento A una vez que haya ocurrido un evento E, se conoce como probabilidad condicional y se define con la siguiente relación:

$$P(A|E) = \frac{P(A \cap E)}{P(E)}$$

El reverendo Thomas Bayes (1702 – 1761) desarrollo una fórmula que simplifica el cálculo de las probabilidades condicionales.

La formula de Bayes permite calcular la probabilidad de que ocurra el evento B, si se sabe que ya ocurrió el evento A, esto es P(B|A). Para ello se requiere conocer la probabilidad simple de que ocurra el evento A, la probabilidad simple de que ocurra el evento B y la probabilidad de que ocurra el evento A si se sabe que ya ocurrió el evento B.

La regla de Bayes es la siguiente:

$$P(B/A) = \frac{P(A/B) * P(B)}{P(A)}$$

Interpretemos la regla de Bayes con un ejemplo:

Si 55.26 % de los automóviles de un estacionamiento son de cuatro puertas (P(A)), se sabe que el 21.27 % de todos los automóviles son blancos (P(B)) y también se sabe que de los autos de cuatro puertas el 59.77 son blancos (P(A/B)). ¿Cuál es la probabilidad de que el próximo auto que salga del estacionamiento sea blanco y de cuatro puertas (P(B/A))?

A = Porcentaje de autos de cuatro puertas
B = Porcentaje de autos blancos
A | B = Porcentaje de autos de cuatro puertas que son blancos

P(B | A) = ¿?Porcentaje de autos blancos que son de cuatro puertas.

P(B | A) = (0.5977*0.2127)/0.5526 = 0.2301

El teorema de Bayes es fundamento de mochos algoritmos de clasificación y parte importante de la Teoría de Decisiones.

3.1.3. Estadística

Para Seymour la estadística es una ciencia, la rama de la matemática que organiza, analiza e interpreta datos aún no procesados.

La estadística se aplica en muchas áreas para ayudar a la toma de decisiones.

3.1.4. Media y desviación estándar.

El **promedio aritmético** de una lista de valores o **media** se define como la suma de los valores divididos por el número de valores. Su relación es la siguiente:

$$\bar{x} = \frac{\sum x_i}{n}$$

A la diferencia que existe entre el un elemento de un conjunto de datos y su media se le llama **desviación.** Al promedio de los cuadrados de las desviaciones se le llama la **varianza** de los datos, y a la raíz cuadrada de la varianza se le llama la **desviación estándar.**

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - media)^2}{n} \quad ; \text{Varianza}$$

$$\sigma = \sqrt{\sigma^2} \quad ; \text{Desviación estándar}$$

La media es una medida de tendencia central, la varianza y la desviación estándar son medidas de dispersión. Con estas medidas es posible comparar entre sí varios grupos de datos.

Su aplicación al campo del aprendizaje automático y el reconocimiento de patrones se verá mas adelante al estudiar los diferentes algoritmos de clasificación o clustering.

El objetivo de este apartado es el de recordar algunos conceptos útiles en la minería de datos, el lector interesado en profundizar en temas de probabilidad y estadística deberá consultar algún texto más especializado.

3.2. Aprendizaje automático

3.2.1. Introducción al Aprendizaje Automático

La manera en que los desarrolladores determinan la inteligencia de un sistema se basa en la buena o mala toma de decisiones que hacen comparadas con las que podría hacer un experto.

El conocimiento se define como la cantidad de información que se adquiere en base a la experiencia, estudio y observación del medio. Sin embargo estas formas de adquirir la información son demasiado lentas y funcionan a lo largo del desarrollo humano. Para el desarrollo de sistemas inteligentes los especialistas han propuesto una serie de técnicas para simular el aprendizaje de manera más rápida. A las técnicas que simulan el aprendizaje en las máquinas se le llama **Aprendizaje Automático**.

Existen diferentes y muy variadas estructuras computacionales que hacen que las máquinas aprendan, entre otras podemos mencionar a las redes neuronales, los árboles de decisión, las técnicas de clustering, lógica difusa y la combinación de éstas.

La historia del Aprendizaje Automático se puede dividir en tres periodos el primero de principios de los 50's a mediados de los 60's donde el objetivo principal era conseguir sistemas de aprendizaje general, utilizando una variedad de sistemas neuronales. El segundo se desarrollo entre los años 60´s y finales de los 70´s y se buscaba simular el aprendizaje humano. El tercer periodo que es el más reciente donde se trata el aprendizaje de conceptos a través de ejemplos.

3.2.2. Árboles de decisión

Un árbol de decisión es una estructura que representa los procesos involucrados en las tareas de clasificación. El resultado puede entonces representarse como un conjunto de reglas Si – Entonces.

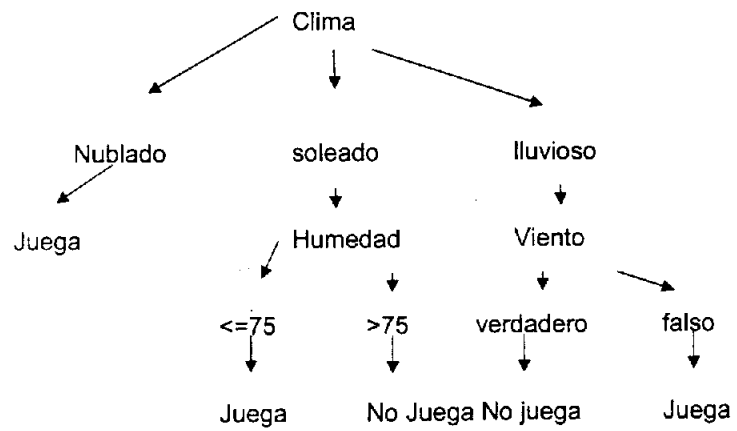
"El aprendizaje en árboles de decisión es uno de los más fáciles de implementar pero a su vez de los más poderosos.

Un árbol de decisión toma de entrada un objeto o situación descrita por un conjunto de atributos y regresa una decisión "falso o verdadero".

Cada nodo interno corresponde a una prueba en el valor de uno de los atributos y las ramas están etiquetadas con los posibles valores de la prueba.

Cada hoja especifica el valor de la clase."[29]

A continuación un ejemplo de árbol de decisión:



[29] F. Morales Eduardo, http://dns1.mor.itesm.mx/~emorales/Cursos/KDD01/principal.html, ITESM, consultada el 20 de marzo de 2004

3.2.2.1. Algoritmo ID3

La función de este algoritmo es construir árboles de decisión, partiendo de una colección de datos de ejemplo o conjunto de entrenamiento.

Datos ejemplo----------→ ID3----------→ Árbol de decisión

El ID3 es un algoritmo de aprendizaje de árboles de decisión propuesto por Roos Quinlan en 1979.

Mostremos como funciona el algoritmo ID3 a través de un ejemplo[30]:

Si tenemos dos clases conocidas C1 y C2 organizados de la siguiente forma:

CLASE	ELEMENTO	ALTURA	CABELLO	OJOS
C1	1	BAJO	RUBIO	AZULES
	2	ALTO	PELIRROJO	AZULES
	3	ALTO	RUBIO	AZULES
C2	4	ALTO	RUBIO	MARRONES
	5	BAJO	CASTAÑO	AZULES
	6	ALTO	CASTAÑO	AZULES
	7	ALTO	CASTAÑO	MARRONES
	8	BAJO	CASTAÑO	MARRONES

De esta manera los individuos 1,2 y 3 pertenecen a la clase 1 y los individuos 4, 5, 6, 7 y 8 pertenecen a la clase 2. No debemos perder de vista que estos son los datos de entrenamiento sobre el cual se construirá nuestro árbol de decisión.

3 basado en el ejemplo mostrado en [ref1]

El primer paso a realizar es identificar el atributo que mejor clasifica a todos los elementos del conjunto de ejemplo. En este ejemplo se tienen como atributos a la ALTURA, CABELLO y OJOS.

El mejor atributo será aquel que ofrezca mayor ganancia de información. Para entender la ganancia revisamos primero el concepto de ***Entropía.***

"De manera general, definimos entropía como la medida de la incertidumbre que hay en un sistema. Es decir, ante una determinada situación, la probabilidad de que ocurra cada uno de los posibles resultados"[31]

La expresión de la entropía más usada es la denominada binaria. Su expresión es:

$$H2(P,^{1-}P) = P\log_2(P)+^{1-}P\log_2(^{1-}P)$$

donde *P* es la probabilidad de que tal evento ocurra.[32]

Ahora bien la ganancia es la diferencia entre la entropía de un nodo y la de uno de sus descendientes. Es una heurística que nos servirá para la elección del mejor atributo clasificador en cada nodo. Formalmente definimos la ganancia como:

Ganancia(S,A)=Entropía(S) - ∑v |Sv| / |S| Entropía (Sv)

Regresando a nuestro ejemplo vamos a seleccionar el mejor atributo clasificador.

Tenemos

S={{1,2,3},{4,5,6,7,8}}

[31] [ref1]
[32] se utiliza la entropia binaria ya que el algoritmo original de ID3 solo clasificaba dos clases diferentes más adelante se verá una ampliación del modelo.

La entropía del conjunto de ejemplo es:

$$\text{Entropia(S)} = -\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right) = 0.954$$

Calculamos la entropía para el atributo CABELLO, primero agrupamos en los siguientes conjuntos:

RUBIO={1,3,4,8}, CASTAÑO={5,6,7}, PELIRROJO={2}

$$\text{Entropía(Rubio)} = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$$; *2 elementos pertenecen a cada clase*

$$\text{Entropía(Castaño)} = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) = 0$$; *los 3 elementos pertenecen a la clase 2*

$$\text{Entropía(Pelirrojo)} = -\frac{1}{1}\log_2\left(\frac{1}{1}\right) = 0$$; *el único elemento pertenece a la clase 1*

La ganancia del atributo CABELLO queda como sigue:

$$\text{G(S, CABELLO)} = E(S) - [\frac{4}{8}E(rubio) + \frac{3}{8}E(castaño) + \frac{1}{8}E(pelirrojo)]$$

$$\text{G(S, CABELLO)} = 0.95443 - \left[\frac{4}{8}1 + \frac{3}{8}0 + \frac{1}{8}0\right] = 0.45443$$

Aplicando el mismo procedimiento a los dos atributos restantes (OJOS, ALTURA) obtenemos:

G(S, OJOS)=0.348

G(S, ALTURA)=0.003

Como podemos observar (figura 14) la ganancia del atributo CABELLO es mayor que la de las otras dos, por lo tanto lo tomamos como atributo raíz de nuestro árbol de decisión.

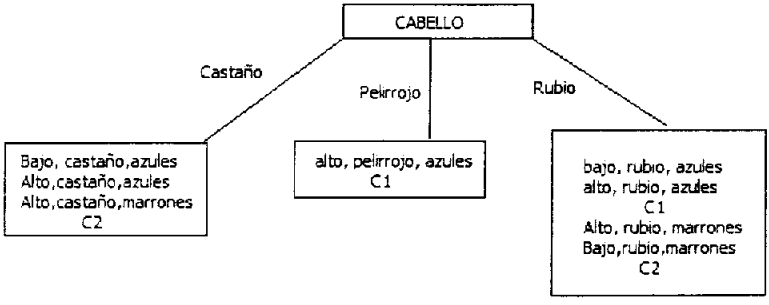


Figura 14. Árbol de decisión generado en el algoritmo ID3

Los elementos dentro de las hojas que salen de las ramas etiquetadas como castaño y pelirrojo pertenecen a una sola clase por lo que el algoritmo se detiene, sin embargo los elementos de la hoja de la rama etiquetada como rubio pertenecen a las dos clases por lo que seleccionamos otro atributo (el atributo OJOS es el de mayor ganancia después de CABELLO) para clasificar, originando un subárbol, figura 15.

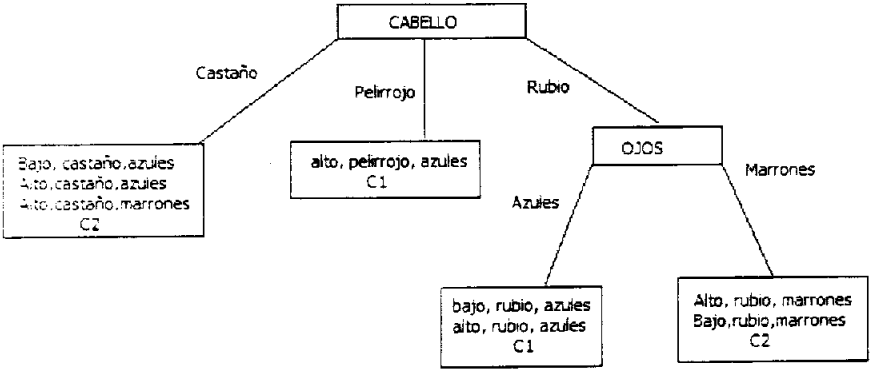


Figura 15. Subárbol generado por el algoritmo ID3.

El árbol resultante clasifica todas las tuplas del conjunto de entrenamiento. Las reglas generadas para el árbol de decisión serían:

Si CABELLO = Castaño Entonces
I pertenece a la clase C2
Sino Si CABELLO = Pelirrojo Entonces
I pertenece a la clase C1
Sino Si CABELLO = Rubio Entonces
    Si OJOS =Azules Entonces
    I pertenece a la clase C1
    Sino Si OJOS = Marrones Entonces
    I pertenece a la clase C2
Fin

Los árboles de decisión se ubican dentro de los modelos de aprendizaje supervisado, ya que los datos del conjunto de ejemplo son como un maestro para el algoritmo.

3.2.2.2. Algoritmo C4.5

El algoritmo C4.5 propuesto también por Quinlan es una extensión del algoritmo ID3. Recibe como entrada un conjunto de ejemplos de entrenamiento a los cuales se les asigna una clase. El objetivo del algoritmo es encontrar una función que permita clasificar un nuevo ejemplo en alguna de las clases marcadas. La manera en como se expresa esta función es lo que diferencia a cada una de las estructuras de clasificación. En este caso el algoritmo C4.5 devuelve una estructura de árbol de decisión.

El aprendizaje en los árboles de decisión se basan en la técnica de divide y vencerás, seleccionando los atributos que mejor clasifiquen un conjunto de

ejemplo. Se utilizan dos criterios para la clasificación de atributos el criterio de ganancia de información o information gain y el criterio de proporción de ganancia.

" Si RF(Cj,S) denota la frecuencia relativa al número de casos en S que caen dentro de la clase Cj. La cantidad de información de un atributo que identifica la clase de una caso de S es entonces:

$$I(S) = -\sum_{j=1}^{x} RF(Cj,S)\log(RF(Cj,S)).$$

Después de que S es dividida en subconjuntos S1,S2,S3...St por un atributo B, la ganancia de información es entonces:

$$G(S,B) = I(S) - \sum_{i=1}^{t} \frac{Si}{S} I(Si).$$

El criterio de ganancia elige el atributo B que maximiza la función G(S,B).

Un problema con este criterio es que favorece a los atributos con un número muy grande de valores iguales."[33]

Para evitar esto la proporción de ganancia o gain ratio ofrece una solución y además toma encuenta la información potencial de la división en si misma.

$$P(S,B) = -\sum_{i=1}^{t} \frac{Si}{S}\log(\frac{Si}{S}).$$

Gain ratio selecciona el atributo B que maximiza la función G(S,B)/P(S,B).

[33] KOHAVI, Ron and QUINLAN, Ross. *Decision Tree Discovery*, http://robotics.stanford.edu/~ronnyk/treesHB.pdf , 1999. Consultada el 7 de enero de 2005

3.2.3. Clustering

Los algoritmos de agrupación (o *clustering,* empleando un término sin traducir que se ha impuesto en la literatura técnica anglosajona y que se refiere a su capacidad de creación de *clusters,* es decir, clases o patrones) se salen de los métodos de aprendizaje supervisado.

"La única información que requieren los algoritmos de agrupación es la definición previa del vector de características. Algunos de estos algoritmos, a lo sumo precisan conocer también el número de clases."[34]

Un procedimiento de agrupación de clases recibe los datos de entrada y a partir de estos, sin supervisión de ningún tipo y de manera autónoma, genera grupos o clases (*clusters* o nubes). Por esta razón también se les ha denominado algoritmos de clasificación autoorganizada.

"Las técnicas de agrupación se utilizan cuando no se tiene un conocimiento suficiente acerca de las clases en que se pueden distribuir los objetos de interés. Esto puede acontecer en ciertas aplicaciones de biología, medicina, sociología, etc., en donde no se encuentran bien definidas las clases."[35]

Los algoritmos de agrupación varían entre sí por el mayor o menor grado de reglas heurísticas que utilizan e, inversamente, por el nivel de procedimientos formales involucrados.

De menor a mayor grado de complejidad los algoritmos a analizar son cuatro:

- Algoritmo de las distancias encadenadas (*chain map*)
- Algoritmo max – min

[34] [Reconocimiento de formas y visión artificial, 166]
[35] [ítem, 167]

- Algoritmo de las K – medias
- Algoritmo ISCDATA y la variante del mismo K – SODATA.

3.2.3.1. Algoritmo de las distancias encadenadas (chain map).

Es un algoritmo muy simple y no requiere ningún tipo de información *a priori*. Aunque los resultados pueden no ser, para determinadas situaciones, los óptimos, es recomendable como procedimiento inicial para tantear la agrupación de los objetos.

Retomemos el ejemplo de las características físicas visto con anterioridad donde tenemos la siguiente tabla:

ELEMENTO	ALTURA	CABELLO	OJOS
1	BAJO	RUBIO	AZULES
2	ALTO	PELIRROJO	AZULES
3	ALTO	RUBIO	AZULES
4	ALTO	RUBIO	MARRONES
5	BAJO	CASTAÑO	AZULES
6	ALTO	CASTAÑO	AZULES
7	ALTO	CASTAÑO	MARRONES
8	BAJO	CASTAÑO	MARRONES

Asignemos valores numéricos para las variables Bajo = 1, Alto = 2, Rubio = 1, Pelirrojo = 2, Castaño = 3, Azules = 1, Marrones = 2

ELEMENTO	ALTURA	CABELLO	OJOS
X1	1	1	1
X2	2	2	1
X3	2	1	1
X4	2	1	2
X5	1	3	1

X6	2	3	1
X7	2	3	2
X8	1	3	2

Cambiemos el orden de los vectores, ya que el algoritmo los ordenará (recordemos que de antemano conocemos a que clases pertenecen los objetos y el objetivo de este apartado es encontrar esta clasificación en datos donde no sea muy claro el patrón de comportamiento):

ELEMENTO	ALTURA	CABELLO	OJOS
X1	1	3	2
X2	1	1	1
X3	2	2	1
X4	2	3	2
X5	2	1	1
X6	2	3	1
X7	2	1	2
X8	1	3	1

El algoritmo se basa en el cálculo de la distancia Euclídea entre vectores; cuya formula es:

$$d_E(X_i, X_j) = \sqrt{\sum_{k=1}^{n} (X_{ik} - X_{jk})^2}$$

El primer paso consiste en tomar un vector de características al azar y calcular el valor de la distancia euclídea de el a los demás puntos de tal forma que la menor distancia resultante nos indicará el siguiente punto de la sucesión:

Tomemos X6:

$$d_E(X6, X1) = \sqrt{(2-1)^2 + (3-3)^2 + (1-2)^2} = 1.4142$$

de igual forma calculamos la distancia de X6 a los demás puntos

$$d_E(X6, X2) = 2.2360$$
$$d_e(X6, X3) = 1$$
$$d_e(X6, X4) = 1$$
$$d_e(X6, X5) = 2.2360$$
$$d_e(X6, X7) = 2.2360$$
$$d_e(X6, X3) = 1$$

Tomamos el primer punto que genera el valor mínimo en este caso X3 y calculamos las distancias con los puntos restantes:

$$d_E(X3, X1) = 1.73205$$
$$d_E(X3, X2) = 1.4142$$
$$d_E(X3, X4) = 1.4142$$
$$d_E(X3, X5) = 1$$
$$d_E(X3, X7) = 1.4142$$
$$d_E(X3, X8) = 1.4142$$

El punto que genera la distancia mínima es el X5, el proceso se repite hasta que ordenamos todos los puntos de acuerdo a la distancia mínima:

$$d_E(X5, X2) = 1$$
$$d_E(X2, X7) = 1.4142$$
$$d_E(X7, X4) = 2$$
$$d_E(X4, X1) = 1$$
$$d_E(X1, X8) = 1$$

La clasificación de los puntos se resume en la siguiente tabla:

CLASE	ELEMENTO	DISTANCIA
C1	X6	Elemento inicial
	X3	1
	X5	1
	X2	1
	X7	1.4142
C2	X4	2
	X1	1
	X8	1

Se detecta una clase cada que se produce un salto significativo en el valor de la correspondiente distancia euclídea.

Veamos la representación gráfica en el siguiente histograma:

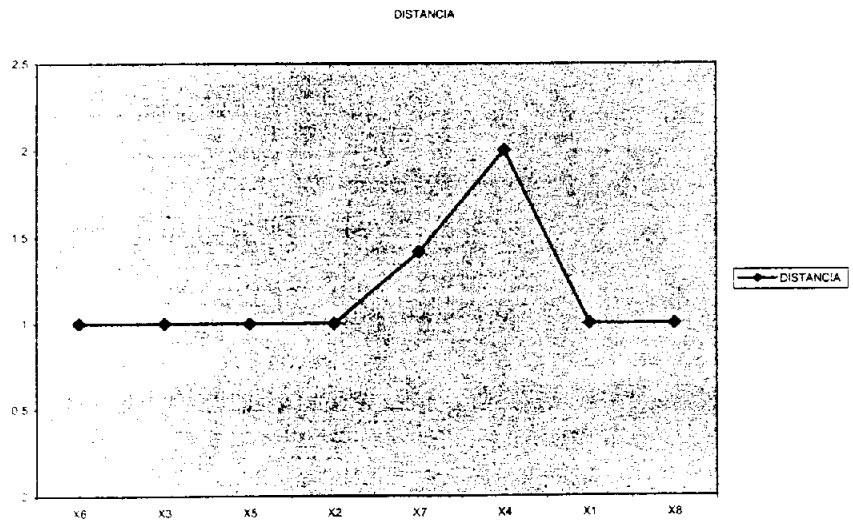


Figura 16. Resultado del algoritmo de distancias encadenadas

Si bien el clasificador no coloca los objetos como en la clasificación original si nos dice que existen dos clases y que uno se compone de tres elementos y otra de cinco elementos.

Es importante subrayar que el parámetro más delicado del algoritmo es el umbral de detección de una nueva clase. Un umbral excesivamente bajo dará lugar a clases ficticias, mientras que uno muy alto tenderá a agrupar objetos de diferentes clases en una misma clase.

En cuanto al primer elemento de la cadena el algoritmo no es muy sensible a esta elección salvo en distribuciones de clases muy caprichosas.

3.2.3.2. Algoritmo Max – Min.

Es un algoritmo heuristico que emplea como único elemento formal la distancia euclídea. Tampoco requiere ninguna información *a priori* respecto al número de clases existentes.

Tomando los datos de entrada siguientes:

ELEMENTO	A	B
X0	0	0
X1	1	1
X2	1	2
X3	4	3
X4	5	2
X5	5	3
X6	2	4
X7	3	5

Primero elegimos un elemento al azar; digamos X4, y generamos la primera clase. Después calculamos las distancias euclídeas del punto que elegimos a los demás puntos. Tomamos el punto que genera la distancia máxima como la segunda clase.

Para nuestro ejemplo:

$$d_{E\max} = d_E(X4, X0) = 5.3849$$

Así que tomamos X0 como nuestra segunda clase. Contamos ya con dos prototipos de clases X4 y X0.

Hacemos X4 = Z1 y X0 = Z2

Vamos a agrupar los puntos restantes a cada uno de los prototipos de las clases identificadas, para esto realizamos dos operaciones:

1.- Obtener la distancia euclídea, para cada vector X no agrupado a cada uno de los prototipos (Z1 y Z2).

2.- A continuación se toman las distancias mínimas generadas para cada clase y de éstas se toma la máxima. Si esta distancia es superior a una determinada fracción de la distancia $d_E(Z1, Z2)$ entre los prototipos de las dos clases previamente formadas, entonces se crea una tercera clase. Es decir:

sii $d_{\max} > d(Z1, Z2) * f \rightarrow$ se crea una clase Z3

con $0 < f < 1$. El prototipo (y único elemento hasta ahora) de la clase3 es el elemento correspondiente a la distancia máxima $d_{\max}$.

Regresando al ejemplo:

Los valores de las distancias generadas para cada uno de los prototipos de clase son:

Xn	d(Xn,Z1)	d(Xn,Z2)	Dmín
X1	4.1231	1.4142	1.4142
X2	4	2.2360	2.2360
X3	1.4142	5	1.4142
X5	1	5.8390	1
X6	3.605	4.4721	3.605
X7	3.605	5.8309	3.605

Tomamos la distancia máxima d(X7,Z1) y verificamos si:

d(X7,Z1) > d(Z1,Z2) * f ; hacemos f = 0.5

3.605 > 5.3851 * 0.5

3.605 > 2.6925

La condición se cumple y se genera una siguiente clase con prototipo X7; entonces X7 = Z3.

Volvemos a calcular la distancia de los puntos restantes a cada uno de los prototipos de las clases, tomamos el máximo elemento, del las mínimas distancias generadas y verificamos si se genera otra clase como se muestra a continuación.

Xn	d(Xn,Z1)	d(Xn,Z2)	d(Xn,Z3)	Dmín
1	4.1231	1.4142	4.4721	1.4142
2	4	2.2360	3.605	2.2360
3	1.4142	5	2.2360	1.4142
5	1	5.8390	2.8284	1
6	3.605	4.4721	1.4142	1.4142

La máxima distancia la genera el punto X2, así que verificamos si forma otra clase:

d(X2,Z2) > 1/3 [d(Z1,Z2) + d(Z1,Z3) + d(Z2,Z3)] ; esto es el promedio de de las distancias entre clases.

2.2360 > 1/3 [14.8209] * f ; donde f = 0.5

2.2360 > 2.47015

La condición no se cumple así que ya no se generan más clases.

Finalizamos la clasificación agrupando los elementos a los prototipos de las clases generadas de acuerdo a la mínima distancia.

Z1	Z2	Z3
X4	X0	X7
X3	X1	X6
X5	X2	

La siguiente figura muestra la distribución de las clases.

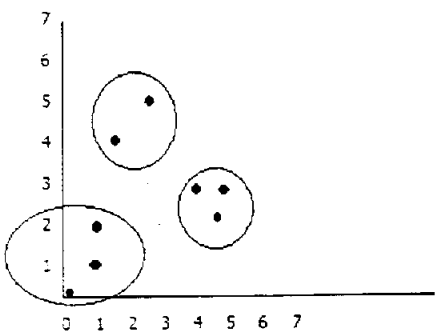


Figura 17. Clases generadas por el algoritmo max-min

El inconveniente principal de este algoritmo es la elección del factor f ya que es muy sensible. Una mala elección de f podría generar resultados no satisfactorios, no obstante al ser una técnica interactiva, es posible ensayar sucesivos valores hasta que se obtenga una agrupación final correcta.

3.2.3.3. Algoritmo K – medias

Tomaremos la descripción que se hace en el libro Reconocimiento de formas y visión artificial:

"El nombre de este algoritmo está ya consagrado en la literatura especializada y hace referencia a que existen k clases o patrones, siendo necesario, por tanto, conocer a priori el número de clases existentes.

Es un algoritmo sencillo, pero muy eficiente, siempre que el número de clases se conozca *a priori* con exactitud. Por su sencillez y robustez, ha sido muy utilizado.

Partiendo de un conjunto de objetos a clasificar X1, X2, ..., Xp el algoritmo de las k – medias realiza las siguientes operaciones.

Paso 1. Establecido previamente el número exacto de clases existentes, digamos k, se escogen al azar entre los elementos a agrupar k, vectores, de forma que van a constituir los centroides (al ser los únicos elementos) de las k clases. Es decir:

$$\alpha_1 : Z_1(1); \alpha_2 : Z_2(1) \ldots \alpha_k : Z_k(1)$$

en donde se ha introducido entre paréntesis el índice iterativo de este algoritmo.

Paso 2. Como se trata de un proceso recursivo con un contador n, en la iteración genérica n se distribuyen todas las muestras $\{X\}_{1\le j\le p}$ entre las k clases, de acuerdo a la siguiente regla:

$$X \in \alpha_j(n) \quad \text{sii} \quad \|X - Z_j(n)\| < \|X - Z_i(n)\|$$
$$\forall i = 1,2,...,k \,/\, i \neq j$$

en donde se han indexado las clases (que son dinámicas) y sus correspondientes centroides.

Paso 3. Una vez redistribuidos los elementos a agrupar entre las diferentes clases, es preciso recalcular o actualizar los centroides de las clases. El objetivo en el cálculo de los nuevos centroides es minimizar el índice de rendimiento siguiente:

$$J_1 = \sum_{X \in \alpha_i(n)} \|X - Z_i(n)\|^2 ; i = 1,2,...,k$$

Este índice se minimiza utilizando la media muestral de $\alpha_1(n)$:

$$Z_i(n+1) = \frac{1}{N_i(n)} \sum_{X \in \alpha_i(n)} X ; i = 1,2,...,k$$

Siendo $N_i(n)$ el número de elementos de la clase α_1 en la iteración n.

Paso 4. Se comprueba si el algoritmo ha alcanzado una posición estable. Es decir, si se cumple:

$$Z_i(n+1) = Z_i(n) \qquad \forall i = 1,2,...,k$$

Si se cumple, el algoritmo finaliza. En el caso contrario se va al paso 2."

Vamos a ver como trabaja el algoritmo, retomando el ejemplo utilizado en el algoritmo max - min:

Datos de entrada.

ELEMENTO	A	B
X0	0	0
X1	1	1
X2	1	2
X3	4	3
X4	5	2
X5	5	3
X6	2	4
X7	3	5

1.- Escogemos al azar tres elementos como prototipos de las clases iniciales (recuérdese que el algoritmo supone que se sabe el número de clases a organizar), tomamos X0, X1 y X2, hacemos:

Z1(1)=X0=(0,0)
Z2(1)=X1=(1,1)
Z3(1)=X2=(1,2)

2.- Distribuimos los elementos alrededor de las clases en base al principio de la mínima distancia euclídea:

Elemento	d(Xn,Z1(1))	d(Xn,Z2(1))	d(Xn,Z3(1))
X3	5	3.6055	3.1622
X4	5.3851	4.1231	4
X5	5.8309	4.4721	4.1231
X6	4.4721	3.1622	2.2360
X7	5.8309	4.4721	3.6055

La clasificación en esta primera iteración es la siguiente:

C1={X0} , C2={X1} , C3={X2, X3, X4, X5, X6, X7}

3.- Calculamos nuevamente los centroides de cada clase, en este caso los de las clases C1 y C2 no se mueven puesto que solo tienen un elemento; no ocurre lo mismo para C3 cuyo nuevo prototipo es la media de los elementos que le conforman:

Z3(2)=1/6[[1+4+5+5+2+3][2+3+2+3+4+5]]=(3.3333,3.1666)

4.- Como la condición de Z1(1),Z2(1),Z3(1) = Z1(2),Z2(2),Z3(2) no se cumple procedemos a repetir el paso 2, o sea calcular nuevamente las distancias de los puntos a clasificar a los nuevos centroides Z1(2), Z2(2) y Z3(2). Las distancias se muestran a continuación:

Elemento	d(Xn,Z1(2))	d(Xn,Z2(2))	d(Xn,Z3(2))
X2	2.2360	1	2.6083
X3	5	3.6055	0.6872
X4	5.3851	4.1231	2.0344
X5	5.8309	4.4721	1.675
X6	4.4721	3.1622	1.5723
X7	5.8309	4.4721	1.8634

La clasificación resultante de la segunda iteración es:

C1={X0} , C2={X1,X2} , C3={X3,X4,X5,X6,X7}

Volvemos a calcular los centroides (paso 3):

Z2(3)=1/2[[1+1][1+2]] =[1,1.5] y Z3=1/5[[4+5+5+2+3][3+2+3+4+5]]=[3.8,3.4]

Como la condición Z1(2),Z2(2),Z3(2) = Z1(3),Z2(3),Z3(3) (paso 4) no se cumple volvemos a calcular las distancias de los puntos a cada uno de los centroides.

Elemento	d(Xn,Z1(3))	d(Xn,Z2(3))	d(Xn,Z3(3))
X1	1.4142	0.5	3.6878
X2	2.2360	0.5	3.1304
X3	5	3.3541	0.4472
X4	5.3851	4.0311	1.8439
X5	5.8309	4.2720	1.2649
X6	4.4721	2.6925	1.8973
X7	5.8309	4.0311	1.7885

La clasificación resultante de la tercera iteración es igual a la anterior:

C1={X0} , C2={X1,X2} , C3={X3,X4,X5,X6,X7}

Esta vez la condición de Z1(2),Z2(2),Z3(2) = Z1(3),Z2(3),Z3(3) se cumple, así que el algoritmo para. Gráficamente la clasificación de los elementos se puede ver en la siguiente figura:

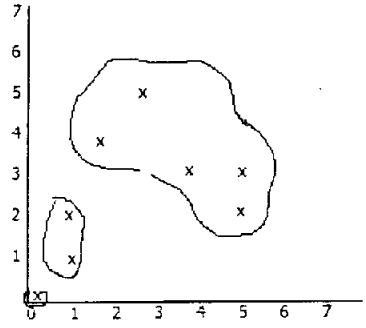


Figura 18. Clases generadas por el algoritmo de K - medias

3.2.3.4. Algoritmo ISODATA

El algoritmo ISODATA basa su estructura en el algoritmo K – medias, pero agrega más heurísticas. El algoritmo utiliza los siguientes parámetros:

Nc: Número de clases que se van formando.

K: Número estimado o deseado de clases a formar.

θ_N: Número mínimo de miembros que debe tener una clase para constituirse como tal.

θ_S: Desviación estándar máxima. Servirá para aplicar un criterio de división de un grupo en dos, cuando la desviación estándar σ del grupo sea superior.

θ_c : Es un parámetro de unión de dos clases, si la distancia entre dos clusters es menor que este valor.

L: Limita el número de fusiones que pueden realizarse en una iteración.

I: número máximo de iteraciones que puede ejecutar el algoritmo. Es el parámetro de parada.

Una vez definidos sus variables describimos el algoritmo:

Paso 1. Inicialización de variables. Se recomienda hacer Nc=k. Se escogen k elementos de los P elementos y se generan los primeros clusters de acuerdo a la mínima distancia euclídea.

Paso 2. Si existen cluster con un número de miembros inferior a θ_N se eliminan y se modifica el parámetro Nc.

Paso 3. Actualizar el valor de los centroides de las clases, calculando la media muestral de cada grupo o clase:

$$Z_i = \frac{1}{N_i}\sum_{j=1}^{N_i} X_j ; i = 1,2,...,Nc$$

siendo Ni el número de elementos de la clase.

Paso 4. Calcular la distancia media de cada cluster, mediante la siguiente expresión:

$$\overline{D_i} = \frac{1}{N_i}\sum_{j=1}^{N_i} \left\| X_j - Z_i \right\| ; i = 1,2,...Nc$$

Esta medida se utilizará posteriormente, junto con otras condiciones, para la posible división de un grupo.

Paso 5. Calcular la distancia media de todas las clases:

$$\overline{D} = \frac{1}{Nc}\sum_{i=1}^{Nc} N_i \overline{D}_i$$

Paso 6. Comprobar si es la última iteración, en cuyo caso θ_c se hace 0 y se va al paso 10.

Se realiza un *test* de posible unión de cluster siendo si $Nc \geq 2k$ en cuyo caso se salta al paso 10. Sino se cumple la condición se continúa con la secuencia.

Paso 7. Se calcula el vector de desviaciones típicas o estándar de cada grupo, según la siguiente expresión:

$$\sigma_i = \begin{pmatrix} \sigma_{i1} \\ \sigma_{i2} \\ \cdots \\ \cdots \\ \sigma_{in} \end{pmatrix} \quad ; \quad \sigma_{ij} = \sqrt{\frac{1}{N_i}\sum_{k=1}^{N_i}(X_{kj} - Z_{ij})^2}$$

i = 1, 2 ... Nc (clases)
j = 1, 2 ... n (características)
k = 1, 2 ... Ni (elementos de la clase Ci)

Paso 8. Obtener la desviación estándar máxima de cada grupo y formar el conjunto:

$$\{\sigma_1 \max, \sigma_2 \max, ..., \sigma_{Nc} \max\}$$

Paso 9. Posible división de clases, si se cumple la condición que $\sigma_j \max > \theta_S$ y además se cumpla una o las dos siguientes condiciones:

a) $D_{ij} > \overline{D}_j$ y $N_j > 2(\theta_N + 1)$

b) $N_j \leq K + 2$

entonces se divide la clase. Para esto calculamos dos nuevos centroides a partir del Zj, con la siguiente fórmula:

$$Z_{jk}^{+} = Z_{jk} + \gamma\sigma_j \max$$

$$Z_{jk}^{-} = Z_{jk} - \gamma\sigma_j \max$$

Paso 10. Calcular la distancia entre parejas de clusters:

$$D_{ij} = D_{ji} = \|Z_i - Z_j\|$$

i=1,2,...Nc-1; j=i+1,i+2...Nc

Paso 11. Comparamos estas distancias Dij con el parámetro θc, de forma que se toman (si existen) las L más pequeñas en orden creciente:

{D1,D2,...DL} con D1<D2<....<DL

Paso 12. Proceso de unión, comenzando con las parejas de clusters con las distancias menores, solo si ninguna de estas dos clases ha sido fusionada con otra en esta misma iteración, entonces se forma un cluster único cuyo centroide es:

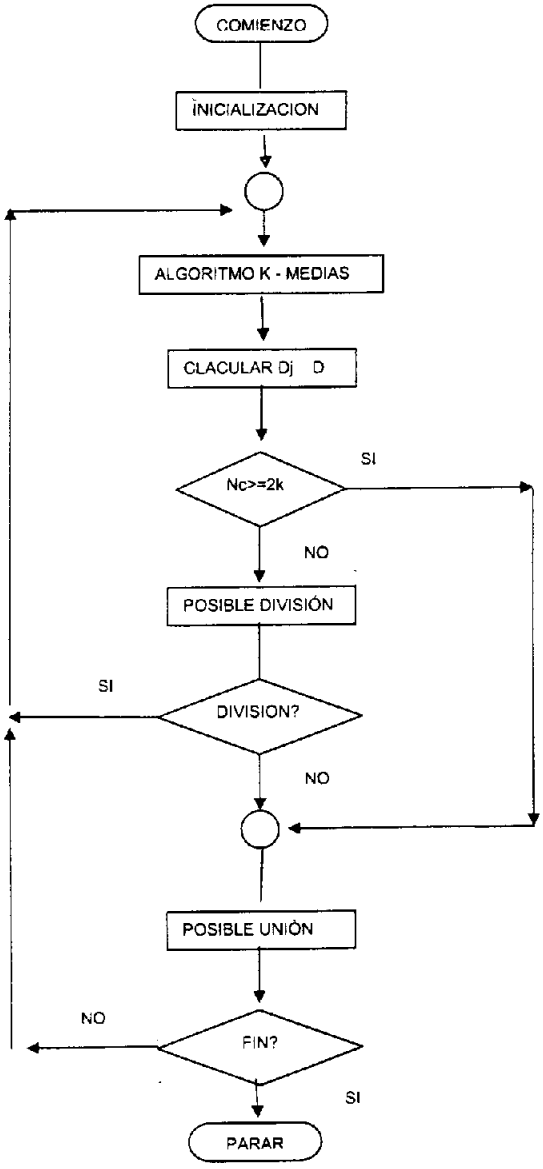
$$Z_{ij} = \frac{1}{N_i + N_j} * (N_i Z_i + N_j Z_j)$$

siendo Ni y Nj el número de muestras de los clusters C1 y C2 respectivamente, antes de la fusión.

Actualizar el parámetro Nc.

Paso 13. Comprobar si es la última iteración I.

El siguiente diagrama de flujo muestra el resumen del algoritmo ISODATA.


COMIENZO
INICIALIZACION
ALGORITMO K - MEDIAS
CLACULAR Dj D
Nc>=2k
SI
NO
POSIBLE DIVISIÓN
SI
DIVISION?
NO
POSIBLE UNIÓN
NO
FIN?
SI
PARAR

El método ISODATA se publicó por vez primera en 1965 y desde entonces se ha aplicado profusamente.

Los algoritmos de clasificación y agrupación han tenido un notable avance en los últimos diez años con la inclusión de ramas como las redes neuronales y la lógica difusa.

En muchos de los casos los proyectos de minería de datos son construidos en bases a dos o tres algoritmos distintos. Depende del diseñador la elección del método que más se adapte a sus necesidades y conocimientos.

4. El proceso de Minería de Datos

4.1. ¿Qué es IBusiness?

"Algo peor que no tener información disponible es tener mucha información y no saber que hacer con ella."[36]

La inteligencia de negocios se puede definir como el proceso de analizar los bienes o los datos acumulados en la empresa para extraer un conocimiento. Pongamos como ejemplo un hotel que tiene franquicias a nivel nacional y utiliza aplicaciones de BI para llevar un registro estadístico del porcentaje promedio de ocupación del hotel, así como los días promedio de estancia de cada huésped, considerando las diferencias entre temporada. Con esta información ellos pueden:

- Calcular rentabilidad en cada temporada del año
- Determinar el segmento de mercado.
- Calcular la participación de mercado de la franquicia y de cada hotel.
- Identificar oportunidades y amenazas.

Una aplicación de I Business puede realizar las siguientes funciones:

- Generar reportes globales o por secciones
- Crear una base de datos de clientes
- Crear escenarios con respecto a una decisión.
- Hacer pronósticos de ventas y devoluciones
- Compartir información entre departamentos
- Análisis multidimensionales.
- Generar y procesar datos
- Cambiar la estructura de toma de decisiones.

[36] SANCHEZ Montoya. Ricardo. Business intelligence... BI or not to BI. http: www. monografias.com trabajos14/bi/bi.shtml

- Mejorar el servicio al cliente.

4.2. ¿Por qué emprender un proceso de minería de datos?

El resultado final de la minería de datos es entregar información útil y clara a los administradores de la empresa o negocio. Para que este fin se cumpla, los desarrolladores del sistema deberán identificar los datos de la empresa que en verdad son relevantes.

Estos datos se han ido acumulando a través de los años y se encuentran en los sistemas de información de la empresa.

Emprender un proyecto de minería de datos no es tarea fácil y en muchos casos no está justificado plenamente.

Según el Lic. Ricardo Sánchez, Asistente del Departamento Académico de Mercadotecnia del ITESM Campus Monterrey, si se responde afirmativamente a por lo menos una de las siguientes preguntas entonces se es candidato a beneficiarse con soluciones de I Business.

a) ¿Pasa más tiempo recolectando y preparando información que analizándola?

b) ¿En ocasiones le frustra el no poder encontrar información que usted está seguro que existe dentro de la empresa?

c) ¿Pasa mucho tiempo tratando de hacer que los reportes en Excel luzcan bien?

d) Quisiera tener una guía sobre las cosas qué han sucedido cuando los administradores anteriores implementaban alguna estrategia

e) No sabe qué hacer con tanta información que tiene disponible en la empresa.

f) ¿Quiere saber qué productos fueron los más rentables durante un periodo determinado?

g) ¿No sabe cuáles son los patrones de compra de sus clientes dependiendo de las zonas?

h) ¿Ha perdido oportunidades de negocio por recibir información retrasada?

i) ¿Trabaja horas extras el fin de mes para procesar documentos o reportes?

j) ¿No sabe con certeza si su gente está alcanzando los objetivos planeados?

k) ¿No sabe si mantiene una comunicación estrecha entre las diversas áreas de su empresa hacia una estrategia común?

l) ¿No tiene idea de por que sus clientes le regresan mercancía?

"La minería de datos no es la respuesta a todos los problemas... En realidad el minado de datos es solo una pequeña parte de todo el proceso. Uno necesita contestar preguntas como: ¿Hay necesidad de minar los datos?, ¿Tiene los datos correctos en el formato correcto?, ¿Tiene las herramientas adecuadas?, mas importante aun ¿Dispone del personal adecuado para realizar el trabajo?, ¿Dispone de los fondos suficientes para realizar el proyecto?"[37]

[37] THURAISINGHAM, Bhavani. *Data Mining Technologies, techniques, tool and trends.* CRC Press USA 1999. P. 93

Contestar estas preguntas es importante antes de embarcarse en un proyecto de minería de datos.

El data mining se ha utilizado en varias áreas y como ejemplo encontramos sistemas que realizan diagnósticos médicos, análisis financiero, **desempeño estudiantil**, análisis de marketing y ventas.

Los proyectos donde la minería de datos ha proporcionado resultados exitosos han considerado los siguientes criterios:

Criterios prácticos:

- Impacto significativo
- No existen métodos alternativos
- Existe soporte para su desarrollo
- No existen problemas de legalidad o violación a información privilegiada

Criterios técnicos:

- Atributos relevantes
- Poco ruido en los datos
- Conocimiento del dominio

Si bien los problemas en las empresas han existido siempre, es hasta ahora cuando las computadoras han alcanzado un desarrollo considerable, los sistemas de bases de datos son mejores y encontramos lenguajes de programación los suficientemente robustos y estructurados para emprender la programación de estos sistemas de minado de datos. La minería de datos es una realidad ahora y siempre que la cantidad de datos lo justifique debemos dar el paso.

4.3. Pasos de la Minería de Datos.

Bien después de encontrar una justificación para iniciar un proyecto de minería y de dejar bien claros los objetivos que queremos lograr los pasos a seguir según Thuraisingham son:

+ Identificar los datos
+ Preparar los datos
+ Minar los datos
+ Obtener información útil
+ Identificar acciones
+ Implementar acciones
+ Evaluar los beneficios
+ Determinar las acciones siguientes
+ Iniciar un nuevo ciclo

¿Dónde encontrar los datos?, estos podrían estar alrededor del mundo y no solo en la empresa, podrían estar en papel e incluso podrían estar en la cabeza de la gente. Así que necesitamos estructurar que datos necesitamos y donde pueden estar, entonces tomarlos y utilizarlos.

Una vez que hemos obtenido los datos debemos prepararlos. En muchos aspectos esta es la tarea más difícil de todo el proceso ya que datos erróneos o incompletos arrojaran resultados no validos. Es aquí donde se diseña el **data warehouse.**

La existencia de valores faltantes es un hecho negativo ya que disminuye la calidad de los datos. Las tres técnicas básicas para tratar valores faltantes son las siguientes:

- Eliminar todas las filas que contengan algún valor faltante, con la consecuente reducción de la información.

- Recodificar los valores faltantes a una constante que sea tratable por el algoritmo aplicado.

- Sustituir los valores faltantes por alguno que no perturbe la información proporcionada por los valores presentes en cada atributo.[38]

Bien tenemos los datos y ya los colocamos en el formato adecuado, debemos entonces determinar la técnica y las herramientas que utilizaremos para minar los datos. Las técnicas de **aprendizaje automático** como árboles de decisión y clustering, redes neuronales y reconocimiento de patrones están incluidas en este punto.

Después de obtener patrones y tendencias necesitamos implementar acciones en beneficio del negocio. El factor humano es muy importante en aquí.

"Ser realista con lo que podemos minar y que podemos hacer nosotros... determinar si ¿tenemos el poder humano?, si ¿entrenaremos a la gente o la contrataremos fuera?... Los clientes, los contratistas y los desarrolladores del sistema deben trabajar muy de cerca para hacer un minado exitoso."[39]

Si el proyecto fracasa se debe tener cuidado de no señalar a nadie ya que no podemos olvidar que estamos ante una nueva tecnología. Aprender de la experiencia, hablar con gente que ha pasado por este proceso y ver que podemos hacer diferente para evitar estos errores. En muchos casos es mejor iniciar un proyecto piloto o un prototipo antes de echar a andar un monstruo para la corporación.

[38] BERGOS, Massagué Jordi. BEAN: Behavior Analyser. 21 de mayo de 2004 memoria.pdf , P. 11
[39] THURAISINGHAM, Bhavani.Op cit. P. 100

4.4. Algunos datos curiosos y proyectos exitosos de Minería de Datos.

Dentro del ambiente de la minería de datos hay algunos ejemplos de resultados que son constantemente utilizados en la literatura sobre el tema.

- Casi el 5% de los clientes de un banco nacieron el 11 de noviembre de 1911. Razón: El campo de fecha de nacimiento es obligatorio y la manera más fácil para pasar al siguiente campo de captura es tecleando 111111.

- Clientes con nombres cortos en un banco tienden a ahorrar grandes cantidades de dinero y luego retirarlas. Razón: Nombres orientales.

- Un caso que ha cobrado fama es el de una compañía de seguros que después de invertir millones de dólares en ejercicios de minería de datos, logró obtener una regla de asociación de la más alta certeza, pero que decía que el 95% de los esposos eran hombres.

- Los que compran coches de color rojo en Francia tienden a no pagar su préstamo de coche.

- Clientes que compran pañales tienden a comprar cerveza.

- Personas mayores (arriba de 65) no responden a ofertas de cuentas de retiro. Es obvio pero entonces por que se les envía propaganda.[40]

Como podemos observar algunas de las reglas de asociación mostradas podrían tener una interpretación difícil de encontrar e incluso alguna de ellas podría no tener ningún significado.

[40] BERGOS. Massagué Jordi. Op cit P.8

Dentro de la página web de la compañía SAP, empresa fundada en 1970 y dedicada a brindar soluciones de negocios, para todo tipo de industria y mercado, encontramos las siguientes empresas (algunas muy conocidas) que han utilizado herramientas de I Business de manera exitosa.

Air Products. Compañía con un valor de 6 billones de dólares, implemento procesos de I Business en 2002, facilitando a sus gerentes y analistas información relevante para la toma de decisiones alrededor del mundo.

BMW Grup. Es uno de los principales constructores de automóviles y motocicletas. En la actualidad aproximadamente 2,100 usuarios monitorean el desarrollo de la producción y los costos de material, con una herramienta de reportes basada en el web, que les permite reaccionar rápidamente cuando la acción es requerida.

Coca – Cola. Buscaba un entorno sencillo para hacer la lectura de los datos financieros accesible a los gerentes ejecutivos alrededor del mundo. Coca – Cola usa ahora consolidación de negocios y planeación a través de un data warehouse.

Valvoline. Utiliza herramientas de data warehouse para que sus empleados tengan mejor acceso a información crítica. Valvoline Company ayuda a mantener al mundo rodando suavemente produciendo lubricantes de autos. Tiene presencia alrededor del mundo con 4,600 empleados en más de 140 países y las ventas en 2002 fueron de cerca de $1.2 billones de dolares.

Algunas otras empresas que utilizan I Business y minería de datos:

- **Bayer AG**, Alemania
- **F. Hoffman – La Roche AG**, Suiza
- **Deutsche Telekom AG**, Alemania
- **Hercules Inc.** Estados Unidos

- **Nestles,** Suiza
- **Oxford University Press,** Inglaterra
- **Siemens,** Alemania
- **Swiss Army,** Suiza
- **Tellabs,** Estados Unidos
- **Texaco Ltd,** Inglaterra
- **Xerox Ltd.,** Estados Unidos

4.5. Retos

Podemos resumir el proceso de minería de datos en tres momentos obtención y preparado de los datos, análisis y obtención de patrones y análisis de resultados y toma de decisiones.

Cada momento en el proceso tiene retos a vencer por parte de los desarrolladores de sistemas y demás trabajadores del conocimiento.

" Uno de los primeros retos es la facilidad en que se puede caer en una falsa interpretación de los resultados...La estadística es una herramienta poderosa, y es elemento crucial en el análisis de datos. Sin embargo, a veces enfrentamos problemas muy serios en la interpretación de sus resultados."[41]

El uso de software de Minería de Datos puede poner a disposición de un "analista" (o minero de datos), la posibilidad de crear fácilmente indicadores, resúmenes, gráficas, y aparentes tendencias, sin un verdadero entendimiento de lo que se está reflejando. Es decir, resulta más fácil hacer creíble una falsedad, posiblemente porque la produjo una computadora.

[41] ESTIVILL Castro, Vladimir., *Tres retos de la mineria de datos.*
www.lania.mx biblioteca newsletters 1999-otono-invierno retos_mineria.html
Laboratorio Nacional de Informática Avanzada A. C. 1999, Consultada el 3 de febrero de 2005

¿Cómo lograr que las herramientas de minado de datos sean accesibles a cualquiera incluso a quienes no dominan la estadística y como al mismo tiempo lograr que las inferencias que se produzcan sean realmente validas?

Es muy fácil equivocarse al hacer minería esto es debido a que ésta es una herramienta explorativa no explicativa, esto es explora los datos para sugerir una hipótesis, es un error tomar esta hipótesis como una explicación. Los más duros criticos de la minería de datos toman este hecho como el principal argumento en contra del data mining, ya que se cae muy fácilmente en la especulación.

Otro reto tiene que ver con la granularidad en el tiempo. Dentro del data warehouse se almacena por periodos de tiempo, por ejemplo meses o quincenas, así que si necesitamos analizar el comportamiento diario o semanal sería imposible. ¿Cómo dividir el tiempo para arrojar relaciones significativas y no perder detalle.

La privacidad de los datos es un reto a vencer pero más que nada en el ámbito legal, ¿Quién puede hacer uso de los datos? y ¿en que medida las acciones tomadas afectan a terceras personas?

"Otros factores que pueden crear una desilusión de las promesas de la Minería de Datos son:

1) que se requiera de mucha experiencia para utilizar herramientas de la tecnología, o que sea muy fácil hallar patrones equívocos, triviales o no interesantes,

2) que no sea posible resolver los aspectos técnicos de hallar patrones en tiempo o en espacio,

3) que exista una reacción del publico por el uso indiscriminado de datos personales para ejercicios de Minería de Datos, que obligue a los legisladores a imponer restricciones exageradas (y tal vez absurdas) al uso de la tecnología."[42]

Las siguientes son opiniones de ejecutivos de empresas estadounidenses sobre lo que esperan que suceda en el corto o mediano plazo con respecto a BI y la Minería de Datos:

En aproximadamente cinco años, veremos un incremento dramático del 40%, en el número de usuarios finales que utilicen herramientas de BI... - Frank Gelbart, CEO, Appfluent Technology Inc., Arlington, Va.

En pocos años, las ventajas competitivas vendrán del uso de BI para entender el comportamiento y preferencias del consumidor a un nivel de segmentación angosto, incluso individual para hacer ofertas a la medida...- Jeff Zabian, Vice President, Seurat Co., Boulder, Colo.

Dentro de dos o tres años, las compañías abandonarán el método tradicional de hacer negocios con ajustes trimestrales. En vez de eso utilizarán la BI y desarrollarán herramientas administrativas como estrategia para responder a cambios en tiempo real de mercado. – Rob Ashe, President & Chief Operating Officer, Cognos Inc., Burlington, Mass.

Los usuarios demandarán mayor integración entre los números y su interpretación. Así mismo, todas las aplicaciones de BI incluirán herramientas de administración de contenido o bien administración de conocimiento. – Brian Hartlen, Señor Vice President, Comshare Inc., Ann Arbor, Mich.

¡Los negocios son una guerra! Como en cualquier guerra, ,sobrevivir depende de la capacidad para actuar rápidamente en un ambiente cambiante. BI será como un

[42] ESTIVILL Castro, Vladimir, *item.*

comando de control central para rastrear variables como el desarrollo operacional, las condiciones del mercado y el desarrollo de los competidores, todas ellas en tiempo real – Sol Klinger, Director, Sterling Management Solutions Inc., Princetown, N. J.

Al mejorar la selección de a quién dirigir los mensajes de mercadotecnia, BI puede ahorrar más de $200 billones de dólares al año por desperdicio de publicidad y mercadotecnia directa... - Dave Morgan, CEO, Tacoda Systems Inc., New Cork.

La Información de BI permite a una compañía crecer y explotar futuras oportunidades y al mismo tiempo, es el blanco para espionaje corporativo, crimen y terrorismo computacional... - Ryon Packer, Vice President, Intrusion Inc., Richardson, Texas.

4.6. La interfase de usuario.

Como en cualquier sistema tener una buena interfase de usuario es crítico para minar datos. Algunos de los primeros sistemas manejadores de base de datos como DBase o Clipper tenían interfaces muy primitivas. Los usuarios gastaban gran cantidad de tiempo escribiendo consultas SQL o escribiendo programas. En la actualidad las herramientas para generar reportes, consultas, programas de aplicación tienen excelentes interfaces de usuario.

Sin embargo las interfases para realizar minería de datos en la actualidad dejan mucho que desear.

Los componentes que debe incluir una interfase para un programa de minería de datos se observan en la figura 19:

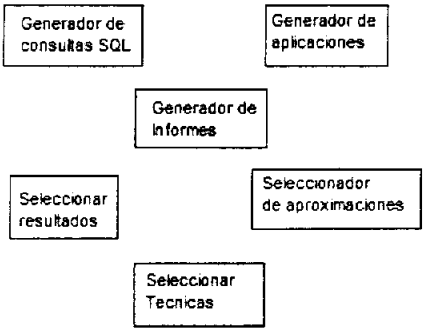


Figura 19.Ejemplo de una Interfase de Usuario para minar datos

El ambiente del mundo de los negocios exige una aplicación cada vez más eficiente de la información disponible. BI genera conocimiento del negocio y pone a disposición de los usuarios la información correcta en el lugar correcto, ofrece muchos beneficios generando una ventaja competitiva. Una aplicación de BI debe reunir cuatro componentes: multidimensionalidad, data mining, agentes y data warehouse.

5. Herramientas comerciales para Minería de Datos

5.1. Empresas e instituciones académicas dedicadas a la minería de datos.

A continuación se listan algunas empresas dedicadas a crear aplicaciones de minería de datos. La información ha sido tomada del web y no pretende ser más que una referencia para el lector interesado en el tema. También se pretende dar un vistazo a las instituciones académicas que se encuentran realizando tareas de investigación sobre este rubro.

5.1.1. Empresas que desarrollan soluciones de minería de datos.

HP México. Filial de HP con un extenso portafolio de aplicaciones para brindar soporte a las necesidades de negocios, poniendo énfasis en:

- Determinar el alcance y los objetivos de sus necesidades de negocios
- Instalar y configurar mercados de datos y almacenes de datos
- Proseguir los servicios de soporte y administración continua [43]

Pearson es una agencia de investigación "full service" con áreas Cualitativa, Cuantitativa, "business to bussiness", Internet, Minería de Datos, Telefónica, Internacional, Investigación en el punto de venta e Investigación Documental. Es una de las empresas con más experiencia en investigación de mercado y opinión pública en México: opera desde 1988 y es una de las empresas fundadoras de la Asociación Mexicana de Agencias de Investigación de Mercado y

[43] HP México. http:// www3.hp.com/servicios/aplicaciones_empresariales/enter_negocios_mineria.html. Consultada el 30 de junio de 2005

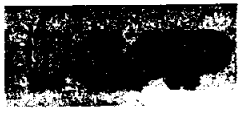
Opinión Pública (AMAI). Hemos recibido el premio agencia del año "AD Cebra" en dos ocasiones (1997 y 2000). [44]

MB Systems de México es una empresa creada con el respaldo de una de las corporaciones líder en el mercado de servicios en América Latina: **MB & Associates**.

Con más de 4 años en el mercado **MB Systems de México** ofrece tecnologías de información, centros de servicio autorizados, comunicaciones, desarrollo y soporte, ofreciendo bajo un mismo techo el más amplio inventario de productos necesarios para el optimo funcionamiento de su organización en tecnologías que van desde una computadora personal hasta la integración de sus oficinas en México con aplicaciones en tiempo real a otras partes del mundo.[45]

INFOTEC. Se dedica a mejorar la competitividad de las organizaciones públicas y privadas a través del uso estratégico de la tecnología, y operamos como una institución pública que desarrolla negocios rentables, globales e innovadores.

INFOTEC es un Centro Público de Innovación y Desarrollo Tecnológico autosuficiente que eleva la competitividad de las organizaciones a través del desarrollo e implantación de conceptos, modelos y sistemas estratégicos con tecnologías innovadoras basadas en Internet y con capacidad de convertirse en estándares de mercado.[46]



INVAP fue creada en 1976, mediante un convenio entre la Comisión Nacional de Energía Atómica de Argentina y el Gobierno de la Provincia de Río Negro. En la actualidad sus oficinas y talleres cubren una superficie de más de 10.000 metros cuadrados.

[44] Pearson. http: www.pearson research.com nuestra-empresa.phtml. Consultada el 30 de junio de 2005
[45] MBSystems de México. http: www.mbsystems.com.mx somos.html. Consultada el 30 de junio de 2005
[46] INFOTEC. http: www.infotec.com.mx wb2 infotec into_quienes_somos.htm Consultada el 30 de junio de 2005

La sede principal de INVAP se encuentra en uno de los mayores centros turísticos argentinos: la ciudad de San Carlos de Bariloche, dentro del Parque Nacional Nahuel Huapi, provincia de Río Negro.
INVAP ocupa a más de 360 empleados, los que, sumados a las empresas asociadas, contratistas y proveedores, implica un total de unas 700 personas.
INVAP cuenta con

- Un cuerpo altamente experimentado en el desarrollo de sistemas tecnológicos así como en el manejo de proyectos complejos.
- Un sistema de calidad que responde a las normas mas exigentes.
- Los sistemas técnicos y administrativos necesarios para control de proyectos.
- Mas de veinte años de experiencia exitosa en el gerenciamiento de proyectos que involucran desarrollos novedosos.

INVAP ha hecho un esfuerzo humano y económico importante y exitoso en la apertura de mercados; es así que hoy la Argentina es conocida como exportadora confiable de instalaciones nucleares y los, equipos y sistemas de control asociados.tecnología nuclear, También ha exportado equipos de Cobaltoterapia, así como equipamiento y sistemas de automatización para proyectos industriales.

En el área de la Tecnología Espacial, INVAP es la única empresa argentina calificada por la NASA para la realización de proyectos espaciales, y como tal ha demostrado su capacidad para el diseño, construcción, ensayo y operación de satélites.[47]

DÆDALUS

DAEDALUS desarrolla productos y sistemas para facilitar y potenciar la compartición de información en organizaciones y la utilización de Internet como fuente de conocimiento. La tecnología disponible incluye

[47] INVAP. http: www.invap.net about perfil.html. Consultada el 30 de junio de 2005

recuperación y extracción automática de información, categorización automática de documentos, gestión de perfiles de intereses de usuarios, robots de Internet y automatización del flujo electrónico de documentos.[48]

DATOLOGIA tiene como misión proveer de tecnologías de información de alto valor agregado a aquellas personas que participan en lograr los objetivos estratégicos de su organización.

Con mas de diez años de experiencia en el área de sistemas y en la extracción de datos para el análisis, tiene como propósito servir a las organizaciones en estos tres importantes conceptos: Procesos, Información y Conocimiento.

Construye almacenes de datos que permiten a las organizaciones tener su información histórica, con la que es posible combinar y medir variables de distintas fuentes (Data Warehousing). También a través de algoritmos estadísticos o inductivos buscamos relaciones ocultas (Data mining) ambas técnicas usadas para apoyar el proceso de toma de decisiones de mediano y largo plazo.

Estos servicios agregan valor porque permiten a nuestros clientes acceder y manipular información de manera rápida y precisa para apoyar el logro de sus objetivos estratégicos, como lo es incrementar ganancias.[49]

INFOMEDIA

INFOMEDIA. Es una empresa que cuenta con una amplia experiencia en la implantación de proyectos de bases de datos usando diferentes tecnologías. Es la única empresa de América Latina en formar parte del *IBM Data Management Business Partner Advisory Borrad.*

[48] DAEDALUS. http: www.daedalus.es
[49] DATOLOGIA. http: www.datologia.com/empresa.html

Son los primeros consultores en México certificados en DB2 UDB en sus diferentes plataformas además colaboró en el desarrollo del procedimiento de certificación de IBM para DBA a nivel mundial y tiene la coautoría del libro *Fundamentos de las Estructuras de Datos Relacionales*, actualmente en uso en varias instituciones de América Latina.[50]

PRO CALIDAD

Se anuncian como la organización de más rápido crecimiento en la industria de **Business Intelligence**. Con capacidad de entregar soluciones gerenciales con resultados tangibles en términos de productividad y retorno de la inversión que los han llevado a ser líderes el mercado.[51]

PROFIN
Inteligencia en tus decisiones

PROFIN México es una empresa de consultoría en Inteligencia de Negocios y de capacitación.

Asine

ASINE S.A. es una empresa especializada en el desarrollo de soluciones informáticas para todo tipo organizaciones, que tiene como criterio principal de éxito, **la satisfacción total de sus clientes**.

Su ámbito de negocios cubre desde servicios de soporte y desarrollo de sistemas

[50] INFOMEDIA. http: www.infomedia.com.mx
[51] PROCALIDAD. http: procalidad.com/compania/index.aspx

computacionales, hasta reingeniería de negocios y asesoría en el uso estratégico de la Tecnología de la Información.

Con una política de asociación con empresas líderes en productos informáticos, **ASINE** ofrece soluciones, productos y servicios de la más alta calidad. **ASINE** es socio de **Microsoft**, **Lumigent** y **Megaputer** y en colaboración con ellos está en condiciones de ofrecer las soluciones más avanzadas y completas en aplicaciones de tecnología de la información y de seguridad informática basada en la Web, así como en la utilización de herramientas **OLAP** y de **Minería de Datos** en el ámbito de la toma de decisiones de alto nivel y de la inteligencia de negocios.[52]

El crecimiento de las empresas que se dedican a la gestión del conocimiento está creciendo día a día. La lista anterior desde luego podría ser muy extensa.

Dentro de esta lista también se encuentran transnacionales como IBM con los productos IBussines y Microsoft con la gama de servidores de base de datos SQL Server 2000 que incluye los Analysis Services que se estudiarán más adelante.

5.1.2. Instituciones educativas que realizan proyectos y estudios sobre minería de datos.

El ambiente académico sigue aportando infinidad de documentos y sitios de interés en Internet la lista siguiente es una selección de las referencias encontradas en el Buscador de Yahoo.

Red española de minería de datos y aprendizaje automático

Grupo AWEG (*Adaptive Web Engineering Group*)

UNIVERSIDAD INDUSTRIAL DE SANTANDER, ESCUELA DE INGENIERIA DE SISTEMAS E INFORMATICA

[52] ASINE S. A. http: www.asine.cl quienessomos.aspx

. Grupo MINERVA - Universidad de Sevilla Coordinador: José C. Riquelme Santos (Coordinador de la Red)

Grupo GICAP (Inteligencia Computacional Aplicada) - Universidad de Burgos
Coordinador: Emilio S. Corchado Rodríguez

Grupo de Ciencias de la Computación y Sistemas Inteligentes - Universidad de Cantabria
Coordinador: Eduardo Mora Montes

Grupo Sistemas Inteligentes y Minería de Datos - Universidad de Castilla-La Mancha
Coordinador: José A. Gámez Martín

Grupo AYRNA (Aprendizaje y Redes neuronales Artificiales) - Universidad de Córdoba
Coordinador: Cesar Hervás Martínez

Grupo de Investigación en Sistemas Inteligentes - Universitat de Girona
Coordinador: Joaquim Melendez Frigola

Grupo SCI2S (Soft Computing y Sistemas de Información Inteligentes) Universidad de Granada Coordinador: Francisco Herrera Triguero

Grupo IDBIS- Intelligent Databases and Information Systems - Universidad de Granada
Coordinador: Juan Carlos Cubero

Grupo de investigación Sistemas Inteligentes. Universidad de Jaén
Coordinador: María José del Jesus Díaz

Grupo de Reconocimiento de Formas y Visión por Ordenador – Universidad Jaume I de Castellón Coordinadores: Filiberto Pla y J. Salvador Sánchez

Grupo de Inteligencia Artificial y Sistemas - Universidad de Las Palmas de Gran Canaria
Coordinador: Javier Lorenzo

Grupo de Investigación y Aplicaciones en Ingeniería Artificial (IA)2 - Universidad de Málaga Coordinador: Francisco A. Triguero Ruiz

Grupo de Sistemas Inteligentes - Universidad de Murcia Coordinadores: Antonio Gómez-Skarmeta y Juan A. Botía

Grupo de investigación en Descubrimiento y Representación del Conocimiento - Universidad Pública de Navarra Coordinador: Ramón Fuentes González

Grupo de Aprendizaje Automático - Universidad de Oviedo Coordinador: Antonio Bahamonde Rionda

Intelligent System Group - Universidad del Pais Vasco - Euskal Herriko Unibertsitatea
Coordinador: Pedro Larrañaga Múgica

Grupo GREC (Grup de Recerca en Enginyeria del Coneixement) - Universitat Politécnica de Catalunya - ESADE Coordinador: Andreu Catalá Mallofré

Grupo SOCO (Soft Computing Research Group) - Universitat Politécnica de Catalunya
Coordinadora: Àngela Nebot Castells

Grupo LARCA (Laboratorio de Algoritmica Relacional, Complejidad y Aprendizaje) - Universitat Politécnica de Catalunya Coordinador: José Luis Balcázar

Grupo KEMLG (Knowledge Engineering and Machine Learning Group) - Universitat Politécnica de Catalunya Coordinador: Miquel Sànchez-Marrè

Grupo de Técnicas Híbridas de Data Mining - Universitat Politécnica de Catalunya Coordinadora: Karina Gibert

Grupo KD&DM (Knowledge Discovery and Data Mining) - Universidad Politécnica de Madrid - Universidad Carlos III de Madrid Coordinadora: Ernestina Menasalvas Ruiz

Grupo MIP (Multi-paradigm Inductive Programming) - Universitat Politécnica de Valencia
Coordinador: Mª José Ramírez Quintana

Grupo de Investigación en Sistemas Inteligentes - Universitat Ramon Llull Coordinador: Josep M. Garrell Guiu

Universidad de Salamanca Coordinadora: Vivian F. López Batista

Grupo de Reconocimiento de Formas y Aprendizaje - Universitat de Valencia Coordinador: Francesc J. Ferri

Grupo de Sistemas Inteligentes - Universidad de Valladolid Coordinador: Carlos Alonso González

El área de Ciencias de la Computación del CIMAT es actualmente uno de los grupos en computación más importante del país. Sus actividades se dirigen a la investigación, el desarrollo tecnológico y a la formación de recursos humanos. Forma parte de la Red de Desarrollo e Investigación en Informática (REDII) auspiciada por CONACYT, la cual agrupa a los 11 principales grupos nacionales de investigación y docencia en esta disciplina.

El área cuenta con un programa de Doctorado en Ciencias, un programa de Maestría y apoya la Licenciatura en Computación de la Universidad de Guanajuato.

La Universidad Nacional Autónoma de México cuenta también con investigadores dedicados a esta rama de la informática al igual que el Instituto Politécnico Nacional a través del CINVESTAV.

5.2. Sistemas comerciales y académicos para realizar minería de datos.

El área comercial y académica ha incursionado en el campo de la minería de datos con una serie de productos.

Estas son algunas de las aplicaciones más nombradas en el ambiente dedicado al desarrollo de aplicaciones de gestión del conocimiento; su inclusión en este trabajo es para brindar una panorámica de los alcances y las nuevas tendencias de investigación y desarrollo.

COGNOS

COGNOS. Es el principal proveedor del mundo en software de planeación y consolidación empresarial, Bussines Intelligence y administración empresarial.

Los clientes usan las soluciones de IBussines de COGNOS para el soporte de decisiones, minería de datos y creación de informes.

DataEngine.

DataEngine es una herramienta de Mineria de Datos desarrollada por la empresa alemana MIT GmbH. DAEDALUS es distribuidor de DataEngine en España y en la mayor parte de los paises iberoamericanos.

DataEngine es una buena herramienta para extraer información útil a partir de datos. Porque los datos en sí mismos no sirven para sacar conclusiones ni para tomar decisiones.

DataEngine es un entorno potente, abierto y económicamente competitivo. Incluye gran variedad de modelos de aprendizaje y métodos de entrenamiento: redes neuronales, reglas borrosas (fuzzy), mapas autoorganizativos, etc.

Synaptris ,dos veces Beacon Award Winner, por la mejor solución herramienta/utilitario con mas de 1800 clientes (y 200.000 usuarios) en 44 países del mundo. Las soluciones Synaptris proveen un alto retorno de la inversión (ROI) para sus clientes con soluciones de consulta y reporteadores para la plataforma Lotus-Domino. Los productos Synaptris mas populares son IntelliPRINTPLUS, IntelliVIEW y PrintCAL estos productos facilitan las tareas de reporteo, impresión , análisis y colaboración para los usuarios Lotus Notes

bpDATAEXPLORER

Se trata de un sistema de explotación de datos, basado en las tecnologías OLAP y minería de datos.

La aplicación está diseñada para que sea agradable, fácil e intuitiva para el usuario.

Será una aplicación indispensable para directivos y controllers, ya que podrán realizar un seguimiento exhaustivo del estado y funcionamiento de su empresa, a partir de los datos introducidos en el trabajo diario de la empresa. Sin necesidad de realizar introducciones de datos accesorias ni de complicados procesos de configuraciones.

Con **bpDATAEXPLORER** el usuario dispondrá de la posibilidad de prever o simular cualquier circunstancia o evento, relativo a ventas, compras, contabilidad o producción. [53]

[53] Bussines Progress. http:// www.b-progress.com/productos.html

Datahouse Company ahora ofrece el sistema **Pi Five** de **data warehousing** y **data mining**; **Pi Five** no es simplemente un mancomunador de información. Es un software que le brinda cruzamiento de datos y armado de cubos multidimensionales de análisis para que directivos, analistas y especialistas encuentren, entre la multitud de datos operativos que brindan sus sistemas administrativos, las tendencias comerciales de su organización, el descubrimiento de comportamientos y estadísticas, la simulación de operaciones basadas en datos históricos ciertos.

Oracle de México

Datawarehouse Consulting es distribuidor de los siguientes productos:
Manejadores de Bases de Datos
Software para la Administración
Herramientas de Desarrollo
Productos BI
Aplicaciones: ERP, MRP, CRM, E-commerce
Soluciones

Computer Associates

Los productos de Computer Associates, permiten crear Portales de Negocio Inteligentes.
Clever Path Portal es el lugar de trabajo de e-business personalizado para empleados, socios y clientes (B2E, B2B Y B2C), permitiendo la interacción con todos los requerimientos de información.
CEM
Log Analyzer for Oracle

Microstrategy

Productos para Business Intelligence:

Web Reporting MicroStrategy Web™ Interfaz basada en HTML para acceso vía WEB, a través de un Browser, con acceso interactivo y de consultas no planeadas.

Desktop Reporting MicroStrategy Agent™ es un ambiente cliente-servidor con capacidades de análisis funcional integrando minería de datos.

Information Delivery MicroStrategy Narrowcast Server™ permite crear formatos y enviar mensajes personalizados desde la plataforma Microstrategy 7 a otras Fuentes de información.

Transactions MicroStrategy Transactor™ proporciona capacidades transaccionales a través de web, acceso inalámbrico y voz.

Performance, Scalability & Security MicroStrategy Intelligence Server™ . Es el centro de la plataforma Microstrategy. Construido para soportar la escalabilidad y tolerancia a fallas que requiere el análisis de terabytes de información.

Design & Construction MicroStrategy Architect™ es el componente en el que se modelan las aplicaciones de Microstrategy 7.

Administration MicroStrategy Administrator™ permite generar un ambiente de desarrollo, implantación y mantenimiento de aplicaciones de business intelligence a gran escala.

Development MicroStrategy SDK™ Conjunto de herramientas para el desarrollo de aplicaciones.

5.3. SQL Server 2000 y Analysis Services

SQL Server es un servidor de bases de datos desarrollado por Microsoft, esta versión recoge funciones interesantes para el desarrollo de Data warehouses y minería de datos, que se irán describiendo.

"La función Servicios OLAP de SQL Server versión 7.0 se llama ahora Analysis Services de SQL Server 2000. Se ha sustituido el término Servicios OLAP por el término Analysis Services. Analysis Services incluye también un nuevo componente de minería de datos."[54]

5.3.1. Características

CUBOS. Una de las características es que permite crear e incrementar cubos de datos en tiempo real, utilizando OLAP relacional (ROLAP). Permitiendo la construcción de estos de manera dinámica.

ALMACENAMIENTO DISTRIBUIDO DE DATOS. Un cubo de dato puede ser almacenado a través de varios analysis servers usando la nueva característica de particiones distribuidas.

ACCIONES. Son conjuntos de operaciones que han sido definidas de antemano, y permiten realizar tareas externas como recibir o enviar datos a alguna aplicación. Supongamos que al analizar un cubo de ventas un analista se percata que un producto está teniendo ventas bajas en alguna ciudad inmediatamente podría invocar la acción de DESCUENTO con lo cual los precios se modificarían en los sistemas de Ventas y Mercadotecnia.

[54] SETH, Paul et al, *"Preparing and Mining Data with Microsoft SQL Server 2000 and Analysis Services"*, www.microsoft.com , Microsoft Online Books.

ACCESO DESDE APLICACIONES DE CLIENTE. La nueva versión de SQL Server también incluye novedades para los clientes que navegan cubos desde aplicaciones de cliente. Dependiendo de la aplicación los clientes pueden trabajar con conjuntos de resultados desde la fuente de datos. Los diseñadores OLAP pueden ahora ocultar cubos enteros o algunas medidas al usuario final.

DIMENSIONES. Se han hecho muchos avances en el área de las dimensiones también. Se incluyen varios tipos de dimensiones tales como dimensiones desiguales o ragged, las cuales permiten al pariente lógico de un item residir fuera del siguiente nivel dentro de la jerarquía. Por ejemplo, una dimensión ragged permitiría crear una línea directa de empleados para ser administrada por el director de la empresa sin tener que pasar por los mandos medios. Analysis Services además introduce la dimensión padre/hijo la cual usa dos columnas de la tabla de dimensiones para crear relaciones entre sus miembros, por ejemplo identificar el pago de un empleado y el pago de la persona sobre de él dentro de la jerarquía de la empresa.

El tipo de almacenamiento ROLAP permite dimensiones más grandes. Además es posible modificar alguna dimensión sin necesidad de procesar todo el cubo.

SEGURIDAD. Incluye gran flexibilidad para acceder a los cubos. El acceso puede ser controlado desde los niveles más bajos hasta las dimensiones del mismo cubo. Además de la autenticación tradicional, los clientes pueden conectarse y usar los analysis servers via HTTP o HTTPS.

HERRAMIENTAS DE DESARROLLO. SQL Server 2000 hace la vida más fácil al integrar un número de productos que eran adquiridos aparte para la versión 7.0. Además incluye un generador de sentencias MDX (sentencias multidimensionales).

MINERIA DE DATOS. SQL Server 2000 es la primera versión del producto que soporta minería de datos. Aunque este componente de Analysis Services está en la infancia se continúa desarrollando para dar un paso mayor. No solo permite generar modelos de minería de datos utilizando algoritmos de clustering o árboles de decisión, si no que además provee de soporte para OLE DB, lo que permite que una tercera parte los proveedores de algoritmos de Data Mining integren sus productos con Analysis Services.

5.3.2. Arquitectura de los Analysis Services

Los servicios de análisis de Microsoft SQL Server 2000 están conformados por un conjunto de componentes, a través del servidor y los clientes, que se describirán a continuación.

ANALYSIS SERVER

Analysis Server es el corazón de los Análisis Services, figura 20, éste procesa los cubos y envía los resultados a los clientes. Al igual que cualquier servicio dentro de SQL Server, el Análisis Server puede ser pausado, iniciado o terminado desde el panel de control.

Cada Análisis Server tiene un depósito el cual contiene las definiciones de los objetos pertenecientes al servidor.

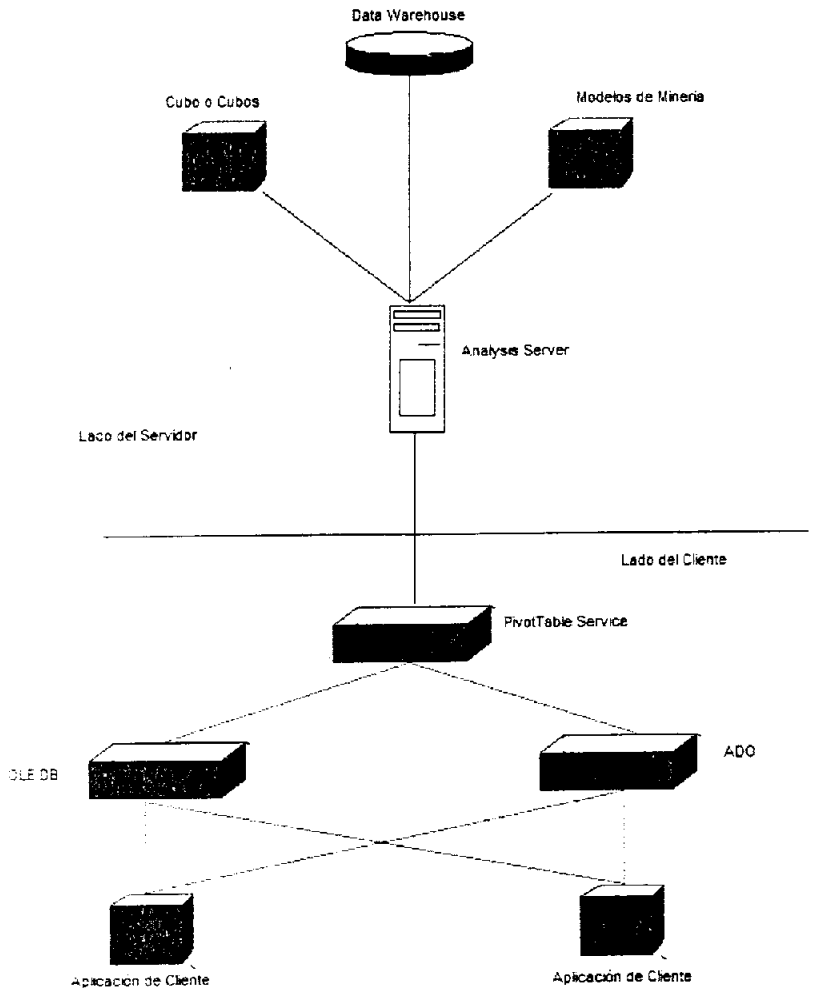


Figura 20. Estructura de Analysis Services

ANALYSIS MANAGER

El Analysis Manager es una herramienta que permite el acceso al Analysis Server. Es una consola de administración, con una interfase de usuario intuitiva y de fácil uso.

Analysis Manager presenta un nodo para cada Analysis Server y un subnodo para cada base de datos.

Contiene cinco carpetas bajo cada base de datos, figura 21, estas son:

Data Sources: Esta carpeta contiene información relacionada a la conexión del servidor, seguridad y proveedor OLE DB.

Cubes: Esta carpeta contiene todos los cubos que pertenecen a la base de datos.

Shared Dimensions: Contiene las dimensiones que comparten más de un cubo. Una dimensión común podría ser el tiempo.

Mining Models: Guarda los modelos de minado de datos basados en la información de los cubos.

Database Roles: Almacena cuentas de seguridad para permitir o restringir el acceso a los usuarios que intentan usar componentes del Analysis Services tales como cubos y modelos de minería.

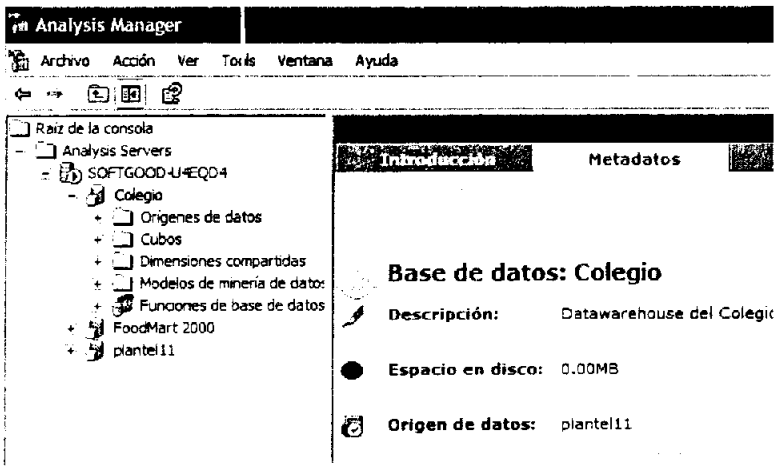


Figura 21. Pantalla de Analysis Manager

CUBES.

La información dentro de los modelos de OLAP es vista conceptualmente como cubos. Los cubos son estructuras multidimencionales que contienen conjuntos organizados de datos de un almacén de datos o data warehouse. Cada cubo consiste de ***dimensiones*** que son esencialmente categorías descriptivas, tales como el tiempo o el área geográfica, y ***miembros***, que son valores numéricos como unidades de venta y población.

Analysis Services permite incorporar a los cubos: celdas calculadas, actualización de un cubo en tiempo real, utilizar el cubo al mismo tiempo que se realizan operaciones de agregación al mismo, ocultar elementos del cubo, funciones de navegación como enrollar y desenrollar.

MINING MODELS

Los modelos de minería de datos son el corazón del concepto de Minería de Datos y son estructuras virtuales que representan grupos de datos para análisis de predicciones. Los modelos están formados por dos partes: columnas y algoritmo de minado. Analysis Services soporta dos algoritmos descritos con anterioridad *Arboles de Decisión* y *Clustering*.

PIVOT TABLE SERVICE

Es la interface principal entre los Analysis Services y las aplicaciones de cliente. Básicamente es un proveedor OLE DB y puede ser usado en muchas plataformas de desarrollo como Visual Basic o Visual C++.

DECISION SUPPORT OBJECTS

Es una librería de clases e interfaces que provee acceso a un Analysis Server. DSO determina los mecanismos básicos de almacenamiento y los elementos usados por los Analysis Services, permitiendo que puedan ser controlados y programados por cualquier lenguaje que implemente Microsoft COM. Dichos elementos están representados en una estructura jerárquica y son base de datos, data sources, dimensiones, cubos, modelos de minería y roles. El objeto Servidor se encuentra en la cima de esta jerarquía.

5.3.3. MINERIA DE DATOS EN SQL SERVER.

Recordemos un poco el objetivo de la Minería de Datos:

"La Minería de Datos es utilizada para desenterrar patrones en los datos. Estos patrones pueden ser obvios en retrospectiva, pero la mayoría de las veces podrían no ser notados sin las herramientas como las que ofrece Analysis Services."[55]

Dentro de Analysis Services se encuentran diversas herramientas para la creación de Cubos y consultas multidimensionales, de igual forma incluye herramientas para construir modelos de minería de datos.

Los modelos que Analisys Manager puede construir están basados en dos algoritmos Microsoft Decision Trees y Microsoft Clustering. Para ambos casos se puede realizar el análisis sobre Cubos o sobre Tablas Relacionales.

Aunque los Analysis Services solo incluyen estos algoritmos, su interfase permite que aplicaciones ajenas diseñadas para el análisis de datos puedan trabajar con las bases de datos y cubos creados.

Analysis Manager incluye un tutorial sobre el concepto de Minería de Datos y ejemplos de cómo funcionan sus diversas características, dentro de las cuales se encuentran:

- Creación, edición y eliminación de Cubos.
- Ayudantes para el diseño de Cubos y modelos de minería.
- Una interfase intuitiva y de fácil manejo.
- Conexión de aplicaciones ajenas.
- Permite la programación de sus componentes.

Los Analysis Services proveen de una herramienta de fácil operación al usuario novel en el campo del análisis de datos. Además de que al estar incluidos en SQL Server 2000 permiten de manera más cómoda que los operadores de bases de datos accedan a esta tecnología.

[55] SETH, Paul et al. Op cit. p. 485

Sin embargo no todas las versiones de SQL Server 2000 incluyen las herramientas para minado de datos. De las versiones Personal, Enterprice, y Estándar, solo las últimas dos cuentan con esta utilidad.

6. Caso Práctico.

Como se ha revisado a lo largo de este trabajo, la importancia de la información dentro de las diferentes organizaciones cobra cada vez mas fuerza como motor principal de cambio de las estrategias implementadas.

Dentro del ámbito académico la necesidad de entender cuales son los factores que más influyen para el buen logro de los objetivos educativos de los alumnos, como la aprobación y aprovechamiento, hace necesario utilizar las herramientas tecnológicas de análisis de datos, que brinda la informática, para visualizar de manera más clara ¿qué ocurre? y ¿qué acciones tomar para mejorar los indicadores?

6.1. Panorama actual.

El Colegio de Bachilleres es una institución con más de treinta años de funcionamiento. Fundado por decreto del C. Presidente Luis Echeverría Álvarez en el año 1973, cuenta actualmente con 20 planteles distribuidos en el área metropolitana, con una población estudiantil de aproximadamente 20,000 alumnos inscritos, una plantilla de cerca de 3,000 trabajadores administrativos y académicos.

"A 31 años de su creación, los avances y logros de la institución son importantes y reconocidos por diferentes instancias. En el último año, la SEP nos incorporó al Programa Nacional de Becas a la Excelencia Académica y el Aprovechamiento Escolar, otorgándonos 834 becas a partir de septiembre de 2004; y, además, nos solicitó incrementar nuestra matrícula de nuevo ingreso en 4 mil lugares en el ciclo 2004-2005 y autorizó los recursos para ampliar nuestra capacidad instalada en tres planteles. Por su parte, la Secretaría de Relaciones Exteriores, por conducto del Instituto de los Mexicanos en el Exterior, eligió a nuestro bachillerato para atender a los mexicanos que radican en Estados Unidos y Canadá; y el bachillerato en línea del Colegio es parte de la oferta del CONEVYT-INEA y está en sus plazas comunitarias de todo el país."[56]

[56] Colegio de Bachilleres, www.cbacilleres.edu.mx, consultada el 29 de mayo de 2005

Como se puede ver en su pagina web la misión del Colegio es la de "Formar ciudadanos con un proyecto de vida basado en competencias académicas y laborales y una vocación profesional definida, con alta autoestima y compromiso consigo mismos, su familia y la sociedad; mediante procesos educativos eficientes que, con libertad y calidad, propicien su inventiva, comprensión, creatividad y crítica; y con hábitos de trabajo y principios éticos que normen su conducta para su incorporación productiva a la sociedad y a la educación superior."

Su visión es la de "Ser una institución pública de calidad, moderna, flexible y orientada a la formación pertinente de sus estudiantes, que use las nuevas tecnologías para ampliar y diversificar las oportunidades de avance académico y egreso en sus modalidades escolar y abierta; que certifique las competencias laborales relacionadas con las capacitaciones impartidas; que utilice con eficiencia su infraestructura y que cuente con una planta de personal académico preparada y comprometida con su función; todo ello para que sus egresados sean reconocidos y aceptados en su grupo social, las instituciones de educación superior y en el campo de trabajo."

Dejando de lado la función social que las instituciones de educación cumplen dentro de un país, revisemos los datos dentro del marco del impacto económico que generan.

La inclusión de México dentro de la OCDE[57] le obliga a ser comparado constantemente con los demás países miembros de este organismo. Sin embargo los resultados no han sido muy alentadores. Según la representante en México del Fondo de las Naciones Unidas para la Infancia (Unicef), Yoriko Yasukawa

"... el país todavía no está a la altura de su imagen de novena economía mundial, ni responde a su categoría de pertenecer a la Organización para la Cooperación y el Desarrollo Económicos (OCDE)... es importante que las autoridades y la sociedad conocieran las bajas calificaciones que han otorgado los organismos internacionales en materia de educación, a fin de que se redoblen esfuerzos para mejorar la calidad de la enseñanza... si bien se registran avances en materia de educación y salud, en relación

[57] Organización para la Cooperación y el Desarrollo Económicos.

con los países del resto del América Latina, el gasto per cápita que destina a programas sociales es inferior al promedio latinoamericano, que se ubica en 500 pesos por persona... el gobierno mexicano requiere de más recursos para impulsar una política social más agresiva." [58]

Dentro de los resultados del examen de matemáticas y lectura que aplico este organismo a niños de entre 14 y 15 años de edad México obtuvo el último lugar.

El problema de la educación no solo en México sino en todo el mundo está ligado a infinidad de factores sociales, económicos, culturales. Y es deber de los encargados de su operación implementar las estrategias que conduzcan a lograr los objetivos planteados, aun y sobre las problemáticas del país.

La educación superior atraviesa en el país por una etapa difícil.

"El elevado porcentaje de reprobación en el bachillerato llegó a niveles que pueden "impedir que una persona de nuevo ingreso pueda tener una oportunidad" en el sistema educativo..."

"De acuerdo con estimaciones del costo de reprobación en México, realizadas a partir de los índices que registra el Panorama Educativo de México 2004, indicadores del Sistema Educativo Nacional –documento elaborado por el INEE y la SEP-, en 2004 más de 3 millones de estudiantes de primaria, secundaria y preparatoria reprobaron un grado escolar. Lo que se traduce en una inversión de más de 45 mil 770 millones de pesos, equivalente al 11.6 por ciento del presupuesto público para la educación del año pasado."

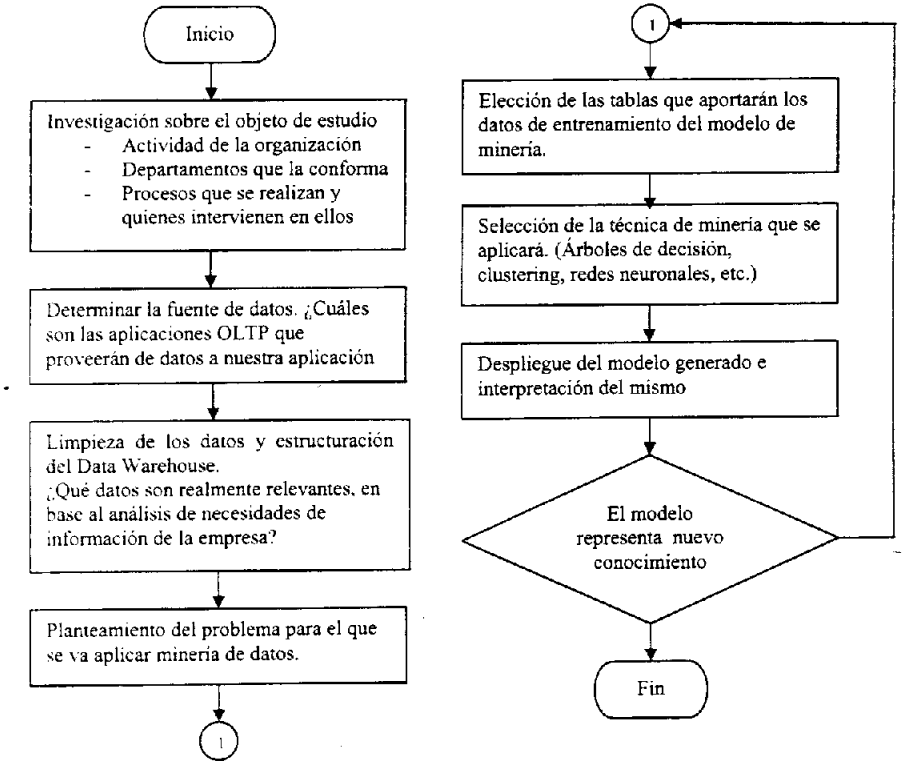
"...en el caso de la educación básica, a niveles de reprobación que son "normales" o aceptables a nivel internacional, esto porque se encuentran por debajo del 20 por ciento. Lo preocupante está en el bachillerato donde el índice de reprobación alcanza el

[58] http://www.economista.com.mx/online4.nsf/all/5F2F4217CFA9502806256FEF0073F3F9?OpenDocument, Notimex, publicado el 26 de abril de 2005

36 por ciento, sólo en una parte refleja que en los "niveles más bajos de la educación hay una formación cultural empobrecida".[59]

6.2. Plan de trabajo.

Los pasos que se siguieron para el desarrollo del proyecto se encuentran en el siguiente diagrama a bloques:



[59] http://www.sep.gob.mx/wb2/sep/sep_Resumen16Febrero2005, Resumen en la página de la SEP del artículo de MARTINEZ, Nuria. "Grave Reprobación en el Bachillerato". EL UNIVERSAL p. 8 y 20, publicado el 16 de febrero de 2005

6.3. ¿Qué se mide en el Colegio de Bachilleres?

Basado en el POA 2005[60] del Colegio Bachilleres los siguientes son los indicadores que se miden en esta institución.

- Egreso. Representa la cantidad de alumnos que concluyen su bachillerato cada semestre.

- Aprobación. El número de alumnos que aprueban cada materia.

- Permanencia. La cantidad de alumnos que inician al inicio del semestre y que llegan al final del mismo.

- Superación. El número de alumnos que ingresan a la institución y que terminan el primer semestre.

- Excelencia Académica. Alumnos que tienen un promedio superior al 9.0

Los supervisores académicos del Colegio de Bachilleres tienen el objetivo de incrementar en niveles aceptables el porcentaje de estas medidas.

Tener a la mano los datos para implementar acciones inmediatas, es una necesidad. Imaginemos el valor de un reporte que nos permita saber cual es la tendencia de aprobación de tal o cual grupo, y no solo eso si no poder conocer las características de cada profesor, el horario en que se imparte tal o cual materia y conocer el impacto real de cada atributo.

Además es tan dinámico el proceso de la educación y tantos los factores que lo afectan que se requiere de nuevas herramientas para apoyar las decisiones que toman los

[60] Programa Operativo Anual, por ley desde 1994, todas las instituciones de gobierno deben elaborar uno al iniciar el año y tratar de cumplir sus objetivos.

directores académicos y diseñar estrategias congruentes con las problemáticas detectadas.

El modelo educativo del Colegio de Bachilleres[61] , coloca al alumno como un individuo que construye su propio conocimiento en base a los estímulos que recibe del exterior y al profesor solo como un auxiliar o guía para optimizar este proceso. En este sentido es primordial conocer las características familiares, económicas y culturales de cada generación que reside en las aulas del Plantel.

De esta manera las herramientas que proporciona el análisis estadístico, incluyendo la minería de datos, pueden ayudar a lograr los objetivos planteados por la institución. De hecho existe un documento, generado por el Departamento de Análisis y Estadística en cada semestre, titulado "Estadística Básica"[62] en el que se concentran datos referentes a la población estudiantil, tales como: promedio de secundaria, sexo, número de alumnos por grupo, preferencia por el plantel y número de aciertos obtenidos en el examen de ingreso. Sin embargo no logra el detalle requerido para entender tal o cual comportamiento de un grupo o de una generación de alumnos, ya que no incluye información sobre hábitos de estudio y condiciones socioeconómicas.

6.4. Estructura Organizativa.

Aunque el objetivo principal del Colegio de Bachilleres es el académico; dentro de su constitución se encuentran una serie de departamentos que coordinan su funcionamiento. Estos podemos dividirlos en dos grupos los académicos y los administrativos. Dentro de los departamentos administrativos tenemos la Unidad Administrativa, que coordina a los subdepartamentos de Servicios, Personal, Caja y Apoyo Audiovisual. En los departamentos académicos tenemos a la Unidad de Servicios de Apoyo Académico, encargada de coordinar a los laboratorios de computo, biología, física, química y a la biblioteca, también se encuentran las Jefaturas de

[61] Basado en un enfoque constructivista.
[62] Este documento es generado semestralmente por: y es coordinado por:

Materia que coordinan el quehacer académico del Plantel en coordinación con la Subdirección del Plantel. A la cabeza de la organización en planteles se encuentra la Dirección.

La Unidad de Registro y Control Escolar merece una mención aparte ya que en ella se realizan trámites para los alumnos como emisión de constancias de estudio, certificados de terminación y credenciales y se lleva el registro de avance académico de los mismos, se capturan calificaciones y se dan las fechas que rigen el calendario escolar.

Bien de esta manera será la URCE la principal proveedora de información para desarrollar nuestro data warehouse, aunque no podemos omitir que un Almacén de Datos guardará en algún momento información de todos los departamentos de nuestra organización.

6.5. Cifras y datos estadísticos relevantes.

La elaboración de informes útiles es en la mayoría de los casos una tarea compleja y tediosa que en muchos momentos inhibe que los administradores escarben más allá de lo que pueden hacer. La emisión de información que sea realmente significativa puede volverse en determinadas ocasiones un arte ya que se requiere conocer cada una de las fuentes de donde será obtenida.

Según el documento interno de la Coordinación Norte del Colegio de Bachilleres titulado "Estadísticas 2005"[63], se observa la necesidad de "... instrumentar un documento semestral con información escolar oportuna que permita mejorar la administración de la matrícula y el rendimiento académico de los planteles."[64]

"La estrategia adoptada ha consistido en integrar periódicamente a los Jefes de las Unidades de Registro y Control Escolar de los planteles, en un equipo de trabajo que

[63] Coordinación Zona Norte del Colegio de Bachilleres, "Estadísticas 2005". Documento informativo de la situación de la matrícula escolar de los periodos 2003 A – 2005 A, consultado en el Plantel 11 "Nueva Atzacoalco"
[64] item, p. 1

reúne e integra información semestral de la matrícula escolar, así como de resultados académicos. Con base en este trabajo se fortalece el vínculo entre las diferentes áreas académicas con el propósito de contribuir para una mejor planeación y operación académica y administrativa de los planteles. Finalmente, esta estrategia ha dado lugar a reuniones semestrales de análisis y evaluación de la información contenida en estos documentos de **Estadística de Alumnos** en las que participan el Coordinador sectorial, los Directores, Subdirectores y Jefes de las URCES de la Coordinación, llegando a conclusiones y acuerdos en beneficio de los planteles".[65]

A continuación se mostrarán algunas cifras estadísticas interesantes correspondientes al Plantel 11 y el proceso implementado para obtenerlas.

- **Población Estudiantil.** El Plantel cuenta con una población estudiantil de 2,124 alumnos inscritos en los diferentes semestres. De estos 552 no adeuda ninguna asignatura y son conocidos como **alumnos regulares**. 922 son **alumnos irregulares** o que adeudan de una a tres asignaturas y 298 son **alumnos repetidores** o **recursadores** y adeudan más de 4 asignaturas.

Inscritos	Regulares	Irregulares	Recusadores
2,124	552	922	298

Distribución de la población estudiantil respecto a su situación académica.[66]

- **Alumnos Regulares con promedio de calificación de 8.0 a 10 (Excelencia Académica).** Actualmente el Plantel 11 tiene 265 alumnos de alto rendimiento académico, distribuidos en los siguientes semestres como se muestra en la tabla siguiente:

PRIMERO	SEGUNDO	TERCERO	CUARTO	QUINTO	SEXTO	TOTAL
83	34	53	16	47	32	265

Población de excelencia académica y su distribución por semestre.

[65] ibidem
[66] Coordinación Zona Norte del Colegio de Bachilleres, "Estadísticas 2005", Sección Plantel 11, p.1

-Alumnos irregulares que adeudan de 1 a 3 asignaturas. El total de esta población por ambos turnos es de 922 y se distribuyen en los diferentes semestres de la siguiente manera:

SEGUNDO	TERCERO	CUARTO	QUINTO	SEXTO	TOTAL
266	163	188	142	163	922

Distribución de los alumnos deudores en los diferentes semestres.

-Egreso. Es el principal indicador a medir ya que en el se ven reflejados todos los esfuerzos y estrategias implementadas durante una administración escolar. El egreso es la cantidad de alumnos que cada semestre concluyen sus estudios en el Colegio. Este parámetro tiene varias clases: egreso en curso normal, egreso pertinente, egreso con calidad y egreso en evaluación extraordinaria.

El egreso en curso normal es la cantidad de alumnos que egresan al concluir el semestre en curso. Se le llama egreso pertinente al número de estudiantes que concluyen su bachillerato en 6 semestres. Egreso con calidad es aquel que contempla a los alumnos que concluyen sus estudios con promedio mayor a 8.0 y finalmente el egreso en evaluación extraordinaria que contempla la población que no concluye en los 6 semestres respectivos.

Las cifras del último egreso, correspondiente al semestre 04B, se aprecian en el siguiente cuadro:

Periodo	Promedio	E. CN	ER	AE	PAAR	SEA	TOTAL
04B	6-6.9	9	9	3	2	1	24
	7-7.9	46	20	13	40	10	129
	8-8.9	21	3	2	3	1	30
	9-10	2	0	0	0	0	2
	Total	78	32	18	45	12	185

Distribucion de la poblacion estudiantil que egreso en el semestre 04B.

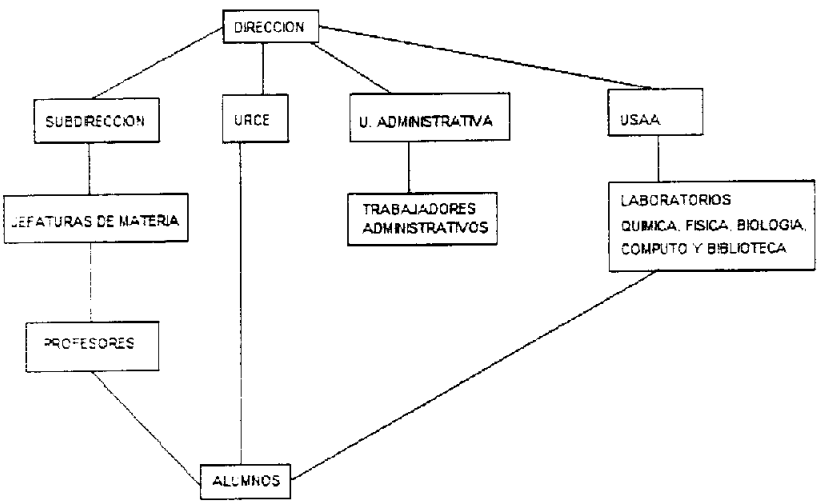
Aunque los cuadros anteriores nos dan una vista de las condiciones en que se encuentra el plantel y cual es la población en la que se debe poner más atención no nos muestran las características de dicha población; de esta forma no se sabe a ciencia cierta de que depende el éxito de uno u otro grupo de estudiantes. Si se pudiera escarbar más y conocerlas tal vez se pudiera pensar en estrategias mejor fundamentadas y ese es uno de los objetivos de esta investigación.

6.6. Estructura de la Unidad de Registro y Control Escolar.

Como parte de la metodología el primer paso es conocer los procesos y el departamento del cual tomaremos los datos.

La Unidad de Registro y Control Escolar trabaja con dos tipos de información. Por una parte el registro académico del plantel y por otro lado información correspondiente a los trámites administrativos de los alumnos.

ORGANIGRAMA DEL PLANTEL 11 "NUEVA ATZACOALCO"
DEL COLEGIO DE BACHILLERES



Se cuenta con el siguiente personal:

- 1 Responsable de Unidad
- 2 Auxiliares de la unidad (1 para cada turno)
- 4 Secretarias (2 por turno)

Los horarios en que funciona esta oficina son de Lunes a Viernes de 9:00 a 13:00 hrs. y de 16:00 a 19:00 hrs.

URCE realiza los siguientes procesos:

- Captura de calificaciones. En este proceso se realiza la captación de los resultados de las diferentes evaluaciones, Curso Normal, Evaluación de Recuperación, Acreditación Especial, Programa de Acreditación con Alto Rendimiento, SEACOBA.[67]
- Alta de Grupos. Proceso en el que se revisa que grupos y que materias han sido autorizados para abrirse, dependiendo entre otras cosas de la demanda por parte de los alumnos. El cupo de alumnos que se recomienda para un grupo es de 50 personas.
- Emisión de Actas de Evaluación. En esta actividad se imprimen las actas de evaluación de las diferentes materias que se imparten en el colegio.
- Solicitud de trámites. Se realizan diversos trámites como emisión de constancias, certificados parciales y totales, bajas temporales, bajas definitivas, historias académicas, renuncia de calificaciones, duplicados de certificado, credenciales.

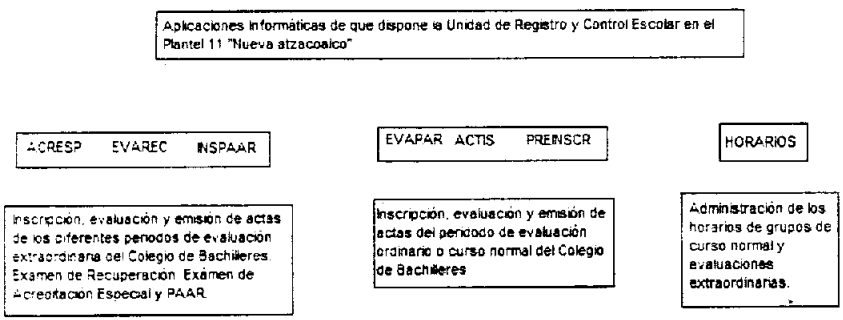
6.7. Estructura del Sistema de Información.

Las instituciones de educación necesitan calificar y cuantificar el comportamiento de la matrícula en los rubros de aprovechamiento, aprobación y ausentismo y es la URCE quien posee esta información.

[67] Con excepción del Curso Normal las demás son conocidas como evaluaciones extraordinarias.

Para la toma de decisiones y elaboración de estrategias orientadas a mejorar los rubros de aprobación, permanencia, egreso y superación se generan diversas estadísticas basadas en los datos contenidos en los sistemas de información siguientes:

EVAREC. Tiene como función la captura de calificaciones y la emisión de actas de los diferentes periodos de evaluaciones y extraordinarias.
EVAPAR. Sistema que captura calificaciones parciales, hasta 5 por semestre.[68] También emite actas finales.
ACRESP. Inscripción de alumnos a examen de acreditación especial.
HORARIOS. Auxiliar en la modificación y consulta de horarios de grupos.
ACTIS. Sistema para modificar y consultar la inscripción de los alumnos.
PREINSCR. Captura la inscripción de grupos de capacitación y materias optativas.
INSPAAR. Genera la inscripción al Programa de Acreditación de Alto Rendimiento (PAAR).[69]



El sistema de información del Colegio de Bachilleres se centra básicamente en captura de calificaciones y emisión de boletas y certificados, desde luego basado y enmarcado en el Reglamento de Inscripciones del mismo. En este sentido no existe hasta el momento una aplicación que ayude a los encargados de la toma de decisiones (Director, Subdirector y Jefes de Departamento) a analizar los datos con rapidez y de manera sencilla.

[68] Sin embargo solo se lleva acabo un evaluación parcial en este plantel.
[69] El PAAR es una evaluación extraordinaria, consiste básicamente en un curso intensivo de alguna materia y se ofrece segun la demanda de los alumnos.

Los informes estadísticos generados cuentan con un solo enfoque y son complejos de emitir, ya que deben ser elaborados manualmente. Imaginemos que a un Jefe de Materia le interesa conocer como se han comportado los grupos de determinado profesor a través de los últimos 3 años y saber o estimar cual será la tendencia de reprobación. Para esto tendría que revisar en su archivo si cuenta con las estadísticas de los últimos años, de lo contrario tendría que revisar las actas de los grupos y del profesor que está consultando, y sí además después de conocer esa tendencia le interesara descubrir cuales son las características de esos alumnos y averiguar cual de ellas afecta más en los resultados obtenidos, entonces la generación de ese reporte se convierte en una tarea muy difícil.

En la Unidad de Registro y Control Escolar se cuenta con la siguiente infraestructura de cómputo:

- Cableado de red con capacidad para 8 nodos
- Seis clientes de red con las siguientes caracteristicas:

    2 Computadoras con procesador Pentium 4 a 3 GHz
    248 MB en RAM
    Disco Duro 74.5 GB
    Sistema Operativo Windows XP

    2 Computadoras Pentium 4 a 2.79 GHz
    240 MB en RAM
    Disco Duro 37.2 GB
    Sistema Operativo Windows XP

    2 Computadoras Pentium 4 a 1.70 GHz
    256 MB en RAM
    Disco Duro 38.3 GB
    Sistema Operativo XP.

- Cinco impresoras, dos de ellas de impresión láser
- Servidor HT
Procesador Celeron
Sistema Operativo Novell 5.0

6.8. Construcción del Data Warehouse.

La búsqueda de información significativa para la construcción de un Data Warehouse lleva a revisar cuidadosamente cada una de las bases de datos de las diferentes aplicaciones con que se cuenta en determinada empresa o institución. Esta búsqueda dio como resultado la elección de las siguientes tablas para armar el correspondiente almacén de datos.

NEWINGXX. Esta base contiene información acerca de la condición socioeconómica de los alumnos de nuevo ingreso, así como datos sobre sus costumbres académicas como ¿cuantas horas lee a la semana? o por ejemplo ¿Cuál es la materia que menos se le dificulta?.

DIRALUM. Base de datos con el directorio histórico de alumnos donde se incluye el nombre, curp y matrícula.

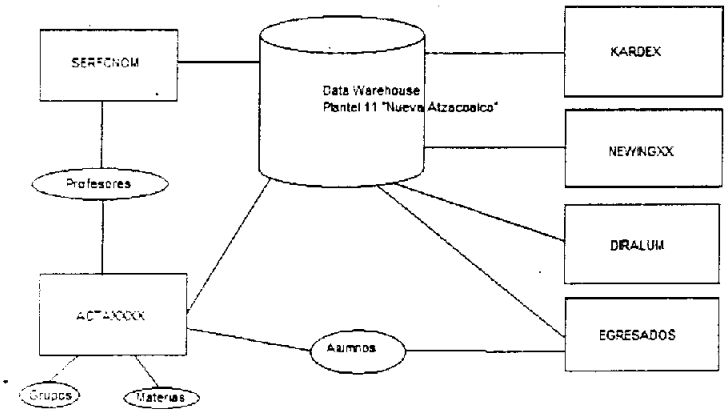
ACTAXXXX. Este archivo se genera cada semestre e incluye las evaluaciones de cada grupo/materia por cada proceso de evaluación (Curso Normal, Examen de Acreditación Especial, Evaluación de Recuperación, PAAR, SEACOBA). Incluye la relación grupo/materia/profesor.

SERFCNOM. Incluye información del personal administrativo y académico que labora en el Plantel.

KARDEX. Directorio de alumnos inscritos con la descripción de los grupos y las materias que están cursando en el semestre vigente. También incluye el horario de clases.

EGRESADOS. Tabla que incluye el nombre y la matrícula para cada alumno que ha egresado en los últimos 4 años.

CONJUNTO DE TABLAS INTEGRADAS AL DW DEL PLANTEL 11 Y COMO SE RELACIONAN CON LOS DIFERENTES ACTORES DENTRO DEL MISMO



6.8.1. Adaptación e integración de los datos encontrados.

Las bases de datos que la URCE maneja se encuentran en formato de DBASE IV, por lo que tuvieron que ser importadas a formato de Access. Por otra parte tablas como actaxxxx no se encontraban normalizadas y por lo tanto no se podían utilizar de manera natural para construir el almacén de datos.

Después de un proceso de normalización y de una elección minuciosa que tratara de cubrir las perspectivas de información que los directivos requieren se llego a la siguiente estructura para armar el data warehouse (la descripción detallada de las tablas, así como su transformación se encuentra en el Anexo 1 y 2):

Se fusionaron los archivos de acta existentes en un solo archivo , normalizado, que lleva por nombre ACTA. De igual forma los archivos newingxxxx se juntaron para formar uno solo llamado NEWING, eliminando los datos que no eran relevantes.

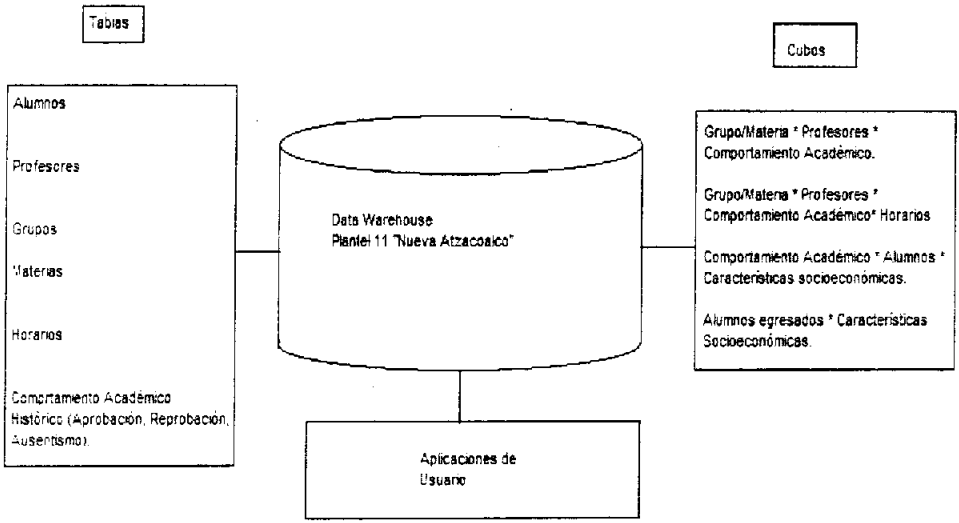
Se genero una tabla llamada ESTADISTICA, que contiene el comportamiento académico histórico de cada grupo / materia en los últimos 4 años por su estructura se considero como tabla de hechos. Más adelante se verá su utilización dentro del almacén de datos.

La tabla DIRALUM, SERFCNOM y EGRESADOS no sufrieron cambios, aunque en el caso de EGRESADOS se debieron completar los registros incompletos correspondientes a las matrículas de los alumnos.

6.8.2. Estructura General del Data Warehouse propuesto.

El Data Warehouse se construyo en base a los datos disponibles en la URCE y se refieren a los resultados académicos obtenidos a través de los últimos cinco semestres escolares. La propuesta de nuestro almacén de datos es cubrir las necesidades de información del área académica para que a través del acceso eficiente a información significativa puedan tomar decisiones que mejoren los resultados obtenidos en los rubros de aprobación, ausentismo y egreso.

La estructura propuesta es la siguiente:



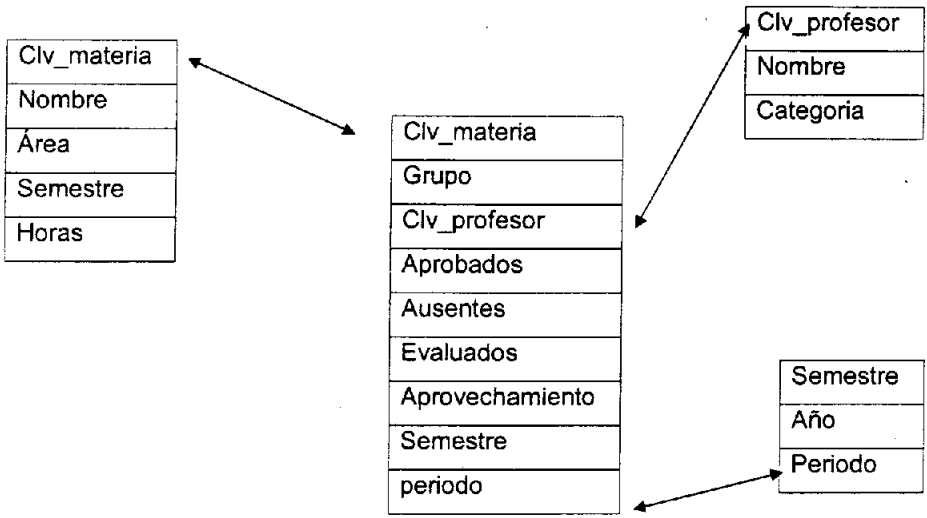
Las tablas propuestas se normalizaron hasta la tercera forma (3FN) para permitir la creación de los diferentes cubos de información los cuales son la estructura central del data warehouse.

Debemos recordar que aunque en el desarrollo del almacén de datos se manejen estructuras definidas siempre será posible desarrollar nuevas estructura que dependerán del punto de vista del analista o del personal directivo.

6.8.3. Desarrollo de los cubos de información que integran el Data Warehouse.

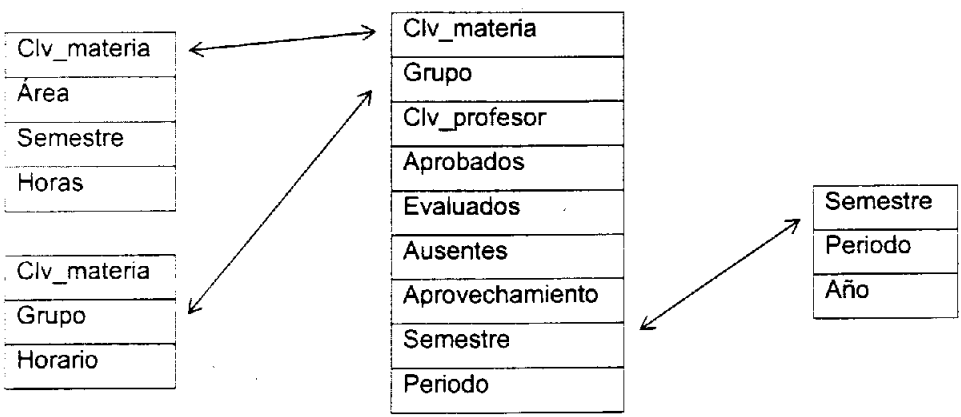
El primer cubo de información incluye las relaciones entre grupo/materia, profesor y comportamiento académico.

Su estructura se observa en la siguiente figura:

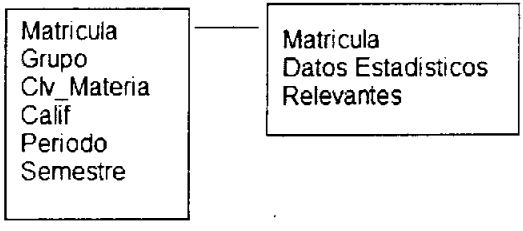


La utilidad principal de esta estructura es la de proporcionar información inmediata al jefe de materia a fin de revisar el comportamiento académico de los grupos de un profesor determinado, también permite observar el comportamiento histórico que estos grupos han tenido a lo largo de los últimos semestres.

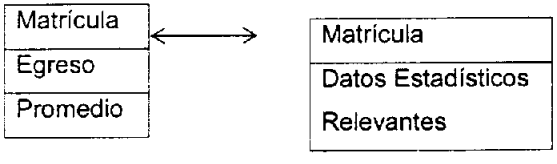
El cubo dos refleja la relación entre el comportamiento académico y la franja horaria en que se imparten las diferentes materias. Su estructura es la siguiente:



El cubo tres muestra las calificaciones obtenidas por cada uno de los alumnos y con que condiciones socioeconómicas ingresaron al plantel. Lo anterior es de utilidad cuando se tienen casos de alumnos que generan problemas es fácil encontrar que aprovechamiento llevan y conocer que hábitos y condiciones sociales tienen; sin embargo es importante señalar que los datos que se tienen no abarcan a toda la población que se encuentra inscrita.



Por último el cubo cuatro guarda la relación entre los alumnos que han egresado en las últimas tres generaciones y los datos socioeconómicos con los que ingresaron a la institución.



6.9. Minando los datos.

Concluido el diseño y estructuración del Almacén de Datos, se procedió a realizar el experimento de minado con los datos del egreso de las generaciones que comprenden los semestres **2001 B – 2004 A, 2002 A – 2004 B, 2002 B – 2005 A.**

El ejercicio de minería planteado se desarrollo siguiendo los seis pasos propuestos en el Online Book *"Preparing and Mining Data with Microsoft SQL Server 2000 and*

Analysis Services", para una solución de minería de datos, los cuales son los siguientes:

- Planteamiento del Problema
- Preparar los datos
- Construcción del modelo
- Validación de los modelos
- Despliegue e interpretación de los modelos
- Mantenimiento de los meta datos asociados al modelo[70], para mejorar el mismo.

6.9.1. Planteamiento del problema.

Dada una nueva generación de alumnos predecir el índice de egreso de la misma tomando en cuenta las características económicas, sociales, académicas y físicas, y los resultados de generaciones anteriores.

6.9.2. Objetivo.

Descubrir los factores que tienen más impacto en el egreso de una generación de alumnos, tomando en cuenta cuatro aspectos: social, económico, hábitos académicos y características físicas. Para lo cual se utilizará el Data Warehouse desarrollado para el Plantel y técnicas de minería de datos.

6.9.3. Preparando los datos.

Nuestro ejercicio de minería de datos utiliza las tablas NEWING01.DBF, NEWING02.DBF, las cuales contienen información sobre las características socio - económicas y académicas de las generaciones de los semestres 2001B – 2004 A, 2002 A – 2004 B y 2002B – 2005 A, y la tabla EGRESADOS.DBF.

[70] Seth, Paul et al. *Preparing and Mining Data with SQL Server 2000 and Analysis Services,* Microsoft SQL Series Online Books, 2003. La referencia de este documento se encuentra en el sitio web de Microsoft *www.microsoft.com*

El primer paso realizado fue el de acomodar los datos existentes en los dos archivos NEWING; ya que no tenían la misma estructura y en el caso del campo VIVE_CON del archivo NEWING01 éste se encontraba dividido en varios campos dentro del archivo NEWING02. Posteriormente se clasificaron los campos dependiendo de la información que contenían en económicos, sociales, académicos y físicos.

La estructura de las tablas NEWING01 y NEWING02 así como la tabla NEWING resultante se encuentran en el anexo 2.

6.9.4. Construcción del modelo de minería.

Una vez que los datos fueron seleccionados y limpiados se utilizo el Analysis Manager para analizar estos.

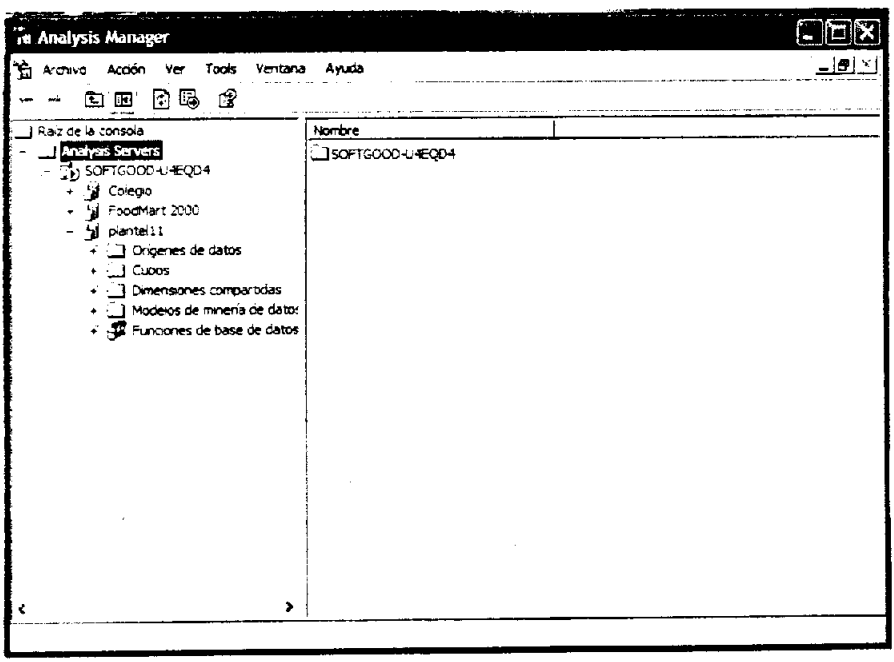


Figura 22. Vista del Analysis Manager.

El proceso que se siguió para crear el modelo fue el siguiente:

1.- Seleccionar la base de datos plantel11

2.- En la opción de modelos de minería de datos dar un clic izquierdo y seleccionar nuevo modelo de minería, figura 23.

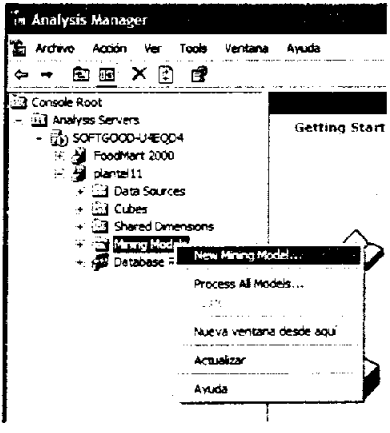


Figura 23. Seleccionar la opción Nuevo Modelo de Datos.

3.- Después de que aparece la pantalla del Mining Model Wizard nos aparece una ventana solicitando el tipo de base de datos, entonces seleccionamos la opción **Relational data**. Figura 24.

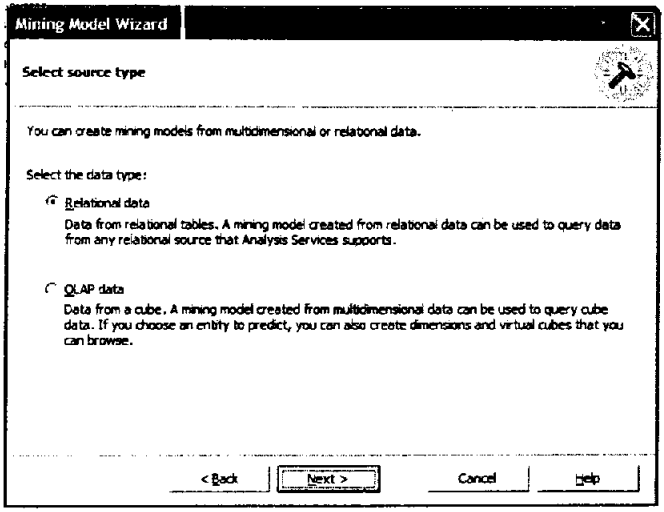


Figura 24. Seleccionar tipo de base de datos.

4.- A continuación seleccionamos la o las tablas donde se efectuará el análisis. En el caso de nuestra aplicación seleccionamos las tablas NEWING y EGRESO. Figura 25.

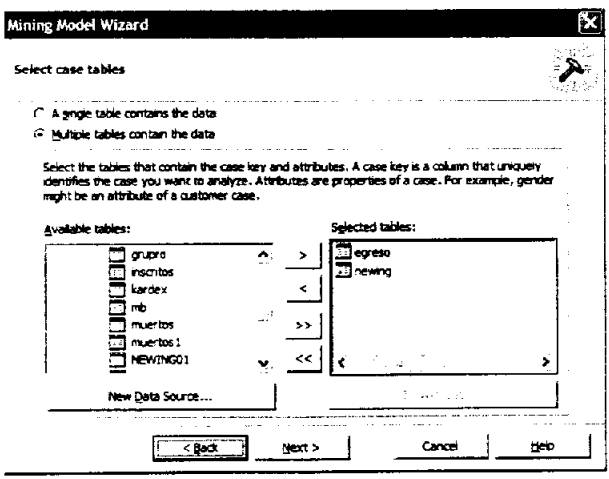


Figura.25. Seleccionar las tablas para crear el modelo.

5.- El siguiente paso es seleccionar la técnica con la que se creará el modelo. Analysis Manager proporciona dos técnicas Microsoft Decision Trees y Microsoft Clustering. Seleccionamos Microsoft Decisión Trees.

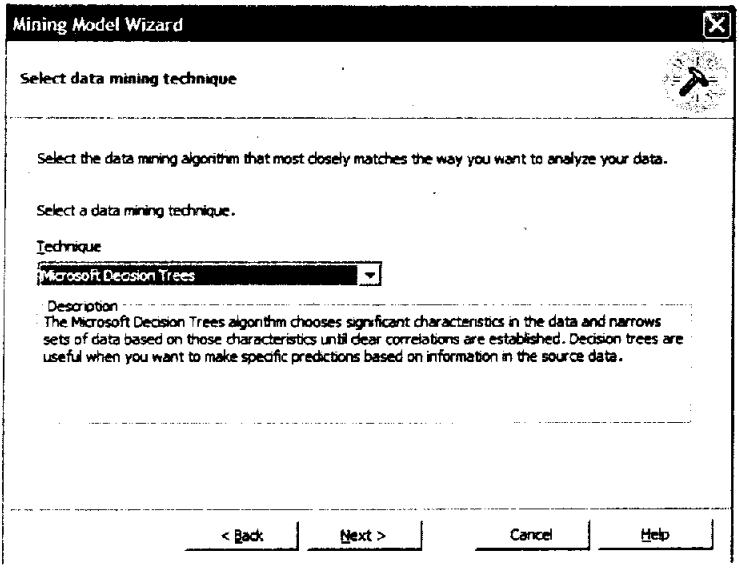


Figura 26. Elegir la técnica de minería.

6.- Después de elegir el algoritmo de minería, se despliega una ventana donde se solicita seleccionar la tabla y el campo llave, donde se encuentran los casos que se desean analizar; en este ejemplo los datos de los alumnos que se encuentran en la tabla EGRESO. Figura 27.

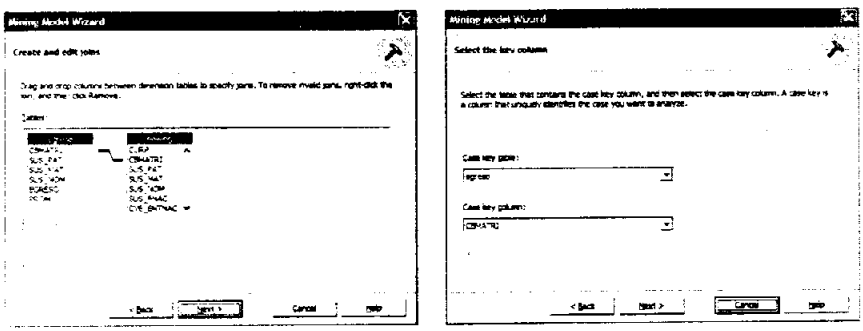


Figura 27. Seleccionar la tabla que contiene los casos a analizar.

7.- Finalmente seleccionamos las columnas de entrada y las de predicción. Una columna de entrada contiene los datos sobre los cuales se quiere basar el análisis. Una columna de predicción contiene las predicciones que el modelo de minería hace basado

en las columnas de entrada. Por conveniencia las columnas de entrada son consideradas de predicción también. Figura 28.

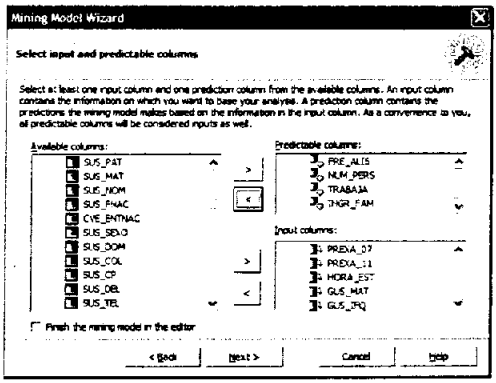


Figura 28. Seleccionar los campos de predicción y de entrada.

8.- Para terminar se le da nombre al modelo y se procesa.

Se construyo otro modelo utilizando la técnica de Microsoft Clustering, el procedimiento es el mismo que el utilizado anteriormente, y el resultado se puede observar en la figura 29.

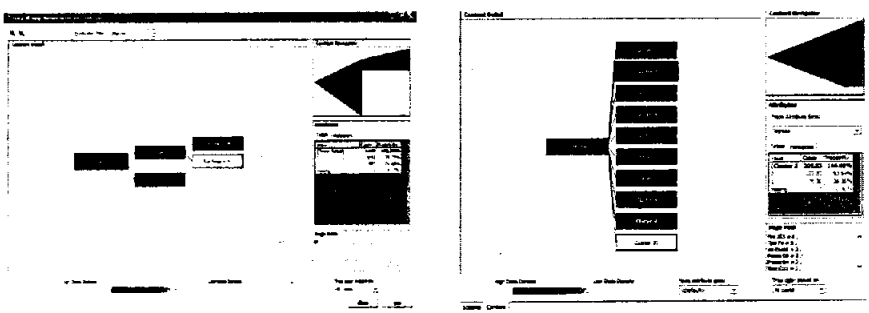


Figura 29. Vista del árbol de decisión y de los clusters generados por el Analysis Manager.

6.9.5. Despliegue e interpretación del modelo.

El modelo generado por la aplicación contiene un conjunto de estructuras de árbol, cada una correspondiente a cada columna que se va a predecir. Para el caso de la columna de egreso se obtuvo una estructura de árbol de tres niveles de profundidad donde el factor que determina que un alumno egrese sería el promedio con que ingresa al bachillerato o lo que es lo mismo el promedio con que concluyo la secundaria. Figura 30.

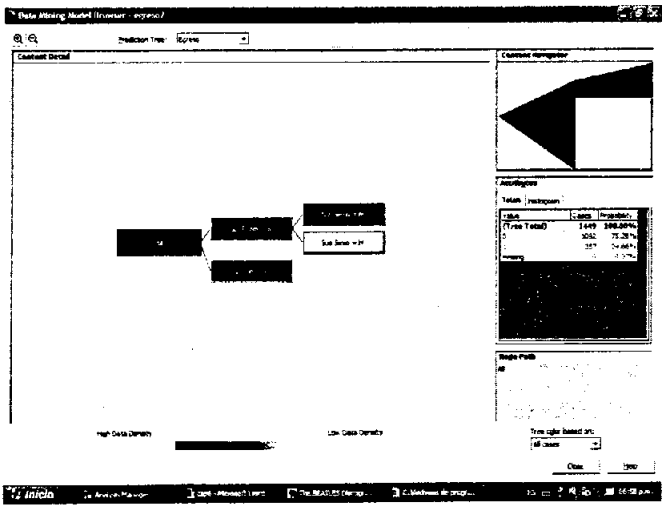


Figura 30. Árbol de predicción generado para la columna de egreso.

En el segundo nivel se observa que el porcentaje de egreso es mayor en el caso de las mujeres.

Podríamos concluir que en efecto el hecho de tener un buen promedio indica en la mayoría de los casos buenos hábitos de estudio y esto incide directamente en la conclusión oportuna del bachillerato.

Para el segundo factor podemos sacar varias conjeturas, primero el hecho de que las mujeres presentan conductas de mayor compromiso con las tareas que emprenden y por lo tanto tienden a obtener mejores promedios, aunque esté demostrado que ambos sexos cuentan con las mismas capacidades intelectuales, otra conclusión podría se que

los hombres en esta etapa se enrolan en trabajos y actividades que les distraen del quehacer académico y finalmente este hecho podría ser meramente una situación demográfica donde la mayoría son mujeres y por lo tanto el número de egresados de este sexo es mayor.

Será tarea del personal directivo diseñar estrategias para reforzar conocimientos de los jóvenes que presentan promedios menores al 8.0, de manera que se lleve un seguimiento del comportamiento académico de los mismos y lograr su egreso. También es necesario implementar un mecanismo que permita obtener datos estadísticos más frescos para cada generación de egreso ya que del momento en el que se obtienen los datos al momento en que un estudiante egresa hay una diferencia de tres años. Esto podria afectar el impacto que puede tener el análisis en un momento dado.

Dentro de las tareas que se están llevando a cabo seria muy recomendable reforzar la estrategia de tutorías en primer semestre de tal manera que los alumnos que ingresan aprendan técnicas y métodos de estudio que les ayuden a mejorar su desempeño académico.

7. Resultados y Conclusiones.

Como resultado de éste trabajo se logro la incorporación de un módulo de análisis estadístico en el nuevo programa informático de la URCE, llamado SADIE, el cual está ya operando en los veinte planteles del Colegio de Bachilleres del área metropolitana.

Este módulo tiene como finalidad facilitar la entrega de resultados académicos a los encargados y responsables del área, para que éstos tomen decisiones oportunas que reditúen en el mejoramiento de la calidad educativa.

Por otro lado se está impulsando el desarrollo de trabajos de minería de datos en las áreas administrativas a fin de conocer el comportamiento en los demás departamentos que conforman el Colegio.

Se está vislumbrando cada vez más la necesidad de tener información significativa, se están haciendo de lado los reportes que solo contienen datos aislados y se requiere que los nuevos informes contengan conocimiento de lo que ocurre en dos o tres áreas diferentes. Como un ejemplo imaginemos un informe sobre un profesor en particular donde se incluyan entre otras cosas la cantidad de horas cursos tomados en los últimos semestres, los resultados académicos de cada uno de sus grupos, los salones donde imparte las clases y los horarios, así como también si ha tenido algún incidente de tipo laboral en fecha reciente. Un reporte así requiere no solo de información académica sino de áreas administrativas también.

Las necesidades y rezagos en materia económica en la educación pública hacen necesaria la inclusión de técnicas de análisis que aprovechen de mejor manera los recursos asignados. En este sentido se debe diseñar un plan para preparar nuevos "trabajadores del conocimiento", de forma que las decisiones tomadas estén cada vez más fundamentadas.

Sin embargo aunque el camino es muy prometedor no debemos olvidar que la Minería de Datos es un campo que falta todavía por explorar y explotar, que requiere de personal capacitado y especializado en el área de bases de datos e inteligencia artificial y que en algunos casos los proyectos pueden ser demasiado caros en tiempo y dinero y no representar ningún provecho tangible.

De hecho el principal obstáculo a vencer es el de encontrar gente especializada en el área de bases de datos. Tal vez la creación de un programa de entrenamiento a los administradores solvente esta carencia de especialistas.

Además un proyecto de estas características debe contar también con un registro minucioso de las acciones que se tomarán en base a la información entregada a fin de corregir las acciones si las tendencias en los índices analizados presentan el mismo comportamiento.

Un trabajo de minería de datos que no es retroalimentado de las nuevas experiencias que generaron ciertas decisiones, está condenado al fracaso o al olvido ya que el objetivo principal es el de permitir diseñar estrategias que permitan alcanzar las metas de la organización. Por lo tanto es vital documentar y registrar que ocurre al implementar o suspender tal o cual acción.

En este sentido los resultados encontrados en el experimento de minado sobre el egreso y las condiciones en que entra un alumno al colegio no fueron, por así decirlo, espectaculares y podrían poner en duda la utilidad del mismo y el gasto de tiempo y esfuerzo por parte del personal. Sin embargo debe de considerarse como el principio de la incorporación de este tipo de aplicaciones a la dinámica del Colegio y como una demostración de que se pueden romper paradigmas donde se pensaba que el gasto excesivo impide aplicar nuevas tecnologías. A partir de este momento la dinámica de la Institución permitirá el mantenimiento de esta aplicación.

Recordemos que la parte investigada es una pequeña porción de las múltiples combinaciones de datos que se pueden realizar, y falta la incorporación de información de otras áreas administrativas a fin de completar el almacén de datos.

Aunque la minería de datos es un área de investigación relativamente vieja, es hasta estos momentos cuando las computadoras cuentan con los recursos suficientes para hacerla una realidad.

Como consecuencia la incorporación y desarrollo de aplicaciones de inteligencia artificial en el sector público es cada vez mayor. Como ejemplo se tiene el desarrollo de un sistema de inteligencia artificial, coordinado por el doctor Enrique Cáceres Nieto del Instituto de Investigaciones Jurídicas de la UNAM, para apoyar a jueces en la toma de decisiones, al resolver problemas que típicamente se encuentran en el razonamiento judicial práctico.[71]

Algo que llama la atención es la cantidad de información que se encuentra en Internet con palabras claves como Búsqueda de Conocimiento en Base de Datos, Gestión del Conocimiento, I Bussines, etc., todas seudónimos para Minería de Datos. Dentro de esta búsqueda de información, sin embargo, no se encontraron comunidades mexicanas especializadas en minería de datos a diferencia de países como España, Argentina y Estados Unidos.

Una asignatura pendiente es la creación de esta comunidad donde se puedan intercambiar experiencias sobre proyectos de minería y su éxito. En este sentido falta mucho trabajo por desarrollar en el área, enfrentando las problemáticas comunes como el llenado automático de datos incompletos, la granularidad, la aplicación de lógica difusa y redes neuronales, algo que ha dado en llamarse Neuro-Fuzzy y que tiene ya un impacto importante en este tiempo; no olvidar los productos electrodomésticos que presentan en la etiqueta Fuzzy Logic.

[71] Desarrollan sistema de inteligencia artificial para apoyar a jueces. La Jornada 1 de septiembre de 2005, p.2a

Se considera que la investigación fue exitosa ya que no represento ningún gasto adicional para la institución, salvo el tiempo de algunas personas y por el contrario aporta un beneficio elevado al proporcionar información con cruces complejos de datos.

Un proyecto de minería de datos podría estar justificado incluso para una mediana empresa donde el volumen de información es lo suficientemente basto (con tablas de más de mil registros) para garantizar resultados confiables sin necesidad de hacer un gran gasto. En el caso de pequeñas empresas no es muy recomendable realizar este tipo de proyectos por que el gasto podría no estar tan justificado, amen de que el pequeño volumen de datos nos llevaría a sacar conclusiones no muy verdaderas.

La manera de pensar de los gerentes y encargados de la toma de decisiones en una empresa deberá empezar a dar un giro. Desde ahora los directivos verán la organización como una entidad dinámica que se nutre de la información recabada por sus sistemas a lo largo de los años y la convierte en experiencia y aprendizaje.

Aunque los beneficios a corto plazo se traducen en entrega de información oportuna no es correcto crear faltas expectativas respecto a un proyecto de minería. El hecho de fortalecer y respaldar la toma de decisiones no significa que se tomará la más acertada, en este sentido no se debe olvidar que a fin de cuentas el éxito de una empresa dependerá del buen tino que se tenga a la hora de calcular riesgos, costos y utilidades.

Cabe señalar que el proceso de diseñar un proyecto de minería es en muchos aspectos un arte para quien lo elabora y que mientras más experimentado sea el diseñador los resultados obtenidos serán mejores, aunque el hecho de seguir una metodología permite el buen logro de las actividades.

7. Resultados y Conclusiones.

Como resultado de éste trabajo se logro la incorporación de un módulo de análisis estadístico en el nuevo programa informático de la URCE, llamado SADIE, el cual está ya operando en los veinte planteles del Colegio de Bachilleres del área metropolitana.

Este módulo tiene como finalidad facilitar la entrega de resultados académicos a los encargados y responsables del área, para que éstos tomen decisiones oportunas que reditúen en el mejoramiento de la calidad educativa.

Por otro lado se está impulsando el desarrollo de trabajos de minería de datos en las áreas administrativas a fin de conocer el comportamiento en los demás departamentos que conforman el Colegio.

Se está vislumbrando cada vez más la necesidad de tener información significativa, se están haciendo de lado los reportes que solo contienen datos aislados y se requiere que los nuevos informes contengan conocimiento de lo que ocurre en dos o tres áreas diferentes. Como un ejemplo imaginemos un informe sobre un profesor en particular donde se incluyan entre otras cosas la cantidad de horas cursos tomados en los últimos semestres, los resultados académicos de cada uno de sus grupos, los salones donde imparte las clases y los horarios, así como también si ha tenido algún incidente de tipo laboral en fecha reciente. Un reporte así requiere no solo de información académica sino de áreas administrativas también.

Las necesidades y rezagos en materia económica en la educación pública hacen necesaria la inclusión de técnicas de análisis que aprovechen de mejor manera los recursos asignados. En este sentido se debe diseñar un plan para preparar nuevos "trabajadores del conocimiento", de forma que las decisiones tomadas estén cada vez más fundamentadas.

Sin embargo aunque el camino es muy prometedor no debemos olvidar que la Minería de Datos es un campo que falta todavía por explorar y explotar, que requiere de personal capacitado y especializado en el área de bases de datos e inteligencia artificial y que en algunos casos los proyectos pueden ser demasiado caros en tiempo y dinero y no representar ningún provecho tangible.

De hecho el principal obstáculo a vencer es el de encontrar gente especializada en el área de bases de datos. Tal vez la creación de un programa de entrenamiento a los administradores solvente esta carencia de especialistas.

Además un proyecto de estas características debe contar también con un registro minucioso de las acciones que se tomarán en base a la información entregada a fin de corregir las acciones si las tendencias en los índices analizados presentan el mismo comportamiento.

Un trabajo de minería de datos que no es retroalimentado de las nuevas experiencias que generaron ciertas decisiones, está condenado al fracaso o al olvido ya que el objetivo principal es el de permitir diseñar estrategias que permitan alcanzar las metas de la organización. Por lo tanto es vital documentar y registrar que ocurre al implementar o suspender tal o cual acción.

En este sentido los resultados encontrados en el experimento de minado sobre el egreso y las condiciones en que entra un alumno al colegio no fueron, por así decirlo, espectaculares y podrían poner en duda la utilidad del mismo y el gasto de tiempo y esfuerzo por parte del personal. Sin embargo debe de considerarse como el principio de la incorporación de este tipo de aplicaciones a la dinámica del Colegio y como una demostración de que se pueden romper paradigmas donde se pensaba que el gasto excesivo impide aplicar nuevas tecnologías. A partir de este momento la dinámica de la Institución permitirá el mantenimiento de esta aplicación.

Recordemos que la parte investigada es una pequeña porción de las múltiples combinaciones de datos que se pueden realizar, y falta la incorporación de información de otras áreas administrativas a fin de completar el almacén de datos.

Aunque la minería de datos es un área de investigación relativamente vieja, es hasta estos momentos cuando las computadoras cuentan con los recursos suficientes para hacerla una realidad.

Como consecuencia la incorporación y desarrollo de aplicaciones de inteligencia artificial en el sector público es cada vez mayor. Como ejemplo se tiene el desarrollo de un sistema de inteligencia artificial, coordinado por el doctor Enrique Cáceres Nieto del Instituto de Investigaciones Jurídicas de la UNAM, para apoyar a jueces en la toma de decisiones, al resolver problemas que típicamente se encuentran en el razonamiento judicial práctico.[71]

Algo que llama la atención es la cantidad de información que se encuentra en Internet con palabras claves como Búsqueda de Conocimiento en Base de Datos, Gestión del Conocimiento, I Bussines, etc., todas seudónimos para Minería de Datos. Dentro de esta búsqueda de información, sin embargo, no se encontraron comunidades mexicanas especializadas en minería de datos a diferencia de países como España, Argentina y Estados Unidos.

Una asignatura pendiente es la creación de esta comunidad donde se puedan intercambiar experiencias sobre proyectos de minería y su éxito. En este sentido falta mucho trabajo por desarrollar en el área, enfrentando las problemáticas comunes como el llenado automático de datos incompletos, la granularidad, la aplicación de lógica difusa y redes neuronales, algo que ha dado en llamarse Neuro-Fuzzy y que tiene ya un impacto importante en este tiempo; no olvidar los productos electrodomésticos que presentan en la etiqueta Fuzzy Logic.

[71] Desarrollan sistema de inteligencia artificial para apoyar a jueces. La Jornada 1 de septiembre de 2005. p.2a

Se considera que la investigación fue exitosa ya que no represento ningún gasto adicional para la institución, salvo el tiempo de algunas personas y por el contrario aporta un beneficio elevado al proporcionar información con cruces complejos de datos.

Un proyecto de minería de datos podría estar justificado incluso para una mediana empresa donde el volumen de información es lo suficientemente basto (con tablas de más de mil registros) para garantizar resultados confiables sin necesidad de hacer un gran gasto. En el caso de pequeñas empresas no es muy recomendable realizar este tipo de proyectos por que el gasto podría no estar tan justificado, amen de que el pequeño volumen de datos nos llevaría a sacar conclusiones no muy verdaderas.

La manera de pensar de los gerentes y encargados de la toma de decisiones en una empresa deberá empezar a dar un giro. Desde ahora los directivos verán la organización como una entidad dinámica que se nutre de la información recabada por sus sistemas a lo largo de los años y la convierte en experiencia y aprendizaje.

Aunque los beneficios a corto plazo se traducen en entrega de información oportuna no es correcto crear faltas expectativas respecto a un proyecto de minería. El hecho de fortalecer y respaldar la toma de decisiones no significa que se tomará la más acertada, en este sentido no se debe olvidar que a fin de cuentas el éxito de una empresa dependerá del buen tino que se tenga a la hora de calcular riesgos, costos y utilidades.

Cabe señalar que el proceso de diseñar un proyecto de minería es en muchos aspectos un arte para quien lo elabora y que mientras más experimentado sea el diseñador los resultados obtenidos serán mejores, aunque el hecho de seguir una metodología permite el buen logro de las actividades.

ANEXO 1.

DESCRIPCIÓN DETALLADA DE LAS TABLAS UTILIZADAS EN LA CONSTRUCCIÓN DEL DATA WAREHOUSE DE LA UNIDAD DE CONTROL ESCOLAR DEL PLANTEL 11 "NUEVA ATZACOALCO" DEL COLEGIO DE BACHILLERES.

a) ACTAXXXX.DBF. Contiene el acta de calificaciones de los diferentes grupos y signaturas para cada semestre. XXXX hace referencia a los 200A, 200B, 201A, 201B, etc.

NOMBRE_CAMPO	TIPO	LONGITUD	DESCRIPCION
FOLIO	Texto	10	FOLIO DEL ACTA
GRUPO	Texto	3	GRUPO EL QUE PERTENECE EL ACTA
ASIGNATURA	Texto	3	ASIGNATURA A LA QUE PERTENECE EL ACTA
NALUM	Doble	8	NUMERO DE ALUMNOS DEL GRUPO
NACTA	Doble	8	NUMERO DE HOJA DEL ACTA
TOLACTAS	Doble	8	NUMERO DE TOTAL DE HOJAS DEL ACTA
MAT1	Texto	9	CAMPOS REFERENTES A LA MATRICULA Y CALIFICACIÓN DE CADA UNO DE LOS ALUMNOS QUE PERTENECEN A ESTE GRUPO, EL NUMERO DE ALUMNOS ES DE 39 POR CADA HOJA DEL ACTA
CAL1	Texto	2	
MAT39	Texto	9	
CAL39	Texto	2	
TIPOACTA	Texto	1	TIPO DEL ACTA NORMAL O ADICIONAL
TIPOEXAM	Texto	2	CORRESPONDE AL PERIODO DE EVALUACIÓN
IMP	Si/No	1	MARCA SI EL ACTA YA FUE IMPRESA
FECHIMPRE	Texto	8	FECHA DE IMPRESIÓN DEL ACTA
PROFESOR	Texto	7	CLAVE DEL PROFESOR QUE EVALUA

b) KARDEX.DBF. Guarda la información de los alumnos inscritos en el semestre actual, la cantidad de materias que cursa los grupos en donde se encuentra, el horario y el salón.

NOMBRE_CAMPO	TIPO	LONGITUD	DESCRIPCION
PTL	Texto	2	NUMERO DE PLANTEL
MATRICULA	Texto	9	MATRICULA DEL ALUMNO
MATRIC8	Texto	8	MATRICULA ANTIGUA DEL ALUMNO
NOMBRE	Texto	45	NOMBRE DEL ALUMNO, INCLUYE APELLIDOS
CAP	Texto	2	CAPACITACION QUE CURSA EL ALUMNO
TNO	Texto	1	TURNO DONDE ESTA INSCRITO
PLT_PROC	Texto	2	PLANTEL DE PROCEDENCIA EN EL CASO DE CAMBIO DE PLANTEL
CMOVE	Doble	8	TIPO DE MOVIMIENTO
PROME	Doble	8	PROMEDIO DEL ALUMNO
CONTMAT	Texto	2	NUMERO DE MATERIAS EN LAS QUE ESTA INSCRITO
MAT1	Texto	3	ASIGNATURA 1
GPO1	Texto	3	GRUPO 1
G_BAND1	Texto	1	HORARIO
LUN1	Texto	4	
MAR1	Texto	4	
MIE1	Texto	4	
JUE1	Texto	4	
VIE1	Texto	4	HORARIO PARA LA MATERIA Y GRUPO
SALON1	Texto	4	SALON
MAT10	Texto	3	
GPO10	Texto	3	
G_BAND10	Texto	1	
LUN10	Texto	4	
MAR10	Texto	4	
MIE10	Texto	4	
JUE10	Texto	4	
VIE10	Texto	4	EL TOTAL DE MATERIAS EN LAS QUE UN ALUMNO PUEDE ESTAR
SALON10	Texto	4	INSCRITO EN UN SEMESTRE ES DE DIEZ
TIP_MOV	Texto	3	MOVIMIENTO 1
TIP_MOV2	Texto	3	MOVIMIENTO 2
TIP_MOV3	Texto	3	MOVIMIENTO 3
ALTA	Sí/No	1	CONFIRMACIÓN DE LA INSCRIPCIÓN
C	Sí/No	1	SIN DESCRIPCIÓN
FP	Sí/No	1	SIN DESCRIPCIÓN
CURP	Texto	18	CURP DEL ALUMNO

c) DIRALUM.DBF. Directorio de los nombres y matrículas de todos los alumnos que han estado inscritos en el Plantel desde 1978.

NOMBRE_CAMPO	TIPO	LONGITUD	DESCRIPCION
MATRICULA	Texto	9	MATRICULA DEL ALUMNO
MATRIC8	Texto	8	MATRICULA ANTIGUA
NOMBRE	Texto	45	NOMBRE DEL ALUMNO INCLUYENDO APELLIDOS
NOM	Texto	35	NOMBRE(S) DEL ALUMNO
PATERNO	Texto	30	APELLIDO PATERNO
MATERNO	Texto	30	APELLIDO MATERNO
PLANTEL	Texto	2	PLANTEL DONDE ESTA INSCRITO
PLTORIG	Texto	2	PLANTEL DONDE SE INSCRIBIO ORIGINALMETE
CAPACITA	Texto	2	CAPACITACION QUE CURSA
TURNO	Texto	1	TURNO DONDE ESTA IINSCRITO
CURP	Texto	16	CURP DEL ALUMNO

d) EGRESO. Tabla generada durante el proceso de análisis, donde se indica la matricula y el nombre de los alumnos y si egresó y con qué promedio.

NOMBRE_CAMPO	TIPO	LONGITUD	DESCRIPCION
CBMATRI	Texto	255	MATRICULA DEL ALUMNO
SUS_PAT	Texto	255	APELLIDO PATERNO
SUS_MAT	Texto	255	APELLIDO MATERNO
SUS_NOM	Texto	255	NOMBRE
EGRESO	Entero largo	4	INDICA SI EL ALUMNO EGRESO O NO (1 PARA EGRESADO, 0 PARA NO EGRESADO)
PROM	Entero largo	4	PROMEDIO AL MOMENTO DE EGRESAR

e) SERFCNOM.DBF. Contiene el directorio de profesores del Plantel donde se incluye su matrícula y su nombre.

NOMBRE_CAMPO	TIPO	LONGITUD	DESCRIPCION
CVE_EMPL	Texto	7	CLAVE DEL PROFESOR
PATERNO	Texto	30	APELLIDO PATERNO DEL PROFESOR
MATERNO	Texto	30	APELLIDO MATERNO
NOMBRE	Texto	30	NOMBRE
NIP	Texto	8	CLAVE DE ACCESO PARA PODER CAPTURAR CALIFICACIONES

f) ACTA. Incluye los datos que contienen los archivos de ACTAXXXX, pero de manera normalizada para su posterior análisis de manera que genera un archivo electrónico histórico de las calificaciones obtenidas por un alumno en distintos periodos de tiempo.

NOMBRE_CAMPO	TIPO	LONGITUD	DESCRIPCION
matricula	Texto	50	MATRICULA DEL ALUMNO
materia	Texto	50	MATERIA QUE CURSO
grupo	Texto	50	GRUPO EN QUE CURSO
calif	Texto	50	CALIFICACION OBTENIDA
semestre	Texto	50	SEMESTRE
periodo	Texto	50	PERIODO DE EVALUACIÓN

g) ESTADISTICA. Tabla creada para la aplicación contiene la información académica de cada uno de los grupos y asignaturas que se imparten en el Plantel por semestre y periodo de evaluación. Es el resultado del análisis del archivo de ACTA descrito anteriormente. Aunque se ha realizado desde tiempo atrás un análisis estadístico de los grupos es hasta este momento que se puede realizar de manera automática. El módulo de análisis automático se incorporo al programa informático de la URCE (SADIE), que se utiliza en los veinte planteles.

NOMBRE_CAMPO	TIPO	LONGITUD	DESCRIPCION
grupo	Texto	4	GRUPO
materia	Texto	4	ASIGNATURA
totalum	Entero largo	4	NÚMERO DE ALUMNOS POR GRUPO
aprob	Entero largo	4	NUMERO DE ALUMNOS APROBADOS
rep	Entero largo	4	NUMERO DE ALUMNOS REPROBADOS
ausentes	Entero largo	4	NUMERO DE ALUMNOS NO EVALUADOS
seis	Entero largo	4	ALUMNOS APROBADOS CON SEIS
siete	Entero largo	4	CON SIETE
ocho	Entero largo	4	CON OCHO
nueve	Entero largo	4	CON NUEVE
diez	Entero largo	4	CON DIEZ
aprovecha	Doble	8	PROCENTAJE DE APROVECHAMIENTO
apropor	Doble	8	PORCENTAJE DE APROBACION
auspor	Doble	8	PORCENTAJE DE REPROBACION
periodo	Texto	3	PERIODO DE EVALUACIÓN
semestre	Texto	6	SEMESTRE

ANEXO 2.

ESTRUCTURA DETALLADA DE LAS TABLAS NEWING01, NEWING02 Y NEWING UTILIZADAS EN EL EJERCICIO DE MINERIA.

METROPOLITANO 2001
Descripción del Archivo Metro2001.dbf

Núm.	Campo	Tipo	Longitud	Descripción	Fuente
1	REGISTRO	Carácter	10	Institución en la que se registro el sustentante	1
2	EXAMINO	Carácter	10	Institución en la que aplicó el sustentante	1
3	CURP	Carácter	16	Clave única de registro poblacional	1
4	TURN_APL	Carácter	1	Turno de aplicación del examen	1
5	CVE_SEDE	Carácter	4	Clave de la sede de aplicación del examen	1
6	CVE_INSAPL	Carácter	1	Clave de la institución aplicadora del examen	1
7	CVE_CCT	Carácter	10	Clave de la escuela secundaria del sustentante	1
8	REGI_SEC	Carácter	3	Regimen de la secundaria de origen PUB = PUBLICA, PRI = PRIVADA	3
9	MODA_SEC	Carácter	1	Modalidad de la secundaria de origen G=General, T=Técnica, W=Trabajadores, V=Telesecundaria, A=Abierta	3
10	MUN_CCT	Carácter	3	Clave del municipio donde se localiza la escuela de procedencia del sustentante	1
11	FOLIO	Carácter	9	Folio del sustentante	1
12	CVE_GPO	Carácter	4	Clave del grupo donde presenta el examen	1
13	SUS_PAT	Carácter	30	Apellido paterno del sustentante	1
14	SUS_MAT	Carácter	30	Apellido materno del sustentante	1
15	SUS_NOM	Carácter	40	Nombre(s) del sustentante	1
16	SUS_FNAC	Fecha	8	Fecha de nacimiento del sustentante	1
17	CVE_ENTNAC	Carácter	2	Clave de la entidad de nacimiento del sustentante	1
18	SUS_SEXO	Carácter	1	Sexo del sustentante (H = Hombre, M = Mujer)	1
19	SUS_DOM	Carácter	40	Domicilio del sustentante, calle y número	1
20	SUS_COL	Carácter	30	Colonia del domicilio del sustentante	1
21	SUS_CP	Carácter	5	Código postal del domicilio del sustentante	1
22	SUS_DEL	Carácter	30	Municipio o delegación del domicilio del sustentante	1
23	SUS_TEL	Carácter	13	Teléfono del sustentante	1
24	SUS_PREG	Carácter	7	Folio del preregistro	1
25	SUS_PROM	Numérico	5.2	Promedio general del tercero de secundaria del sustentante	1
26	SUS_CATG	Carácter	1	Categoría del sustentante egresado, Foraneo o INEA	1
27	FOL_CERT	Carácter	10	Folio del certificado del sustentante	1
28	ANO_CERT	Carácter	4	Año del certificado del sustentante	1
29	SUS_TURN	Carácter	1	Preferencia de turno vespertino (X=turno vespertino)	1
30	OPC_ED01	Carácter	6	Opción educativa solicitada número 1	1
31	OPC_ED02	Carácter	6	Opción educativa solicitada número 2	1
32	OPC_ED03	Carácter	6	Opción educativa solicitada número 3	1
33	OPC_ED04	Carácter	6	Opción educativa solicitada número 4	1
34	OPC_ED05	Carácter	6	Opción educativa solicitada número 5	1
35	OPC_ED06	Carácter	6	Opción educativa solicitada número 6	1
36	OPC_ED07	Carácter	6	Opción educativa solicitada número 7	1
37	OPC_ED08	Carácter	6	Opción educativa solicitada número 8	1
38	OPC_ED09	Carácter	6	Opción educativa solicitada número 9	1
39	OPC_ED10	Carácter	6	Opción educativa solicitada número 10	1

METROPOLITANO 2001 Descripción del Archivo Metro2001.dbf					
Núm.	Campo	Tipo	Longitud	Descripción	Fuente
40	OPC_ED11	Carácter	6	Opción educativa solicitada número 11	1
41	OPC_ED12	Carácter	6	Opción educativa solicitada número 12	1
42	OPC_ED13	Carácter	6	Opción educativa solicitada número 13	1
43	OPC_ED14	Carácter	6	Opción educativa solicitada número 14	1
44	OPC_ED15	Carácter	6	Opción educativa solicitada número 15	1
45	OPC_ED16	Carácter	6	Opción educativa solicitada número 16	1
46	OPC_ED17	Carácter	6	Opción educativa solicitada número 17	1
47	OPC_ED18	Carácter	6	Opción educativa solicitada número 18	1
48	OPC_ED19	Carácter	6	Opción educativa solicitada número 19	1
49	OPC_ED20	Carácter	6	Opción educativa solicitada número 20	1
50	PRE_EXA	Carácter	1	Presentó examen	3
51	PREXA_01	Carácter	1	Al iniciar, identifico lo que necesito estudiar y hago un plan de trabajo	2
52	PREXA_02	Carácter	1	Estudio principalmente con mis apuntes de clase	2
53	PREXA_03	Carácter	1	Utilizo las monografías que venden en las papelerías	2
54	PREXA_04	Carácter	1	Estudio principalmente con el libro de texto de la materia	2
55	PREXA_05	Carácter	1	Utilizo enciclopedias, diccionarios y atlas	2
56	PREXA_06	Carácter	1	Realizo resúmenes y/o cuadros sinópticos	2
57	PREXA_07	Carácter	1	Estudio principalmente con los apuntes de mis compañeros	2
58	PREXA_08	Carácter	1	Resuelvo ejercicios para reafirmar lo estudiado	2
59	PREXA_09	Carácter	1	Solicito apoyo a mis padres o hermanos	2
60	PREXA_10	Carácter	1	Solicito asesoría a mis maestros	2
61	PREXA_11	Carácter	1	Estudio en equipo con mis compañeros de clase	2
62	HORA_EST	Carácter	2	A la Semana, Cuantas horas dedicas al estudio fuera del horario escolar	2
63	EXT_SEC	Carácter	1	Presentaste algún examen extraordinario en la secundaria	2
64	REP_SEC	Carácter	1	Repetiste algún año escolar en la secundaria	2
65	GUS_MAT	Carácter	1	Gusto por las matemáticas	2
66	GUS_IFQ	Carácter	1	Gusto por la introducción a la física y a la química	2
67	GUS_FIS	Carácter	1	Gusto por la física	2
68	GUS_QUI	Carácter	1	Gusto por la química	2
69	GUS_BIO	Carácter	1	Gusto por la biología	2
70	GUS_ESP	Carácter	1	Gusto por el español	2
71	GUS_HIS	Carácter	1	Gusto por la historia	2
72	GUS_GEO	Carácter	1	Gusto por la geografía	2
73	GUS_CIV	Carácter	1	Gusto por el civismo	2
74	GUS_LEX	Carácter	1	Gusto por la lengua extranjera	2
75	GUS_ART	Carácter	1	Gusto por la expresión y apreciación artística	2
76	GUS_EFI	Carácter	1	Gusto por la educación física	2
77	GUS_EDT	Carácter	1	Gusto por la educación tecnológica	2
78	GUS_OED	Carácter	1	Gusto por la orientación educativa	2

METROPOLITANO 2001
Descripción del Archivo Metro2001.dbf

Núm.	Campo	Tipo	Longitud	Descripción	Fuente
79	GUS_OPT	Carácter	1	Gusto por la asignatura optativa	2
80	PRE_MAT	Carácter	1	Preparación en matemáticas	2
81	PRE_IFQ	Carácter	1	Preparación en introducción a la física y a la química	2
82	PRE_FIS	Carácter	1	Preparación en física	2
83	PRE_QUI	Carácter	1	Preparación en química	2
84	PRE_BIO	Carácter	1	Preparación en biología	2
85	PRE_ESP	Carácter	1	Preparación en español	2
86	PRE_HIS	Carácter	1	Preparación en historia	2
87	PRE_GEO	Carácter	1	Preparación en geografía	2
88	PRE_CIV	Carácter	1	Preparación en civismo	2
89	PRE_LEX	Carácter	1	Preparación en lengua extranjera	2
90	PRE_ART	Carácter	1	Preparación en expresión y apreciación artística	2
91	PRE_EFI	Carácter	1	Preparación en educación física	2
92	PRE_EDT	Carácter	1	Preparación en educación tecnológica	2
93	PRE_OED	Carácter	1	Preparación en orientación educativa	2
94	PRE_OPT	Carácter	1	Preparación en las asignaturas optativas	2
95	SEC_EST	Carácter	1	Tienes la intensión de seguir estudios superiores después de la educación media superior	2
96	INS_INGR	Carácter	2	A qué institución de educación superior te gustaría ingresar al terminar tus estudios de educación media superio	2
97	ACT_REA	Carácter	1	Cuál de las siguientes actividades te gusta realizar más	2
98	TPO_LEC	Carácter	2	Cuántas horas de tu tiempo libre inviertes en leer	2
99	TIP_LEC	Carácter	2	Qué tipo de lectura te gusta más	2
100	TPO_TV	Carácter	2	Cuántas horas de tu tiempo libre inviertes en ver televisión	2
101	PROG_TV	Carácter	2	Qué tipo de programas prefieres ver en televisión	2
102	ACT_MT01	Carácter	1	Asiste regularmente a clase	2
103	ACT_MT02	Carácter	1	Se dedica la mayor parte del tiempo de la clase a trabajar con los alumnos	2
104	ACT_MT03	Carácter	1	Califica injustamente a sus alumnos	2
105	ACT_MT04	Carácter	1	Se esfuerzan para que todos los alumnos comprendan lo tratado en clase	2
106	ACT_MT05	Carácter	1	Ayudan a los alumnos en el desarrollo de sus trabajos en clase	2
107	ACT_MT06	Carácter	1	Dan suficientes ejemplos sobre los temas tratados en clase	2
108	ACT_MT07	Carácter	1	Realizan evaluaciones regularmente	2
109	ACT_MT08	Carácter	1	Castigan injustificadamente a los alumnos	2
110	ACT_MT09	Carácter	1	Informan a los alumnos sobre el resultado de sus evaluaciones, señalando sus progresos o fallas de aprendizaje	2
111	ACT_MT10	Carácter	1	Promueven el trabajo en equipo entre los alumnos	2
112	ACT_MT11	Carácter	1	Toman en cuenta la opinión e intereses de los alumnos para organizar la clase	2
113	CAL_FOR	Carácter	1	Califica la calidad de la formación que en general recibiste en tu secundaria	2
114	CUA_HNO	Carácter	2	Cuántos hermanos tienes	2
115	LUG_OCU	Carácter	2	Qué lugar ocupas entre tus hermanos	2
116	EDAD_MAD	Carácter	1	Edad de la madre	2
117	EDAD_PAD	Carácter	1	Edad del padre	2

METROPOLITANO 2001 Descripción del Archivo Metro2001.dbf					
Núm.	Campo	Tipo	Longitud	Descripción	Fuente
118	VIVE_CON	Carácter	2	Con quién vives actualmente	2
119	AS_ESC01	Carácter	1	Me ayudan en mis tareas escolares	2
120	AS_ESC02	Carácter	1	Platican conmigo sobre mis avances o fallas en la escuela	2
121	AS_ESC03	Carácter	1	Me orientan sobre lecturas o actividades que pueden ampliar mi aprendizaje escolar	2
122	AS_ESC04	Carácter	1	Me exigen un determinado tiempo de estudio a la semana	2
123	AS_ESC05	Carácter	1	Se informan con mis maestros sobre mi avance en la escuela	2
124	AS_ESC06	Carácter	1	Me premian o me felicitan cuando me va bien en la escuela	2
125	AS_ESC07	Carácter	1	Me castigan cuando me va mal en la escuela	2
126	ESC_MAD	Carácter	2	Escolaridad de la madre	2
127	ESC_PAD	Carácter	2	Escolaridad del padre	2
128	OCU_MAD	Carácter	2	Ocupación de la madre	2
129	OCU_PAD	Carácter	2	Ocupación del padre	2
130	NUM_FOC	Carácter	2	Cuántos focos tiene tu casa incluyendo lámparas	2
131	FRE_ALI1	Carácter	1	Con que frecuencia comes carne de res, cerdo, pollo o pescado	2
132	FRE_ALI2	Carácter	1	Con que frecuencia comes huevos	2
133	FRE_ALI3	Carácter	1	Con que frecuencia comes leche y derivados	2
134	FRE_ALI4	Carácter	1	Con que frecuencia comes frutas y verduras frescas	2
135	FRE_ALI5	Carácter	1	Con que frecuencia comes frijol, arroz, lentejas, habas, etc.	2
136	FRE_ALI6	Carácter	1	Con que frecuencia comes pan y pastas	2
137	NUM_PERS	Carácter	2	Número de personas que viven en tu casa	2
138	TRABAJA	Carácter	1	Desarrollas algún trabajo por el cual recibes un sueldo	2
139	INGR_FAM	Carácter	2	Cuál es el ingreso familiar mensual	2
140	PREPARA	Carácter	1	Qué también preparado te sientes para presentar con buen éxito tu examen de ingreso a nivel medio superior	2
141	FOR_LLEN	Carácter	1	Cómo llenaste esta hoja	2
142	LUG_LLEN	Carácter	1	Dónde llenaste esta hoja	2
143	CAL_ORIE	Carácter	1	Califica el apoyo que has recibido por parte de tu orientador en este concurso	2
144	EXAMEN	Carácter	2	Identifica poblaciones PRESENTADOS y NO PRESENTADOS del CENEVAL y UNAM	3
145	NGLOBAL	Numérico	3	Número de aciertos totales obtenidos en el examen	5
146	NHV	Numérico	3	Número de aciertos obtenidos en habilidad verbal (24 preguntas totales)	5
147	NESP	Numérico	3	Número de aciertos obtenidos en español (10 preguntas totales)	5
148	NHIS	Numérico	3	Número de aciertos obtenidos en historia (10 preguntas totales)	5
149	NGEO	Numérico	3	Número de aciertos obtenidos en geografía (10 preguntas totales)	5
150	NCIV	Numérico	3	Número de aciertos obtenidos en civismo (10 preguntas totales)	5
151	NHM	Numérico	3	Número de aciertos obtenidos en habilidad matemática (24 preguntas totales)	5
152	NMAT	Numérico	3	Número de aciertos obtenidos en matemáticas (10 preguntas totales)	5
153	NFIS	Numérico	3	Número de aciertos obtenidos en física (10 preguntas totales)	5
154	NQUI	Numérico	3	Número de aciertos obtenidos en química (10 preguntas totales)	5
155	NBIO	Numérico	3	Numero de aciertos obtenidos en biología (10 preguntas totales)	5
156	PNGLOBAL	Numérico	8.2	Porcentaje de aciertos obtenidos en el examen	5

METROPOLITANO 2001
Descripción del Archivo Metro2001.dbf

Núm.	Campo	Tipo	Longitud	Descripción	Fuente
157	PNHV	Numérico	8.2	Porcentaje de aciertos obtenidos en habilidad verbal	5
158	PNESP	Numérico	8.2	Porcentaje de aciertos obtenidos en español	5
159	PNHIS	Numérico	8.2	Porcentaje de aciertos obtenidos en historia	5
160	PNGEO	Numérico	8.2	Porcentaje de aciertos obtenidos en geografía	5
161	PNCIV	Numérico	8.2	Porcentaje de aciertos obtenidos en civismo	5
162	PNHM	Numérico	8.2	Porcentaje de aciertos obtenidos en habilidad matemática	5
163	PNMAT	Numérico	8.2	Porcentaje de aciertos obtenidos en matemáticas	5
164	PNFIS	Numérico	8.2	Porcentaje de aciertos obtenidos en física	5
165	PNQUI	Numérico	8.2	Porcentaje de aciertos obtenidos en química	5
166	PNBIO	Numérico	8.2	Porcentje de aciertos obtenidos en biología	5
167	EXPL_ASI	Carácter	5	Explicación después de haber corrido el proceso de asignación en el CENEVAL	6
168	NOPC_SOL	Carácter	3	Número de opciones solicitadas por el sustentante	6
169	NOPC_ASI	Numérico	2	Número de opción asignada	6
170	COPC_ASI	Carácter	6	Clave de la opción educativa asignada en el proceso de asignación en el CENEVAL	6
171	EXPL_MOD	Carácter	7	Explicación de modificaciones después de aclaraciones y asignaciones en los módulos CDO	6
172	ASIG_FIN	Carácter	6	Clave de opción asignada final después de aclaraciones y asignaciones en los módulos CDO	6
173	EXPL_FIN	Carácter	7	Explicación final de la asignación después de aclaraciones y asignaciones en los módulos CDO	6
174	NOPC_FIN	Numérico	2	Número de opción asignada final después de aclaraciones y asignaciones en los módulos CDO	6

Equivalencias de la columna FUENTE	
Clave	**Descripción**
1	Dirección General de Evaluación, SEP. Proceso de registro (RAM2001.DBF)
2	Hoja de datos generales, CENEVAL
3	Procesos de identificación del CENEVAL
4	Hojas de respuestas, CENEVAL
5	Proceso de calificación, CENEVAL
6	Proceso de asignación, CENEVAL

Tabla de interpretación del campo Expl_Mod	
Expl_Mod	**Descripción**
<31	Menor a 31 aciertos
ASI	Asignado en el proceso de asignación del CENEVAL
ASI_CD	CDO asignado en módulo CDO ○
ASI_CS	Asignados en módulo CDO presentando certificado ○
ASI_NP	NP asignado (si presentó examen)
ASI_RA	Reasignación automática debido a cambio de promedio ■
ASI_RM	Reasignación en módulo CDO debido a cambio de promedio ○
ASI_SC	Asignados al comprobar que si tenían certificado ■
ASI_RI	Reasignación debido a error en instructivo
CDO_SC	CDO debido a que presentó certificado
CDO	Con Derecho a otra Opción
NP	No presentó examen
SC	Sin certificado

Tabla de interpretación del campo Expl_Asi y Expl_Fin	
Expl_Mod	**Descripción**
<31	Menor a 31 aciertos
ASI	Asignado en el proceso de asignación del CENEVAL
CDO	Con Derecho a otra Opción
NP	No presentó examen
SC	Sin certificado

○ Asignado en una opción diferente a las elegidas
■ Asignado en una de sus opciones elegidas

METROPOLITANO 2002
Descripción del Archivo Metro2002.dbf

im.	Campo	Tipo	Longitud	Descripción	Fuente
1	REGISTRO	Carácter	10	Institución en la que se registro el sustentante	1
2	EXAMINO	Carácter	10	Institución en la que aplicó el sustentante	1
3	CURP	Carácter	16	Clave única de registro poblacional	1
4	TURN_APL	Carácter	1	Turno de aplicación del examen	1
5	CVE_SEDE	Carácter	4	Clave de la sede de aplicación del examen	1
6	CVE_INSAPL	Carácter	1	Clave de la institución aplicadora del examen	1
7	CVE_CCT	Carácter	10	Clave de la escuela secundaria del sustentante	1
8	REGI_SEC	Carácter	3	Régimen de la secundaria de origen PUB. = PUBLICA, PRI = PRIVADA	3
9	MODA_SEC	Carácter	1	Modalidad de la secundaria de origen G=General, T=Técnica, W=Trabajadores, V=Telesecundaria, A=Abierta	3
10	MUN_CCT	Carácter	3	Clave del municipio donde se localiza la escuela de procedencia del sustentante	1
11	FOLIO	Carácter	9	Folio del sustentante	1
12	CVE_GPO	Carácter	4	Clave del grupo donde presenta el examen	1
13	SUS_PAT	Carácter	30	Apellido paterno del sustentante	1
14	SUS_MAT	Carácter	30	Apellido materno del sustentante	1
15	SUS_NOM	Carácter	40	Nombre(s) del sustentante	1
16	SUS_FNAC	Fecha	8	Fecha de nacimiento del sustentante	1
17	CVE_ENTNAC	Carácter	2	Clave de la entidad de nacimiento del sustentante	1
18	SUS_SEXO	Carácter	1	Sexo del sustentante (H = Hombre, M = Mujer)	1
19	SUS_DOM	Carácter	40	Domicilio del sustentante, calle y número	1
20	SUS_COL	Carácter	30	Colonia del domicilio del sustentante	1
21	SUS_CP	Carácter	5	Código postal del domicilio del sustentante	1
22	SUS_DEL	Carácter	30	Municipio o delegación del domicilio del sustentante	1
23	SUS_TEL	Carácter	13	Teléfono del sustentante	1
24	SUS_PREG	Carácter	7	Folio del pre-registro	1
25	SUS_PROM	Numérico	5.2	Promedio general del tercero de secundaria del sustentante	1
26	SUS_CATG	Carácter	1	Categoría del sustentante egresado, Foráneo o INEA	1
27	FOL_CERT	Carácter	10	Folio del certificado del sustentante	1
28	ANO_CERT	Carácter	4	Año del certificado del sustentante	1
29	SUS_TURN	Carácter	1	Preferencia de turno vespertino (X=turno vespertino)	1
30	OPC_ED01	Carácter	6	Opción educativa solicitada número 1	1
31	OPC_ED02	Carácter	6	Opción educativa solicitada número 2	1
32	OPC_ED03	Carácter	6	Opción educativa solicitada número 3	1
33	OPC_ED04	Carácter	6	Opción educativa solicitada número 4	1
34	OPC_ED05	Carácter	6	Opción educativa solicitada número 5	1
35	OPC_ED06	Carácter	6	Opción educativa solicitada número 6	1
36	OPC_ED07	Carácter	6	Opción educativa solicitada número 7	1
37	OPC_ED08	Carácter	6	Opción educativa solicitada número 8	1
38	OPC_ED09	Carácter	6	Opción educativa solicitada número 9	1
39	OPC_ED10	Carácter	6	Opción educativa solicitada número 10	1
40	OPC_ED11	Carácter	6	Opción educativa solicitada número 11	1
41	OPC_ED12	Carácter	6	Opción educativa solicitada número 12	1

METROPOLITANO 2002
Descripción del Archivo Metro2002.dbf

im.	Campo	Tipo	Longitud	Descripción	Fuente
42	OPC_ED13	Carácter	6	Opción educativa solicitada número 13	1
43	OPC_ED14	Carácter	6	Opción educativa solicitada número 14	1
44	OPC_ED15	Carácter	6	Opción educativa solicitada número 15	1
45	OPC_ED16	Carácter	6	Opción educativa solicitada número 16	1
46	OPC_ED17	Carácter	6	Opción educativa solicitada número 17	1
47	OPC_ED18	Carácter	6	Opción educativa solicitada número 18	1
48	OPC_ED19	Carácter	6	Opción educativa solicitada número 19	1
49	OPC_ED20	Carácter	6	Opción educativa solicitada número 20	1
50	PRE_EXA	Carácter	1	Presentó examen	3
51	PLAN_TRA	Carácter	1	Al iniciar, identifico lo que necesito estudiar y hago un plan de trabajo	2
52	APUN_CLA	Carácter	1	Estudio principalmente con mis apuntes de clase	2
53	UTIL_MON	Carácter	1	Utilizo las monografías que venden en las papelerías	2
54	EST_TEXT	Carácter	1	Estudio principalmente con el libro de texto de la materia	2
55	UTIL_ENC	Carácter	1	Utilizo enciclopedias, diccionarios y atlas	2
56	UTIL_PC	Carácter	1	Utilizo computadora	2
57	EST_ACOM	Carácter	1	Estudio principalmente con los apuntes de mis compañeros	2
58	EST_EQUI	Carácter	1	Estudio en equipo con mis compañeros de clase	2
59	LEN_IND	Carácter	1	¿Tu lengua materna es indígena?	2
60	TIEM_EST	Carácter	1	A la Semana, ¿Cuántos días estudias fuera del horario escolar?	2
61	HORA_EST	Carácter	2	A la Semana, ¿Cuántas horas dedicas al estudio fuera del horario escolar?	2
62	EXA_SEC	Carácter	2	¿Cuántos exámenes extraordinarios presentaste en la secundaria?	2
63	GUST_MAT	Carácter	1	Gusto por las matemáticas	2
64	GUST_INT	Carácter	1	Gusto por la introducción a la física y a la química	2
65	GUST_FIS	Carácter	1	Gusto por la física	2
66	GUST_QUI	Carácter	1	Gusto por la química	2
67	GUST_BIO	Carácter	1	Gusto por la biología	2
68	GUST_ESP	Carácter	1	Gusto por el español	2
69	GUST_HIS	Carácter	1	Gusto por la historia	2
70	GUST_GEO	Carácter	1	Gusto por la geografía	2
71	GUST_CIV	Carácter	1	Gusto por el civismo	2
72	GUST_EXT	Carácter	1	Gusto por la lengua extranjera	2
73	GUST_ART	Carácter	1	Gusto por la expresión y apreciación artística	2
74	GUST_EDU	Carácter	1	Gusto por la educación física	2
75	GUST_TEC	Carácter	1	Gusto por la educación tecnológica	2
76	EXTR_MAT	Carácter	1	Matemáticas	2
77	EXTR_INT	Carácter	1	Introducción a la física y la química	2
78	EXTR_FIS	Carácter	1	Física	2
79	EXTR_QUI	Carácter	1	Química	2
80	EXTR_BIO	Carácter	1	Biología	2
81	EXTR_ESP	Carácter	1	Español	2
82	EXTR_HIS	Carácter	1	Historia	2

METROPOLITANO 2002
Descripción del Archivo Metro2002.dbf

Im.	Campo	Tipo	Longitud	Descripción	Fuente
83	EXTR_GEO	Carácter	1	Geografía	2
84	EXTR_CIV	Carácter	1	Civismo	2
85	EXTR_EXT	Carácter	1	Lengua extranjera	2
86	EXTR_ART	Carácter	1	Expresión y apreciación artística	2
87	EXTR_EDU	Carácter	1	Educación física	2
88	EXTR_TEC	Carácter	1	Educación tecnológica	2
89	SEG_ESTU	Carácter	1	Tienes la intención de seguir estudios superiores después de la educación media superior	2
90	INST_ING	Carácter	1	A qué institución de educación superior te gustaría ingresar al terminar tus estudios de educación media superior	2
91	TIEM_LEE	Carácter	2	A la semana ¿Cuántas horas de tu tiempo libre dedicas a la lectura, sin tomar en cuenta las que dedicas a tus lib	2
92	LEE_TEX	Carácter	2	¿Cuántos libros completos has leído en los últimos doce meses sin tomar en cuenta tus libros de texto?	2
93	LEE_FREC	Carácter	1	¿Qué es lo que lees con más frecuencia?	2
94	TIPO_LEC	Carácter	1	¿Qué tipo de lectura te gusta más?	2
95	TIEM_TV	Carácter	2	¿Cuántas horas de tu tiempo libre inviertes en ver televisión al día?	2
96	ASIS_REG	Carácter	1	Los maestros faltan a clase o llegan tarde	2
97	BIEN_ALU	Carácter	1	El director o la directora se preocupan por el bienestar de los alumnos	2
98	AMB_AGRA	Carácter	1	El ambiente de tu grupo es agradable y amistoso	2
99	OTRA_ACT	Carácter	1	Se pierde tiempo de clase en otras actividades	2
00	NOR_DICI	Carácter	1	Las normas de disciplina se respetan	2
01	APR_ALUM	Carácter	1	Los maestros se preocupan por el aprendizaje de los alumnos	2
02	RELA_PRO	Carácter	1	Hay buenas relaciones entre los maestros (se llevan bien)	2
03	OBS_CLAS	Carácter	1	El director o la directora visitan tu grupo para observar la clase	2
04	REL_PADR	Carácter	1	Hay buenas relaciones entre los padres de familia y la escuela (maestros, personal directivo y administrativo)	2
05	CUAN_HER	Carácter	1	¿Cuántos hermanos tienes?	2
06	LUG_OCUP	Carácter	1	¿Qué lugar ocupas entre tus hermanos? (de mayor a menor)	2
07	EDAD_MAD	Carácter	1	Edad de la madre	2
08	EDAD_PAD	Carácter	1	Edad del padre	2
09	ACT_TARE	Carácter	1	Me ayudan a mis tareas escolares	2
10	ACT_ESCU	Carácter	1	Me felicitan o premian cuando me va bien en la escuela	2
11	ACT_OPIN	Carácter	1	Respetan mis opiniones sobre lo que ocurre en la escuela	2
12	ACT_DECI	Carácter	1	Promueven que tome mis propias decisiones sobre lo que pasa en la escuela	2
13	VIVE_PAD	Carácter	1	Vives con tu padre	2
14	VIVE_MAD	Carácter	1	Vives con tu madre	2
15	VIVE_HER	Carácter	1	Vives con tus hermanos	2
16	VIVE_FAM	Carácter	1	Vives con otros familiares	2
17	VIVE_CON	Carácter	1	Vives con tu cónyuge o pareja	2
18	VIVE_SOL	Carácter	1	Vives solo	2
19	VIVE_EST	Carácter	1	Vives con tus compañeros de estudio	2
20	VIVE_TRA	Carácter	1	Vives con tus compañeros de trabajo	2
21	VIVE_SIT	Carácter	1	Otra situación	2
22	PADR_GUS	Carácter	1	A mis padres les gustaría que por lo menos yo:	2
23	TIEM_CON	Carácter	1	¿Cuántas horas al dia conviven contigo tus padres en los dias de trabajo?	2

METROPOLITANO 2002
Descripción del Archivo Metro2002.dbf

Núm.	Campo	Tipo	Longitud	Descripción	Fuente
24	ESCO_MAD	Carácter	2	Escolaridad de la madre	2
25	ESCO_PAD	Carácter	2	Escolaridad del padre	2
26	OCU_MADR	Carácter	2	Ocupación de la madre	2
27	OCU_PADR	Carácter	2	Ocupación del padre	2
28	CON_CARN	Carácter	2	Carnes de res, cerdo, pollo o pescado (una porción equivale a 200 gr.)	2
29	CON_HUEV	Carácter	2	Huevos (una porción equivale a un huevo)	2
30	CON_LECH	Carácter	2	Leche (una porción equivale a 1/4 de litro o a un vaso mediano)	2
31	CON_FRUT	Carácter	2	Fruta y verduras frescas (una porción equivale a una fruta o una verdura)	2
32	CON_LENT	Carácter	2	Frijol, arroz, lentejas, habas, etc. (una porción equivale a un plato)	2
33	CON_PAN	Carácter	2	Pan (una porción equivale a una pieza de pan)	2
34	CON_CERE	Carácter	2	Cereales (una porción equivale a un plato)	2
35	NUM_PERS	Carácter	2	Número de personas que vive en tu casa incluyéndote a ti.	2
36	CUAN_CUA	Carácter	2	¿Cuántos cuartos de tu casa se utilizan para dormir?	2
37	TRABAJA	Carácter	1	Actualmente, ¿Desarrollas algún trabajo por el cual recibes un sueldo?	2
38	INGR_FAM	Carácter	2	¿Cuál es el ingreso familiar mensual?	2
39	EXAMEN	Carácter	2	Identifica poblaciones PRESENTADOS y NO PRESENTADOS del CENEVAL y UNAM	3
40	NGLOBAL	Numérico	3	Número de aciertos totales obtenidos en el examen	5
41	NHV	Numérico	3	Número de aciertos obtenidos en habilidad verbal (24 preguntas totales)	5
42	NESP	Numérico	3	Número de aciertos obtenidos en español (10 preguntas totales)	5
43	NHIS	Numérico	3	Número de aciertos obtenidos en historia (10 preguntas totales)	5
44	NGEO	Numérico	3	Número de aciertos obtenidos en geografía (10 preguntas totales)	5
45	NFCE	Numérico	3	Número de aciertos obtenidos en formación cívica ética (10 preguntas totales)	5
46	NHM	Numérico	3	Número de aciertos obtenidos en habilidad matemática (24 preguntas totales)	5
47	NMAT	Numérico	3	Número de aciertos obtenidos en matemáticas (10 preguntas totales)	5
48	NFIS	Numérico	3	Número de aciertos obtenidos en física (10 preguntas totales)	5
49	NQUI	Numérico	3	Número de aciertos obtenidos en química (10 preguntas totales)	5
50	NBIO	Numérico	3	Numero de aciertos obtenidos en biología (10 preguntas totales)	5
51	PNGLOBAL	Numérico	8.2	Porcentaje de aciertos obtenidos en el examen	5
52	PNHV	Numérico	8.2	Porcentaje de aciertos obtenidos en habilidad verbal	5
53	PNESP	Numérico	8.2	Porcentaje de aciertos obtenidos en español	5
54	PNHIS	Numérico	8.2	Porcentaje de aciertos obtenidos en historia	5
55	PNGEO	Numérico	8.2	Porcentaje de aciertos obtenidos en geografía	5
56	PNFCE	Numérico	8.2	Porcentaje de aciertos obtenidos en formacióncívica ética	5
57	PNHM	Numérico	8.2	Porcentaje de aciertos obtenidos en habilidad matemática	5
58	PNMAT	Numérico	8.2	Porcentaje de aciertos obtenidos en matemáticas	5
59	PNFIS	Numérico	8.2	Porcentaje de aciertos obtenidos en física	5
60	PNQUI	Numérico	8.2	Porcentaje de aciertos obtenidos en química	5
61	PNBIO	Numérico	8.2	Porcentaje de aciertos obtenidos en biología	5
62	SIN_CERT	Carácter	2	Identifica a los sustentantes que no tienen certificado de secundaria (S/C)	3
63	NO_PRES	Carácter	2	Identifica a los sustentantes que no presentaron examen (N/P)	3
64	MENOS_31	Carácter	3	Identifica a los sustentantes que obtuvieron menos de 31 aciertos en el examen (<31)	3

METROPOLITANO 2002
Descripción del Archivo Metro2002.dbf

ım.	Campo	Tipo	Longitud	Descripción	Fuente
65	BAJA_BI	Carácter	2	Identifica a los sustentantes que causaron baja del examen por infracción (BI)	3
66	EXPL_ASI	Carácter	5	Explicación después de haber corrido el proceso de asignación en el CENEVAL	6
67	NOPC_SOL	Carácter	3	Número de opciones solicitadas por el sustentante	6
68	NOPC_ASI	Numérico	2	Número de opción asignada	6
69	COPC_ASI	Carácter	6	Clave de la opción educativa asignada en el proceso de asignación en el CENEVAL	6
70	EXPL_MOD	Carácter	7	Explicación de modificaciones después de aclaraciones y asignaciones en los módulos CDO	6
71	ASIG_FIN	Carácter	6	Clave de opción asignada final después de aclaraciones y asignaciones en los módulos CDO	6
72	EXPL_FIN	Carácter	7	Explicación final de la asignación después de aclaraciones y asignaciones en los módulos CDO	6
73	NOPC_FIN	Numérico	2	Número de opción asignada final después de aclaraciones y asignaciones en los módulos CDO	6

Equivalencias de la columna FUENTE	
Clave	**Descripción**
1	Dirección General de Evaluación, SEP. Proceso de registro (RAM2002.DBF)
2	Hoja de datos generales, CENEVAL
3	Procesos de identificación del CENEVAL
4	Hojas de respuestas, CENEVAL
5	Proceso de calificación, CENEVAL
6	Proceso de asignación, CENEVAL

Tabla de interpretación del campo Expl_Mod	
Expl_Mod	**Descripción**
<31	Menor a 31 aciertos
ASI	Asignado en el proceso de asignación del CENEVAL
ASI_CD	CDO asignado en módulo CDO ○
ASI_CS	Asignados en módulo CDO presentando certificado ○
ASI_NP	NP asignado (si presentó examen)
ASI_RA	Reasignación automática debido a cambio de promedio ■
ASI_RM	Reasignación en módulo CDO debido a cambio de promedio ○
ASI_SC	Asignados al comprobar que si tenían certificado ■
ASI_RI	Reasignación debido a error en instructivo
CDO_SC	CDO debido a que presentó certificado
CDO	Con Derecho a otra Opción
NP	No presentó examen
SC	Sin certificado

Tabla de interpretación del campo Expl_Asi y Expl_Fin	
Expl_Mod	**Descripción**
<31	Menor a 31 aciertos
ASI	Asignado en el proceso de asignación del CENEVAL
CDO	Con Derecho a otra Opción
NP	No presentó examen
SC	Sin certificado

○ Asignado en una opción diferente a las elegidas
■ Asignado en una de sus opciones elegidas

NEWING
Descripción del Archivo

im.	Campo	Tipo	Longitud	Descripción	Fuente
1	CURP	Carácter	16	Clave única de registro poblacional	1
2	SUS_PAT	Carácter	30	Apellido paterno del sustentante	1
3	SUS_MAT	Carácter	30	Apellido materno del sustentante	1
4	SUS_NOM	Carácter	40	Nombre(s) del sustentante	1
5	SUS_FNAC	Fecha	8	Fecha de nacimiento del sustentante	1
6	CVE_ENTNAC	Carácter	2	Clave de la entidad de nacimiento del sustentante	1
7	SUS_SEXO	Carácter	1	Sexo del sustentante (H = Hombre, M = Mujer)	1
8	SUS_DOM	Carácter	40	Domicilio del sustentante, calle y número	1
9	SUS_COL	Carácter	30	Colonia del domicilio del sustentante	1
10	SUS_CP	Carácter	5	Código postal del domicilio del sustentante	1
11	SUS_DEL	Carácter	30	Municipio o delegación del domicilio del sustentante	1
12	SUS_TEL	Carácter	13	Teléfono del sustentante	1
13	SUS_PROM	Numérico	5.2	Promedio general del tercero de secundaria del sustentante	1
14	PREXA_01	Carácter	1	Al iniciar, identifico lo que necesito estudiar y hago un plan de trabajo	2
15	PREXA_02	Carácter	1	Estudio principalmente con mis apuntes de clase	2
16	PREXA_03	Carácter	1	Utilizo las monografías que venden en las papelerías	2
17	PREXA_04	Carácter	1	Estudio principalmente con el libro de texto de la materia	2
18	PREXA_05	Carácter	1	Utilizo enciclopedias, diccionarios y atlas	2
19	PREXA_07	Carácter	1	Estudio principalmente con los apuntes de mis compañeros	2
20	PREXA_11	Carácter	1	Estudio en equipo con mis compañeros de clase	2
21	HORA_EST	Carácter	2	A la Semana, Cuantas horas dedicas al estudio fuera del horario escolar	2
22	GUS_MAT	Carácter	1	Gusto por las matemáticas	2
23	GUS_IFQ	Carácter	1	Gusto por la introducción a la física y a la química	2
24	GUS_FIS	Carácter	1	Gusto por la física	2
25	GUS_QUI	Carácter	1	Gusto por la química	2
26	GUS_BIO	Carácter	1	Gusto por la biología	2
27	GUS_ESP	Carácter	1	Gusto por el español	2
28	GUS_HIS	Carácter	1	Gusto por la historia	2
29	GUS_GEO	Carácter	1	Gusto por la geografía	2
30	GUS_CIV	Carácter	1	Gusto por el civismo	2
31	GUS_LEX	Carácter	1	Gusto por la lengua extranjera	2
32	GUS_ART	Carácter	1	Gusto por la expresión y apreciación artística	2
33	GUS_EFI	Carácter	1	Gusto por la educación física	2
34	GUS_EDT	Carácter	1	Gusto por la educación tecnológica	2
35	SEC_EST	Carácter	1	Tienes la intensión de seguir estudios superiores después de la educación media superior	2
36	TPO_LEC	Carácter	2	Cuántas horas de tu tiempo libre inviertes en leer	2
37	TIP_LEC	Carácter	2	Qué tipo de lectura te gusta más	2
38	TPO_TV	Carácter	2	Cuántas horas de tu tiempo libre inviertes en ver televisión	2
39	CUA_HNO	Carácter	2	Cuántos hermanos tienes	2
40	LUG_OCU	Carácter	2	Qué lugar ocupas entre tus hermanos	2
41	EDAD_MAD	Carácter	1	Edad de la madre	2

NEWING

Descripción del Archivo

im.	Campo	Tipo	Longitud	Descripción	Fuente
42	EDAD_PAD	Carácter	1	Edad del padre	2
43	VIVE_CON	Carácter	2	Con quién vives actualmente	2
44	AS_ESC01	Carácter	1	Me ayudan en mis tareas escolares	2
45	AS_ESC03	Carácter	1	Me orientan sobre lecturas o actividades que pueden ampliar mi aprendizaje escolar	2
46	AS_ESC06	Carácter	1	Me premian o me felicitan cuando me va bien en la escuela	2
47	AS_ESC07	Carácter	1	Me castigan cuando me va mal en la escuela	2
48	ESC_MAD	Carácter	2	Escolaridad de la madre	2
49	ESC_PAD	Carácter	2	Escolaridad del padre	2
50	OCU_MAD	Carácter	2	Ocupación de la madre	2
51	OCU_PAD	Carácter	2	Ocupación del padre	2
52	FRE_ALI1	Carácter	1	Con que frecuencia comes carne de res, cerdo, pollo o pescado	2
53	FRE_ALI2	Carácter	1	Con que frecuencia comes huevos	2
54	FRE_ALI3	Carácter	1	Con que frecuencia comes leche y derivados	2
55	FRE_ALI4	Carácter	1	Con que frecuencia comes frutas y verduras frescas	2
56	FRE_ALI5	Carácter	1	Con que frecuencia comes frijol, arroz, lentejas, habas, etc.	2
57	FRE_ALI6	Carácter	1	Con que frecuencia comes pan y pastas	2
58	NUM_PERS	Carácter	2	Número de personas que viven en tu casa	2
59	TRABAJA	Carácter	1	Desarrollas algún trabajo por el cual recibes un sueldo	2
60	INGR_FAM	Carácter	2	Cuál es el ingreso familiar mensual	2

BIBLIOGRAFIA

DATE, C J.: *Introducción a los sistemas de bases de datos.*
Ed. Addison – Wesley Iberoamericana, V. 1, 5ª E, 1993

ENSOR, Dave and STEVENSON, Ian.: *Oracle Design.*
Ed. O´REILLY, EUA, 1997

JARKE, Matthias et al. : *Fundamental of Data Warehouses*
Ed. Springer –Verlag, Berlin Heidelberg, Alemania, 2000.

LIPSCHUTZ, Seymour. Matemáticas para computación,
Ed. Mc. Graw – Hill,1992

SARABIA Ramírez Luis G. y BOLAÑOS Usla Miguel R.: *El Data Warehouse de Bancomext*
En: Política Digital, NEXOS, México, Número 15, febrero 2004

THURAISINGHAM, Bhavani. : *Data Mining Technologies, techniques, tool and trends*
Ed. CRC Press, USA, 1999.

JURGENS, Marcus. *Index Structures for Data Warehouses*,
Ed. Springer – Verlag Berlin Heidelerg Germany 2002

PERNER Petra and Petrou Maria, *Machine Learning and Data Mining in Pattern Recognition*, First International Workshop, MLDM'99 Leipzing, Germany, September 16- 18, 1999 Proceedings.
Ed. Springer – Verlag Berlin 1999.

NEDELLEC, Claire and ROUVEIROL Celine (Eds.), *Machine Learning: ECML – 98*, 10th European Conference on Machine Learning Chemnitz, Germany, April 21-23, 1998 Proceedings.
Ed. Springer – Verlag Berlin Heidelberg Germany 1998

THERRIEN, Charles W., *Decision, estimation and clasification. An Introduction to Pattern Recognition and Related Topics.*
Ed. John Wiley & Sons, New York

Kandel Abraham et al (Eds.), Data Mining and Computational Intelligence,
Ed. Physica – Verlag Heidelberg Germany 2001

LINKS

F. Morales Eduardo, Curso de Minería de Datos,
http://dns1.mor.itesm.mx/~emorales/Cursos/KDD01/principal.html,
ITESM, consultada el 20 de marzo de 2004

CASARES, Claudio.: *Data Warehousing*,
www.programacion.com/bbdd/tutorial/warehouse/1/, Consultada 24 de marzo de 2004

JIMENEZ, Claudia.: *Bases de Datos Multidimensionales*,
www.inf.udec.cl/~basedato/trabajos/multidimensionales.pdf , Departamento de Ingeniería Informática y Ciencias de la Computación Universidad de Concepción, Agosto 2002

KOHAVI, Ron and QUINLAN, Ross, *Decision Tree Discovery*,
http://robotics.stanford.edu/~ronnyk/treesHB.pdf , 1999.

SANCHEZ Montoya, Ricardo. Business intelligence... BI or not to BI. http:// www. monografías.com/trabajos14/bi/bi.shtml

BERGOS, Massagué Jordi, BEAN: Behavior Analyser. 21 de mayo de 2004

ESTIVILL Castro, Vladimir., *Tres retos de la minería de datos,* www.lania.mx/biblioteca/newsletters/1999-otono-invierno/retos_mineria.html Laboratorio Nacional de Informática Avanzada A. C. 1999, Consultada el 3 de febrero de 2005

SETH, Paul et al, *"Preparing and Mining Data with Microsoft SQL Server 2000 and Analysis Services"*, www.microsoft.com , Microsoft Online Books.

Colegio de Bachilleres, www.cbacilleres.edu.mx, consultada el 29 de mayo de 2005

GONZALEZ, José Carlos , Gestión del Conocimiento, http://www.daedalus.es, DAEDALUS – Data, Decisions and Lenguaje, S. A. 2001

http://www.cs.us.es/~delia/sia/html98-99/pag-alumnos/web2/indice.html, Documento sobre árboles de decision. Consultada el 25 de marzo de 2004.

MURILLO GARIBAY, Víctor M, Modelo Multidimencional, http://galeon.com/materiasis/molap.html. Consultada el 25 de marzo de 2004

WOLF, Carmen Gloria, *Documento sobre el modelo multidimencional,* http://revista.inf.udec.cl/ediciones/edicion4/modmulti.PDF, Consultada el 25 de marzo de 2004

www.monografías.com, Inteligencia de Negocios (BI), Consultada el 14 de julio de 2004.

HERNANDEZ NOVICH, Ernesto, http://www.linux.org.ve/archivo/l-linux-2001-December/034453.html

Concepto de Minería de Datos, http://answermath.com/data-mining/mineria-de-datos-2-concepto.htm, Consultada el 20 de mayo de 2004.

Minería de Datos, http://www.crisp-dm.org/. Consultada el 3 de junio de 2004

INFINITA CONSULTORES, http://www.infinitax.com/notas/notas.htm, Consultada el 3 de junio de 2004

Cubos de datos, http://www.prado.com.mx/Cubos/cubos-definiciones-y-conceptos-4.htm, Consultada el 3 de junio de 2004

Data Warehouse, http://www.csi.map.es/csi/silice/DW2251.html, Consultada el 3 de junio de 2004

HP México. http://www3.hp.com/servicios/aplicaciones_empresariales/enter_negocios_mineria.html. Consultada el 30 de junio de 2005

Pearson. http://www.pearson_research.com/nuestra-empresa.phtml. Consultada el 30 de junio de 2005

MBSystems de México. http://www.mbsystems.com.mx/somos.html. Consultada el 30 de junio de 2005

INFOTEC. http://www.infotec.com.mx/wb2/infotec/into_quienes_somos.htm Consultada el 30 de junio de 2005

INVAP. http://www.invap.net/about/perfil.html. Consultada el 30 de junio de 2005

DAEDALUS. http:// www.daedalus.es/ Consultada el 30 de junio de 2005

DATOLOGIA. http://www.datologia.com/empresa.html Consultada el 30 de junio de 2005

INFOMEDIA. http://www.infomedia.com.mx Consultada el 30 de junio de 2005

PROCALIDAD. http://procalidad.com/compania/index.aspx Consultada el 30 de junio de 2005

Tesis de Minería de datos, http://info.pue.udlap.mx/~tesis/msp/pech_p_ma/capitulo2.pdf. Consultada el 15 de abril de 2005

http://www.palermo.edu.ar/ingenieria/downloads/MaterialRMyA.pdf, Consultada el 15 de abril de 2005

http://ccc.inaoep.mx/~jagonzalez/DM/MineriaDatos-BN.pdf, Consultada el 15 de abril de 2005.

http://www.ideasoft.com.uy/consult/dm2.htm, Consultada el 15 de abril de 2005

http://www.olapxsoftware.com/es/WhatIsOlap.asp, Consultada el 15 de abril de 2005.

http://www.olapxsoftware.com/es/support/faq.asp, Consultada el 15 de abril de 2005.

http://www.microsoft.com/latam/technet/articulos/200005/art03/, Consultada el 17 de abril de 2005.

http://www.bitam.com.mx/TecAnalisis.htm#contenerDSS, Consultada el 17 de abril de 2005.

http://www.inf.ufrgs.br/~clesio/cmp151/cmp15120031/Datawarehouse_9pp.pdf, Consultada el 17 de abril de 2005.

http://www.1keydata.com/datawarehousing/inmon-kimball.html, Consultada el 17 de abril de 2005.

http://www.csi.map.es/csi/silice/DW2251.html#MOLAP, Consultada el 17 de abril de 2005.