Instituto Tecnológico de Costa Rica

Departamento de Computación

Programa de Maestría
Énfasis: Sistemas de Información

ESTABLECIMIENTO DE UN GRUPO DE INDICADORES PARA LA TOMA DE DECISIONES EN EL SECTOR VIVIENDA UTILIZANDO TÉCNICAS DE *MINERIA DE DATOS*

Informe final de tesis para optar por el grado de Máster en Computación

Estudiante: Giannina Ortiz Quesada

Profesor Tutor: Dr. Carlos González Alvarado

Cartago, Setiembre 2003

Resumen

El sector vivienda, es un sector de suma importancia para el desarrollo del país y su crecimiento desordenado ha generado un impacto negativo en los últimos años. Si consideramos los últimos datos presentados como parte del trabajo del Plan de Desarrollo Urbano, son evidentes las lamentables tendencias que presenta el desarrollo urbano costarricense.

Muchos de estos problemas se podrían evitar o minimizar, si a la hora de tomar decisiones en este campo se tuviera una visión global de la situación y para ello las instituciones coordinaran entre si, aportando cada una la información que le corresponde según su ámbito de trabajo.

En Costa Rica existe una serie de instituciones que generan y procesan informes referentes al sector vivienda, sin embargo, cada una de estas instituciones trabaja por separado, generando datos de interés particular sin tener una visión global de la utilidad general de esa información recopilada, sobre todo del valor agregado de ésta en la toma de decisiones. Con ello se pierde el valor agregado de la labor desempeñada por los profesionales y los políticos en esta área.

Por ejemplo, el Ministerio de Vivienda tiene información sobre la cual se basa para tomar sus decisiones, pero muchas veces no se asocia esta información con la de la Comisión Nacional de Emergencia, la cual tiene datos sobre las principales zonas de riesgo o bien con Acueductos y Alcantarillados para informarse sobre la distribución de los mantos acuíferos o la disponibilidad de agua potable en una zona determinada. Este problema se presenta en casi todas las instituciones u organizaciones, incluyendo las mismas municipalidades.

Según un análisis preliminar y realizado de forma manual, se notó la presencia en ciertas viviendas de casos de enfermedades tales como alergias, asma y otras enfermedades de tipo respiratorio, relacionadas con el hacinamiento provocado por la construcción actual (casas totalmente pegadas unas de otras, dimensiones muy pequeñas, poca ventilación e iluminación, entre otros).

Esto se podría evitar si, a la hora de establecer las políticas de vivienda, éstas se basaran en la información de los entes de salud relacionada con las enfermedades provocadas por aspectos directamente vinculados a las viviendas.

Ante la situación descrita anteriormente es que se determinó la necesidad de crear una serie de indicadores que sean apoyo para la toma de decisiones. Para ello se presenta el siguiente trabajo, el cual es el informe final de tesis, para optar por el grado de Máster en Computación con énfasis en Sistemas de Información del Instituto Tecnológico de Costa Rica.

En este trabajo se presenta una metodología para la creación de Indicadores para la toma de decisiones en el sector vivienda, utilizando técnicas de minería de datos. Se analizaron las necesidades de información del sector vivienda, la información disponible, las técnicas de minería de datos existentes, se determinaron las más apropiadas y se finalizó con un prototipo utilizando el programa SQL Server 2000.

En este trabajo es de suma importancia resaltar la relación de la ciencia de la computación con otras áreas, para el particular, el sector vivienda del país.

También se deja plasmado en este documento la necesidad de seguir trabajando en el tema y se dan algunas recomendaciones.

Palabras claves: *Minería de Datos, Vivienda, Indicadores, Sistemas de Información, Toma de decisiones.*

Indice General

INDICE DE FIGURAS Y TABLAS 5

CAPITULO 1. INTRODUCCIÓN 6

1.1 EL PROBLEMA Y SU IMPORTANCIA 6
1.2 ANTECEDENTES DEL PROBLEMA 12
1.3 OBJETIVOS 16
1.3.1 OBJETIVO GENERAL 16
1.3.2 OBJETIVOS ESPECÍFICOS: 16

CAPITULO 2. MARCO CONCEPTUAL 17

2.1 FUENTES DE DATOS 17
2.1.1 ESTADÍSTICAS DEL INSTITUTO NACIONAL DE ESTADÍSTICA Y CENSO (INEC) 18
2.1.2 MINISTERIO DE SALUD 25
DEFINICIÓN DE LAS REGIONES DE COSTA RICA SEGÚN MINISTERIO DE SALUD 25
2.1.3 CÁMARA COSTARRICENSE DE LA CONSTRUCCIÓN 26
2.1.4 INFORMACIÓN DISPONIBLE EN OTRAS INSTITUCIONES PARA ANALIZAR 28
2.2 HERRAMIENTAS DE ANÁLISIS 29
2.2.1 MINERÍA DE DATOS 29
2.2.2 REQUERIMIENTOS QUE DEBE CUMPLIR UN SISTEMA DE MINERÍA DE DATOS 31
2.2.3 DATAWAREHOUSE 35
2.2.4 TÉCNICAS DE MINERÍA DE DATOS 37
2.3 INDICADORES PARA LA TOMA DE DECISIONES 52
2.4 MINERÍA DE DATOS CON SQL SERVER 2000 54
2.4.1 EVOLUCIÓN HACIA EL DATA MINING 54
2.4.2 DESARROLLO DE DATA MINING CON SQL SERVER 2000 55
2.4.3 OLE DB PARA DATA MINING 57
2.4.4 EL PROCESO DE DATA MINING 58

CAPITULO 3. METODOLOGÍA 64

CAPITULO 4. INDICADORES PARA EL SECTOR VIVIENDA 73

4.1 NECESIDADES DE INFORMACIÓN DEL SECTOR CONSTRUCCIÓN 73
4.2 CREACIÓN DEL CUBO DE DATOS 75
4.1.1 MODELO CONCEPTUAL 76
4.1.2 USO DE SQL SERVER 2000 – ANALYSIS SERVER PARA EL DESARROLLO DEL CUBO 81
4.2 ESTABLECIMIENTO DE INDICADORES 87

CAPITULO 5. CONCLUSIONES Y RECOMENDACIONES 92

5.1 CONCLUSIONES 92
5.2 RECOMENDACIONES 95

BIBLIOGRAFÍA 97

ANEXOS ¡ERROR! MARCADOR NO DEFINIDO.

Indice de figuras y tablas

Figura 1-1. Ejemplos de ciudades desordenadas, Cartago. 7
Figura 1-2. Ejemplo de manejo inadecuado de desechos, Cartago. 8
Figura 2-1. Esquema de las fuentes de datos a analizar para obtener los indicadores 18
Tabla 2-1. Diccionario de datos de los censos nacionales de población y vivienda de Costa Rica, 2000 19
Tabla 2-2. Diccionario de datos de las proyecciones nacionales de población de Costa Rica, 1970-2100 20
Tabla 2-3. Diccionario de datos de las proyecciones cantonales de población de Costa Rica, 1975-2015 20
Figura 2-4. División de las subregiones de Costa Rica según el INEC 21
Figura 2-5. Subregiones de Costa Rica, según el INEC 22
Figura 2-6. División de las Regiones de Costa Rica según el Ministerio de Salud 26
Figura 2-7. Ciclo de vida de un datawarehouse 36
Figura 2-8. Representación grafica de una regla de asociación 39
Figura 2-9. Representación gráfica de cómo se resuelve una regla de asociación 40
Figura 2-10. Algoritmo Apriori 41
Figura 2-11. Ejemplo de taxonomía 42
Figura 2-12. Ejemplo de información asociada a los datos 42
Figura 2-13. Ilustración de las definiciones del ejemplo 46
Figura 2-14. Identificación de los grupos en los subespacios del espacio de datos original 46
Figura 2-15. Ejemplo de una regla de clasificación, mostrada como un árbol de decisión 48
Figura 2-16. 8 vistas de un cubo de datos para información de ventas 50
Figura 2-17. Ejemplo: Comparación de donaciones para investigación 51
Figura 2-18. Ejemplo: Consulta que crea un modelo de data mining 56
Figura 2-19. Ejemplo análisis invitación a un congreso 60
Figura 2-20. Ejemplo árbol de decisión 61
Figura 2-21. Escenario de implantación de OLE DB para data mining 63
Figura 4-1. Esquema del planteamiento del sistema 76
Figura 4-2. Definición de atributos 78
Figura 4-3. Dependencias funcionales 79
Figura 4-4. Diseño dimensional jerárquico 80
Figura 4-5. Relaciones existente en la base de datos prototipo 81
Figura 4-6. Ejemplo de pantalla una vez creada la base de datos 82
Figura 4-7. Ejemplo de pantalla creación del cubo, tabla de hechos 83
Figura 4-8. Ejemplo de pantalla definición de dimensiones, dimensión área geográfica 84
Figura 4-9. Ejemplo de pantalla definición de jerarquías, dimensión área geográfica 84
Figura 4-10. Ejemplo de pantalla definición de cubo, Indicadores Vivienda 85
Figura 4-11. Procesamiento de un cubo 86
Figura 4-12. Ejemplo pantalla de procesamiento del cubo 86
Figura 4-13. Visualización del cubo “Indicadores” 88
Figura 4-14. Ejemplo, indicador Vivienda y Salud 89
Figura 4-15. Ejemplo, indicador Accesibilidad al Crédito 90
Figura 4-16. Ejemplo, indicador Vulnerabilidad 91
Figura 4-17. Ejemplo, indicador Servicios Básicos y Salud 92

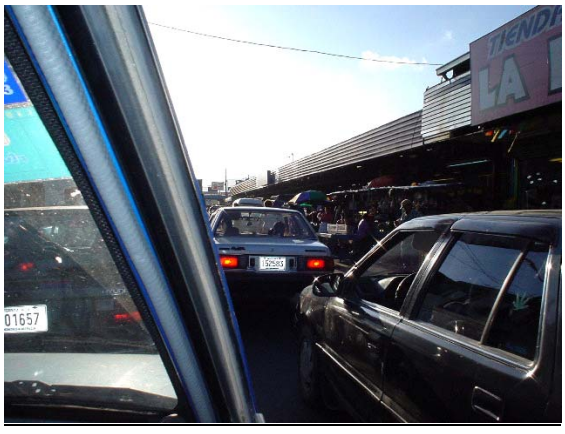
CAPITULO 1. Introducción

1.1 El problema y su importancia

En Costa Rica existe una serie de instituciones, las cuales generan y manipulan información referente al sector vivienda, el cual es de suma importancia para el desarrollo de nuestro país. Principalmente si consideramos los últimos datos presentados como parte del trabajo del Plan Nacional de Desarrollo Urbano [PNDU01], en donde son evidentes las lamentables tendencias que presenta el desarrollo urbano nacional, entre otros se pueden citar los siguientes:

Crecimiento horizontal desordenado

Este se observa en la expansión de las ciudades sin una adecuada planificación, concentración de población en algunas zonas, especialmente en la Gran Área Metropolitana, teniendo esta una densidad promedio de 75 habitantes por kilómetro cuadrado[1].



[1] Según datos del Censo Nacional 2000.

Figura 1-1. Ejemplos de ciudades desordenadas, Cartago.

Manejo inadecuado de recursos y desechos

Al darse una expansión desordenada de las ciudades, se presentan problemas con el manejo de recursos como: agua potable, aguas negras. Además de la generación de desechos, los cuales generan un problema de gran magnitud. Se sabe, por ejemplo, que sólo el 2% de las aguas negras de nuestro país son tratadas adecuadamente.

Figura 1-2. Ejemplo de manejo inadecuado de desechos, Cartago.

Generación de problemas ambientales de alto impacto

La carencia de vivienda e infraestructura básica hace que las poblaciones busquen la manera más sencilla, sin que muchas veces sea la correcta de disponer de sus desechos.

Viviendas que tienden a convertirse en enemigos de la salud (generación de enfermedades)

Los materiales de construcción, el diseño de las viviendas influyen en la salud de las personas, provocando en muchos casos problemas respiratorios, dermatológicos, alergias, entre otros.

Construcción en zonas de riesgo

La falta de planificación y las condiciones socioeconómicas, aunado al problema de inmigración, han hecho que muchas familias construyan en terrenos no aptos y con condiciones de riesgo enormes.

Problemas de accesibilidad a la infraestructura

Al expandirse las ciudades el costo de las obras de infraestructura se hace muy alto y en muchas ocasiones ni siquiera se llevan a cabo las obras, lo cual genera una gran cantidad de problemas. Además el transporte y sus implicaciones toman un papel relevante.

De los factores mencionados anteriormente, es importante resaltar los problemas de salud que podrían estar relacionados con las viviendas, ya que de análisis preliminares se ha obtenido información que relaciona problemas de alergias o problemas respiratorios a la falta de ventilación o bien a problemas de humedad o de hacinamiento.

Uno de los grandes problemas con la información que manejan las instituciones es que cada una trabaja por separado, generando y almacenando datos que llegan a convertirse en “tumbas de datos”, ya que no se tiene una visión global de lo que sirve a las personas que toman decisiones en el sector vivienda. En otras palabras, aunque existen muchos datos, estos no proporcionan información, ni valor agregado a la labor que desempeñan los profesionales y los políticos en esta área.

Por ejemplo, el Ministerio de Vivienda y Asentamientos Humanos cuenta con información sobre la cual se basa para tomar sus decisiones, pero muchas veces ésta no se asocia con la de la Comisión Nacional de Emergencias, la cual tiene datos sobre las principales zonas de riesgo o bien con Acueductos y Alcantarillados para informarse sobre la distribución de mantos acuíferos o la disponibilidad de agua potable en una

zona determinada. Y este tipo de problemas se presenta con casi todas las instituciones, incluyendo por supuesto a las mismas municipalidades.

Adicional a esta falta de integración de la información, se tienen también los “celos” en la información, ya que muchas instituciones creen que si comparten la información, están cediendo terreno en su labor profesional, por lo tanto será de suma importancia para cualquier planteamiento que se haga, convencer a estos entes de la importancia de tener información y no sólo datos almacenados.

Con base en el problema descrito es que se planteó la posibilidad de realizar una investigación en el área de Computación, específicamente en Sistemas de Información la cual propone desarrollar un modelo basado en una serie de indicadores para ayudar en la toma de decisiones en el sector vivienda en Costa Rica. Esta iniciativa no sólo ha sido de interés para realizar el trabajo de tesis presentado, sino que algunas instituciones públicas como el Instituto Tecnológico de Costa Rica, a través del Centro de Investigaciones en Vivienda y Construcción (CIVCO) de la Escuela de Ingeniería en Construcción, han mostrado interés en el tema y en el apoyo a este tipo de iniciativas.

Finalmente, el trabajo que se presenta, además de tener un importante contenido de computación en el área de la Minería de Datos, es importante para mejorar la toma de decisiones en el sector vivienda de nuestro país, lo cual realmente hace falta. En el trabajo se utilizaron algunas técnicas de minería de datos, dándole primordial importancia a las técnicas estadísticas, ya que por el tipo de datos que se manejaron.

Asimismo se profundizó en la metodología para la creación de un depósito de datos y el análisis de requerimientos para establecer realmente cuales son las necesidades a satisfacer, así como pasar de una teoría de cubos de información al planteamiento de pasos para establecer indicadores para la toma de decisiones, lo cual sería un aspecto innovador para el sector construcción y reafirma el hecho de que la computación es una herramienta de gran utilidad para el desarrollo de cualquier campo. Por lo tanto el impacto de esta investigación en el planteamiento de posibles soluciones a uno de los mayores problemas de nuestra sociedad es evidente.

1.2 Antecedentes del problema

Este trabajo se enfocará en demostrar que por medio de las técnicas de Minería de Datos se puede mejorar la toma de decisiones; como ejemplo, se pretende analizar cómo la construcción y diseño de las viviendas pueden tener relación con problemas de salud en la población. Para ello se utilizarán las herramientas tecnológicas que se tienen a disposición.

El utilizar estas técnicas de manera manual no es nuevo; el estudio de un caso de epidemiología en el año de 1854 en Londres, puede considerarse pionero en este campo. Hoy en día todos conocemos que el cólera es una terrible enfermedad causada por una bacteria que afecta los intestinos y se obtiene a través del agua contaminada. Pero en 1854, todas las personas pensaban que las muertes provocadas por esta enfermedad eran causadas por una sustancia que se encontraba en el aire. La gente, además, empezó a observar que las personas que vivían en condiciones antihigiénicas enfermaban. [JOHN01]

En 1850, Londres tenía un grave problema: la contaminación del río Támesis. El problema no era solamente el mal olor, sino que los habitantes bebían el agua del río, la cual se encontraba contaminada con materia fecal y en ese momento sólo dos de las muchas compañías que distribuían agua a la población la purificaban antes de distribuirla a los habitantes.

Aparece entonces el físico John Snow, quien estaba como practicante en Londres, con una “loca teoría” (catalogada así por la gente de ese tiempo) de que el agua contaminada era la causante de la enfermedad, esta teoría no tuvo mucho apoyo, pero a pesar de ello, Snow decide probarla. Construye entonces, lo que podemos llamar un datawarehouse (depósito de datos), recolectando los certificados de defunción; haciendo lo que hoy llamaríamos extracción, transformación y carga de datos (ETL), sobre las muertes de cólera mediante la tabulación de la dirección de las víctimas y la compañía que proveía el servicio de agua. Lo rescatable en este caso, es que al final Snow hizo posible que mediante el análisis y relación de los datos se encontrara la respuesta a la causa de esta enfermedad y se pudiera hacer un cambio y evitar que más gente siguiera muriendo [JOHN01]. Podemos decir entonces, que hasta el día de hoy no ha cambiado nada en lo que al análisis de la información se refiere, ya que la extracción de información a través de la correlación de datos se lleva a cabo desde hace mucho tiempo, lo que sí es nuevo hoy en día es el medio a través del cual se extrae la información.

Anteriormente, toda la información que se iba generando y la forma de extraer los datos valiosos, se realizaba de forma manual, hasta que aparecieron las primeras máquinas que agilizaban la labor de las personas (como la máquina de Babbage en 1830) y luego el surgimiento de la era de la computación, cuando en 1937 John Vicent Atanosoff construye el primer prototipo de computador electrónico.

Se empieza entonces, a generar más datos, por lo tanto surge la necesidad de almacenarlos de forma permanente y es cuando aparece el concepto que nos es muy familiar hoy en día: Bases de Datos (1964).

Luego a mediados de los años 80, surge otro problema ¿qué hacer con la información que se almacena?. Esta pregunta todavía es difícil contestarla, ya que gracias a la tecnología de las bases de datos, las empresas e instituciones han ido almacenando gran cantidad de datos, pero éstos no hacen más que ocupar espacio.

Nacen entonces una serie de conceptos que pretenden encontrarle solución a este problema, nuevas tecnologías que tienen un objetivo similar al de los antiguos mineros: buscar en las grandes montañas de información las pepitas de oro (conocimiento).

Y esto es precisamente lo que se quiere lograr con la presente investigación, donde como punto de partida para la utilización de las técnicas de minería de datos y la generación de los índices planteados, se analizarán bases de datos o simplemente datos de instituciones como el Instituto Nacional de Estadística y Censo (INEC), las estadísticas de construcción obtenidas por la Cámara Costarricense de la Construcción

y el Colegio Federado de Ingenieros y Arquitectos, y las estadísticas de enfermedades del Ministerio de Salud.

1.3 Objetivos

1.3.1 Objetivo general

Establecer un grupo de indicadores para el sector vivienda de Costa Rica, utilizando técnicas de Minería de Datos, con el fin de mejorar la toma de decisiones.

1.3.2 Objetivos específicos:

- Analizar algunas fuentes de datos existentes en el sector construcción del país.
- Utilizar herramientas estadísticas y OLAP, para el establecimiento de los indicadores.
- Crear un cubo (OLAP) de información, mediante SQL Server 2000, para la extracción de la información.

CAPITULO 2. Marco Conceptual

2.1 Fuentes de datos

Las fuentes de datos que se analizarán a continuación son las que tendrán relación directa con el desarrollo del trabajo. Como se mencionó en la introducción, uno de los aspectos que tiene mayor relevancia al analizar el sector construcción, es la posible relación que existe entre la construcción de viviendas y los problemas de salud. De allí que se enfocará el análisis en básicamente 3 fuentes: INEC (capítulo vivienda y salud), Ministerio de Salud y Cámara Costarricense de la Construcción, enfocándose además en una zona específica, en este caso la provincia de Cartago.

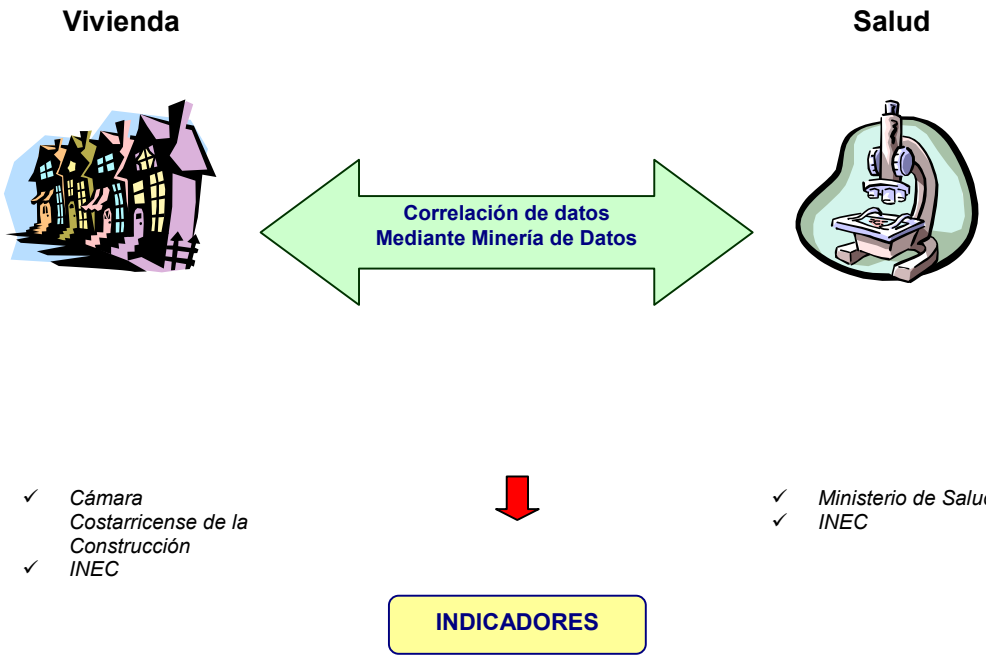


Figura 2-1. Esquema de las fuentes de datos a analizar para obtener los indicadores

2.1.1 Estadísticas del Instituto Nacional de Estadística y Censo (INEC)

El Instituto Nacional de Estadística y Censos (INEC) es el ente rector técnico del Sistema Estadístico Nacional, creado según Ley #7839 del 4 de noviembre de 1998. Con su autonomía funcional y administrativa, suministra diversos servicios, producto de investigaciones estadísticas especializadas.

El Instituto Nacional de Estadística y Censos con la cooperación del Centro Centroamericano de Población (CCP) de la Universidad de Costa Rica ofrece al usuario un servicio de consulta de algunas bases de datos estadísticas. En este servicio es posible obtener tabulaciones de hasta tres variables para una selección de casos determinada. Actualmente están disponibles las bases de datos de los censos de población y vivienda de 1973 y 1984, las estadísticas vitales de 1970 al 2000 y las proyecciones cantonales y nacionales de población. Se puede acceder a la información mediante la siguiente dirección de Internet: www.inec.go.cr

En este servidor se pueden realizar consultas a grandes bases de datos obteniendo los resultados rápidamente. Estas bases de datos son archivos de información original de los individuos de censos, registros vitales y otros similares. Los tabulados pueden obtenerse de manera casi inmediata gracias a una novedosa tecnología desarrollada por: Public Data Queries Inc.

Dentro de la información disponible a los usuarios, se cuenta con un diccionario de datos, el cual resulta valioso para la labor que se desea realizar en el presente trabajo.

JERARQUIA: vivienda			
Variable	**Descripción**	**Rango**	**Códigos**
PROVINCIA	PROVINCIA	1-7	1 San José
CANTON	CANTON	01-20	
DISTRITO	DISTRITO	01-14	
PC	CANTÓN (3 DÍGITOS)	101-706	101 San José
PCD	DISTRITO 5 DÍGITOS	10101-70605	10101 Carmen

Tabla 2-1. Diccionario de datos de los censos nacionales de población y vivienda de Costa Rica, 2000

JERARQUIA: registro			
Variable	**Descripción**	**Rango**	**Códigos**
anio	Año trabajo	1970-2100	
edad	Edad en años simples	0-95	
edadq	Edad en grupos quinquenales	1-20	1 "0-4 años"
sexo	Sexo	1-2	1 Masculino
poblac	Población	0-999999	

Tabla 2-2. Diccionario de datos de las proyecciones nacionales de población de Costa Rica, 1970-2100

JERARQUIA: registro			
Variable	**Descripción**	**Rango**	**Códigos**
provinc	Provincia de residencia	1-7	1 San José
canton	Canton de residencia (3 dígitos)	101-706	101 San José
anio	Año trabajo	1970-2015	
edadq	Edad en grupos quinquenales	1-99	1 0-4 años
sexo	Sexo	1-2	1 Masculino
poblac	Población	0-999999	
subCan	Sub Cantón	1010-7060	
region	Región	1-6	1 Metro S José
subreg	Sub Región	1-22	1 San José

Tabla 2-3. Diccionario de datos de las proyecciones cantonales de población de Costa Rica, 1975-2015

Definición de las sub-regiones de Costa Rica (de acuerdo al INEC)

Otro aspecto que es importante considerar es la división regional que se utiliza para la generación y análisis de información . Esta división se debe comparar posteriormente con la utilizada en las otras fuentes de datos a analizar.

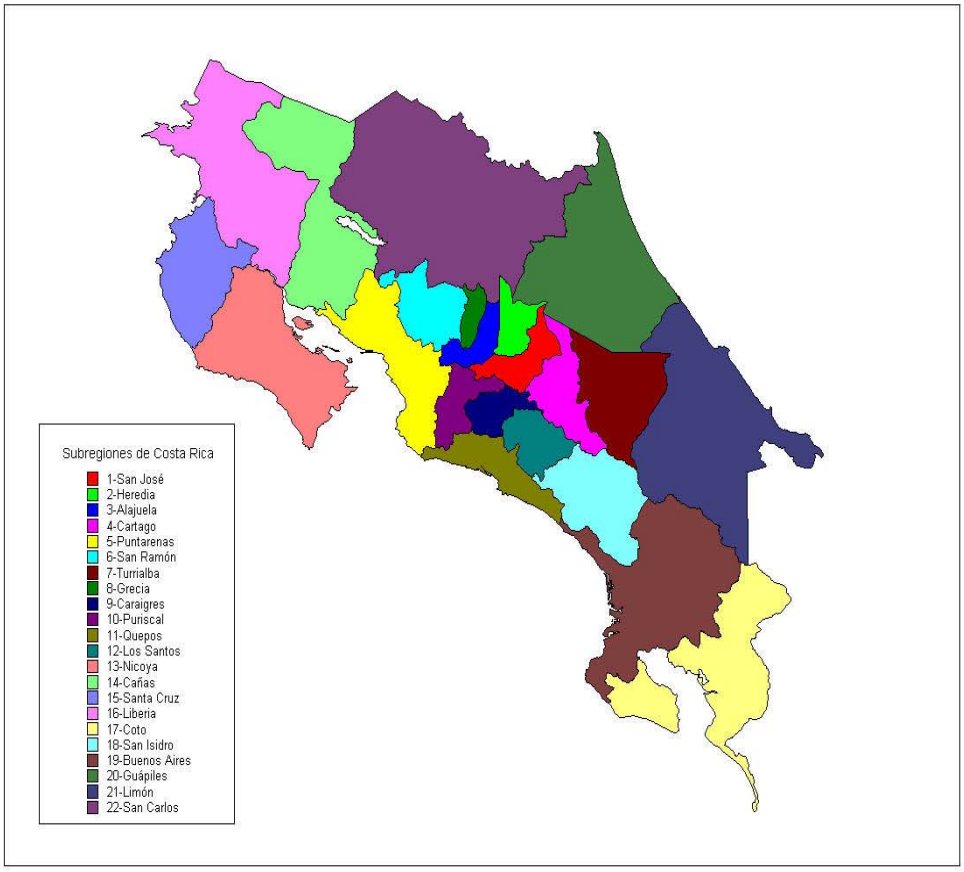


Figura 2-4. División de las subregiones de Costa Rica según el INEC

1. San José	Toda la provincia de San José menos: el cantón de Turrubares los distritos Fráiles, San Cristobal y Rosario del cantón de Desamparados todos los distritos del cantón de Aserrí excepto el propio Aserrí el cantón de Acosta el cantón de Puriscal los distritos Guayabo, Tabarcia, Picagres y Piedras Negras de Mora el distrito Palmichal del cantón de Acosta los cantones Tarrazú, Dota, Pérez Zeledón y León Cortés
2. Heredia	Toda la provincia de Heredia menos: el cantón de Sarapiquí el distrito de Varablanca del cantón de Heredia
3. Alajuela	Los cantones de Alajuela, Atenas y Poás de la provincia de Alajuela
4. Cartago	Los cantones de Cartago, Paraíso, Alvarado, Oreamuno y El Guarco de la Provincia de Cartago
5. Puntarenas	Los cantones de Puntarenas, Esparza, Montes de Oro y Garabito de la Provincia de Puntarenas y los cantones de San Mateo y Orotina de la provincia de Alajuela el cantón de Turrubares de la provincia de San José
6. San Ramón	Los cantones de San Ramón, Naranjo, Palmares y Alfaro Ruiz de la provincia de Alajuela
7. Turrialba	Los cantones de Jiménez y Turrialba de la provincia de Cartago
8. Grecia	Los cantones de Grecia y Valverde Vega de la provincia de Alajuela
9. Caraigres	De la provincia de San José: Los distritos Fráiles, San Cristobal y Rosario del cantón de Desamparados Todos los distritos del cantón de Aserrí excepto el propio Aserrí El cantón de Acosta El distrito de Corralillo del cantón central de Cartago
10. Puriscal	De la provincia de San José: El cantón de Puriscal Los distritos Guayabo, Tabarcia, Picagres y Piedras Negras de Mora El distrito Palmichal del cantón de Acosta
11. Quepos	Los cantones de Aguirre y Parrita de la provincia de Puntarenas
12. Los Santos	De la provincia de San José los cantones de Tarrazú, Dota y León Cortés
13. Nicoya	De la provincia de Guanacaste los cantones de Nicoya, Nandayure y Hojancha Los distritos de Lepanto, Paquera y Cóbano de la provincia de Puntarenas
14. Cañas	Los cantones de Cañas, Abangares y Tilarán de la provincia de Guanacaste El cantón de Upala de la provincia de Alajuela
15. Santa Cruz	Los cantones de Santa Cruz y Carrillo de la provincia de Guanacaste

Figura 2-5. Subregiones de Costa Rica, según el INEC

Información disponible en el INEC para analizar

Dentro de la información disponible en forma tabular (en formato Excel) en el sitio de Internet se tiene lo siguiente[2]:

Encuesta a Hogares [INEC01]

(Sólo se incluye la información considerada relevante para el trabajo)

- Población total por condición de actividad y tasas, según región de planificación y sexo. (C01) 1999-2001
- Población de 12 años o más por condición de actividad y tasas, según zona, sexo y grupos de edad. (C03) 2001
- Distribución relativa de los hogares por declaración de ingreso 1987-2001. (C35)
- Distribución relativa de los hogares con ingreso conocido por nivel de pobreza, según región de planificación y año 1987-2001. (C36)
- Principales características de los hogares y de las personas con ingreso conocido por nivel de pobreza según zona 2001. (C34)
- Principales características de los hogares y de las personas con ingreso conocido por nivel de pobreza según región de planificación 2001. (C33)
- Total de hogares y personas en hogares por número de miembros en el hogar, según región de planificación 2001. (C31)
- Población inactiva por tipo de inactividad, según sexo y grupos de edad 2001. (C30)
- Población ocupada con empleo pleno e ingreso primario mensual conocido por sexo, según rama de actividad, grupo ocupacional, categoría ocupacional y sector industrial 2001. (C25)
- Población ocupada con ingreso conocido e ingreso promedio mensual total por categoría ocupacional, según región de planificación y sexo 2001. (C17)

Adicional a estas tablas de datos en la publicación “IX Censo Nacional de Población y V de Vivienda del 2000: Resultados Generales del INEC”, se encuentra la siguiente información de interés [INEC01]:

[2] Ver anexo N.1

- Población total por sexo, según grupos quinquenales de edad, 2000. Cuadro 4.
- Cantones con mayor crecimiento relativo de población, 1984 al 2000. Cuadro 5.
- Cantones con menor crecimiento relativo de población, 1984 - 2000. Cuadro 6.
- Cantones con densidad de población superior a mil habitantes por kilómetro cuadrado, 2000. Cuadro 8.
- Cantones con densidad de población inferior a 20 habitantes por kilómetro cuadrado, 2000. Cuadro 9.
- Tasas de discapacidad por sexo según tipo de discapacidad, 2000. Cuadro 15.
- Total de viviendas por ocupación de la vivienda y promedio de ocupantes según provincia, 2000. Cuadro 16.
- Viviendas individuales ocupadas según disponibilidad de servicios básicos, 1984 – 2000. Gráfico 8.
- Total de viviendas individuales ocupadas que tienen un artefacto o más y sus relativos, 1984 – 2000. Cuadro 17.
- Población total por sexo, total de viviendas por ocupación de la vivienda y promedio de ocupantes según provincia, cantón y distrito. Cuadro de resultados 1.
- Población total por grupos de edad según provincia, cantón y sexo. Cuadro de resultados 2.
- Población total por zona y sexo según provincia y cantón. Cuadro de resultados 3.
- Población total por provincia o lugar de nacimiento según provincia de residencia actual y sexo. Cuadro de resultados 4.
- Población total por tipo de discapacidad según sexo y grupos de edad. Cuadro de resultados 7.
- Población de 5 años y más por condición de asistencia a la educación regular y sexo según zona y grupos de edad. Cuadro de resultados 8.
- Total de viviendas individuales ocupadas por tipo de vivienda según zona, abastecimiento y procedencia de agua, servicio sanitario y tenencia de electricidad. Cuadro de resultados 11.

Otra información que se tiene disponible en el INEC es:

- Total de viviendas ocupadas y total de ocupantes. Por tipo de vivienda, según zona.
- Total de viviendas ocupadas y promedio de ocupantes. Por tipo de vivienda, según zona.
- Total de viviendas ocupadas, total de ocupantes y promedio de ocupantes por viviendas. Por tipo de abastecimiento de agua, según zona.
- Total de viviendas ocupadas, total de ocupantes y promedio de ocupantes por viviendas. Por procedencia de agua, según zona.
- Total de viviendas ocupadas y promedio de ocupantes por viviendas. Por tenencia servicio de baño y tipo de vivienda, según zona y tipo de vivienda.

2.1.2 Ministerio de Salud

El Ministerio de Salud es el ente rector en materia de salud en Costa Rica y tiene como misión garantizar que la producción social de la salud se realice en forma eficiente y eficaz, mediante el ejercicio de la Rectoría, con plena participación de los actores sociales para contribuir a mantener y mejorar la calidad de vida de la población y el desarrollo del país, bajo los principios de equidad, solidaridad y universalidad [MSCR00].

Definición de las Regiones de Costa Rica según Ministerio de Salud

El Ministerio de Salud para realizar su trabajo ha dividido el país en 9 regiones: Chorotega, Huetar Norte, Central Occidental, Central Norte, Central Sur, Central Este, Pacífico Central, Huetar Atlántica y Brunca. En la siguiente figura se puede observar esa división [MSCR00].

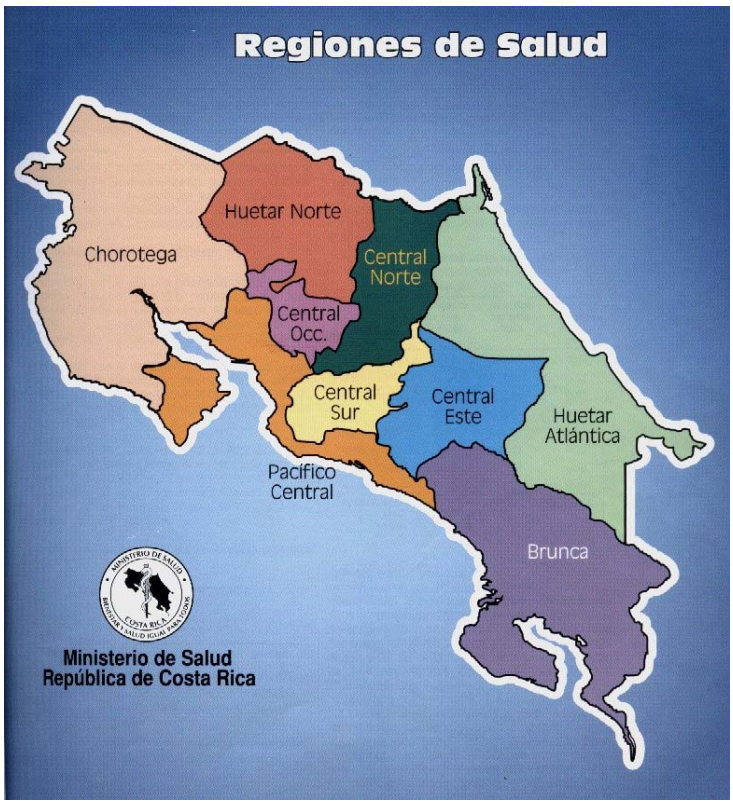


Figura 2-6. División de las Regiones de Costa Rica según el Ministerio de Salud

Información disponible en el Ministerio de Salud para analizar

Dentro de la información disponible en forma tabular con que cuenta el Ministerio de salud y que es de interés para este análisis, se encuentra;

- Mortalidad por los cinco grandes grupos de causas.
- Casos registrados de enfermedades de declaración obligatoria según causa específica por año de ocurrencia. (Importante sería tenerlos por zona o provincia).
- Cobertura Nacional de la Evaluación de la Atención Integral (Cantidad de Ebais). (Importante ubicar Ebais por cantones y distritos).

2.1.3 Cámara Costarricense de la Construcción

La Cámara Costarricense de la Construcción es una asociación civil de interés colectivo que tiene por objeto el fomento, desarrollo, protección y defensa de la industria de la

construcción, de sus industrias conexas, de los intereses profesionales de sus asociados y el armonioso desarrollo económico y social del país.

Creada bajo el amparo de la Ley General de Asociaciones, la Cámara es una organización privada con personalidad jurídica propia, con jurisdicción en todo el territorio nacional, con estatutos legalmente aprobados, que definen su campo de acción, y que le dan forma y estructura.

Están asociadas a la Cámara varios tipos de empresas directamente relacionadas con las actividades de la industria de la construcción; tales como las de consultoría en arquitectura e ingeniería, en construcción de obras de todo tipo, fabricantes de materiales de construcción y comerciantes de esos materiales.

Información disponible en la Cámara para analizar

Estadísticas (Información y gráficos):

- Total de vivienda construida
- Sector vivienda
- Sector comercio y oficinas
- Sector industria
- Otros sectores

Indices de Construcción:

- Indice de vivienda y edificios
- Indice de acueductos y alcantarillados
- Indice de urbanización
- Indice de mano de obra

Gráficos

- Variación del PIB Total vrs PIB de Construcción
- Variación Area de Construcción vrs PIB de Construcción
- Variación Indice de Edificios vrs PIB de Construcción
- Variación del tipo de cambio (venta) vrs PIB de Construcción
- Variación del IMAE vrs PIB de Construcción

2.1.4 Información disponible en otras instituciones para analizar

Observatorio del Desarrollo Universidad de Costa Rica:

- Porcentaje de la población total, extensión y densidad por regiones de planificación.

Ministerio de Planificación Nacional y Política Económica:

MIDEPLAN, a través del tiempo, ha logrado conformar una extensa red de vínculos institucionales que han facilitado el intercambio de la información que se genera a nivel nacional. Recientemente se ha avanzado en la sistematización de la información cuantitativa, con la creación por parte de MIDEPLAN de una base de datos que contiene las principales variables e indicadores sociodemográficos, económicos y ambientales.

De la unión de esos dos elementos, la base de datos y la red de enlaces institucionales, nace el Sistema de Indicadores sobre Desarrollo Sostenible (SIDES), con los siguientes objetivos:

- Contribuir a la difusión de información que permita ampliar y profundizar el análisis del desarrollo nacional por parte de los diferentes actores sociales.
- Servir de enlace entre productores y usuarios de información.
- Avanzar en la elaboración de indicadores agregados sobre desarrollo sostenible.
- Como parte de la estrategia seguida para garantizar el logro de los objetivos propuestos, se definieron mecanismos tanto de intercambio como de divulgación de la información, entre los que se encuentran la red de cómputo institucional, publicaciones específicas y este medio.

En este sistema, específicamente en el capítulo de vivienda y servicios básicos, se puede encontrar información referente a:

- Condición de vivienda, según tipo de tenencia y zona.
- Viviendas con disponibilidad de servicio de agua.

- Vivienda con sistema de disposición de excretas, según tipo de disposición y zona.
- Habitantes con disponibilidad de servicio de agua, según tipo de servicio y zona.
- Habitantes con sistema de disposición de excretas, según tipo de servicio y zona.
- Estado físico de las viviendas, según estado de los materiales y zona.
- Hacinamiento y déficit habitacional, según zona.
- Construcción: área construida, valor de las viviendas, índice de precios y participación en el PIB.
- Bono familiar de vivienda: número otorgado, inversión realizada y montos.

2.2 Herramientas de análisis

2.2.1 Minería de Datos

La minería de datos, en su definición más simple, es un conjunto de técnicas que permite extraer información a partir de un conjunto o depósito de datos. Se puede decir que la minería de datos es un proceso que se lleva a cabo en diferentes pasos o etapas y que permite hallar o extraer conocimiento de una o varias bases de datos. Se analizan los datos existentes en estas bases desde varias perspectivas y se les transforma a través de herramientas en información útil. Esta información va normalmente orientada a los análisis de rentabilidad, incremento de ganancias o reducción de costos en los negocios [DEHA01]

Al igual que la mayoría de nuevas tecnologías o desarrollos, la minería de datos nace por una necesidad. Una vez que surgen las bases de datos, las organizaciones empiezan a acumular estos datos en grandes cantidades y se convierten en la mayoría de los casos en “tumbas” de datos que no hacen nada más que estar allí ocupando espacio. Surge entonces la necesidad de ser cada vez más competitivos y para ello es necesario contar con información para la toma de decisiones, y esa información en

muchos casos se basa en patrones de compras, tendencias y proyecciones, entre otros. Por lo tanto se hace necesario contar con herramientas que permitan obtener lo que se necesita, siendo la minería de datos una de ellas.

Si se analiza la minería de datos como un paso en el proceso de descubrimiento del conocimiento, es necesario identificar una serie de etapas para lograr ese objetivo:

- Primero es necesario efectuar una limpieza de los datos, esto implica la eliminación del ruido que puedan tener y eliminar inconsistencias.
- Luego se deben integrar los datos, ya que en una organización las fuentes se datos generalmente son heterogéneas, provienen de diversos motores de bases de datos o de sistemas diferentes.
- No todos los datos son importantes, por lo tanto se deben seleccionar y recuperar los datos relevantes.
- Una vez seleccionados los datos deben transformarse y consolidarse.
- Lo siguiente es utilizar alguna de las técnicas de minería de datos para obtener la información que se necesita.
- Una vez obtenida la información es importante analizarla o sea identificar y evaluar los patrones.
- Finalmente se debe representar esa información utilizando alguna técnica de visualización o representación de información.

Las técnicas de la minería de datos son el resultado de un largo proceso de investigación y va más allá del acceso y la retrospectiva de los datos, hacia la entrega de información prospectiva y proactiva. Se soporta en tres tecnologías, las cuales han sido objeto de estudio (y lo siguen siendo):

- **Recolección masiva de datos**
- **Potentes computadoras con multiprocesadores**
- **Algoritmos de Data Mining**

2.2.2 Requerimientos que debe cumplir un sistema de Minería de Datos

1. ***Manejar diferentes tipos de datos***

Ya que las aplicaciones que contienen los datos que se desean analizar generalmente son diferentes, un sistema de descubrimiento de conocimiento debe ser capaz de realizar minería sobre datos distintos. Sin embargo, se debe ser realista en el sentido de que no se puede esperar que un solo sistema de minería puede manejar todos los tipos de datos existentes.

2. ***Eficiencia y escalabilidad de los algoritmos para realizar la minería de datos***

Los algoritmos para el descubrimiento de conocimiento deben ser eficientes y escalables en grandes bases de datos. Esto significa, que el tiempo de corrida debe ser predecible y aceptable. Algoritmos de comportamiento exponencial o de complejidad polinomial no son prácticos.

3. ***Utilidad, certeza y expresividad de los resultados de la minería***

Esto motiva un estudio sistemático que mida la calidad del conocimiento descubierto, incluyendo análisis de intereses y rentabilidad, mediante la elaboración de estadísticas, análisis, modelos y herramientas de simulación.

4. ***Expresar de diferentes formas las respuesta y resultados de la minería***

Las tareas de minería de datos pueden ser especificadas por personas no expertas, pero el conocimiento generado debe ser fácil de comprender y poder ser utilizado inmediatamente. Por lo tanto se requiere un sistema de descubrimiento que utilice técnicas de representación fáciles de entender.

5. ***Interactividad con el conocimiento descubierto y múltiples niveles de abstracción***

Como es difícil predecir el nivel que se espera al extraer información, se debe permitir que el usuario interactúe de manera flexible, refinando las respuestas, haciendo cambios dinámicos de enfoque y profundizando en la información.

6. ***Obtener información de diferentes fuentes de datos***

Se debe tratar de que se pueda realizar minería de datos desde fuentes diferentes, esto implica el desarrollo de algoritmos para minería de datos sobre datos paralelos y distribuidos, ya que se sabe que generalmente representan una complejidad computacional.

7. ***Protección, privacidad y seguridad de los datos***

Es importante estudiar cuando el conocimiento descubierto puede invadir la privacidad y cuales medidas de seguridad pueden desarrollarse para prevenir un mal uso de la información [CHEN96] .

La tecnología de bases de datos se ha utilizado con gran éxito en el procesamiento tradicional de datos en los procesos de negocio. Actualmente ha crecido el interés de ampliar el dominio de esta tecnología y sacarle un mayor provecho. Una de la aplicaciones nuevas es la minería sobre esas bases de datos, uso que se considera, será de suma importancia para mejorar la toma de decisiones en las compañías. Ya que las organizaciones han venido acumulando gran cantidad de datos sobre clientes, transacciones, ventas, etc., sin darse cuenta del valor potencial de la información que se puede extraer de ellos [AGRA99] .

Desafortunadamente, las bases de datos que existen actualmente ofrecen poca funcionalidad para el soporte de aplicaciones de minería, en otros casos, las técnicas utilizadas, tienen un comportamiento poco eficiente cuando las bases de datos son de tamaño considerable. Esto se ha convertido en una de las razones por las cuales todavía existen una gran cantidad de bases de datos inexploradas [AGRA99] .

Antes de empezar a exponer las principales técnicas de Minería de Datos y cuáles se utilizarán para correlacionar los datos y obtener la información de deseamos (indicadores), es muy importante conocer el concepto de Datawarehouse.

2.2.3 Datawarehouse

Un *datawarehouse* es un *depósito de datos* y tiene gran relación con los conceptos de minería de datos, ya que es de este depósito de donde se extrae la información para ser analizada. Una definición simple de lo que es un datawarehouse sería: "una colección de datos de los sistemas transaccionales de una empresa en un solo sistema, una de las utilidades que tiene es que muestra a los usuarios información global de la empresa o compañía" [FLAN01] .

Los *datawarehouse*, se originan a partir de datos comunes e históricos usados y generados por los sistemas de producción más algunos datos adicionales de fuentes externas. Los usuarios de los datawarehouse, podrían, mediante técnicas de minería de datos:

- Tomar decisiones de una manera "razonada"
- Tomar justo a tiempo medidas correctivas
- Realizar planeamientos exitosos, entre otros.

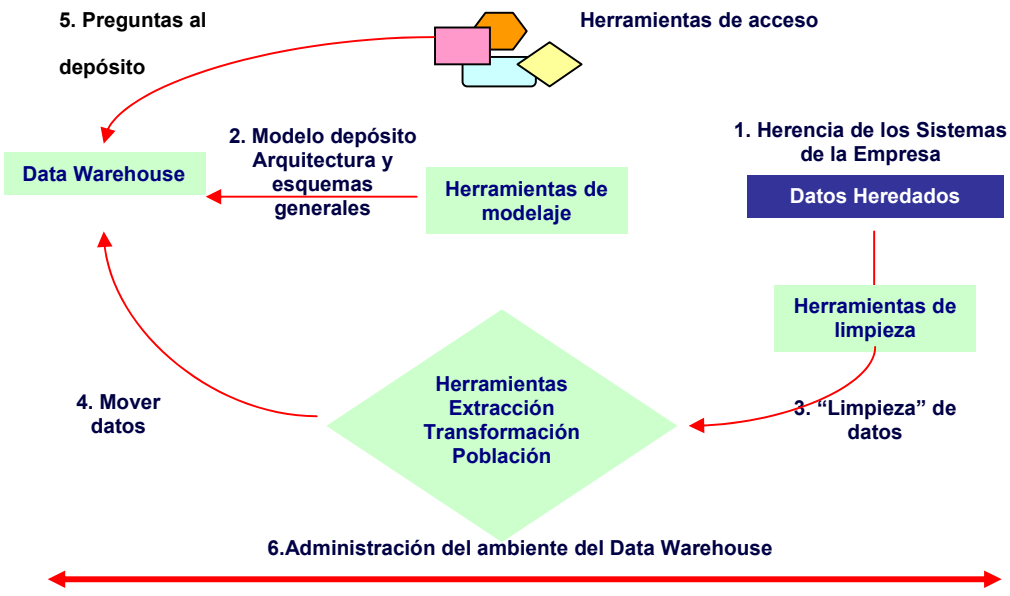


Figura 2-7. Ciclo de vida de un datawarehouse3

De manera muy simple, la figura 2-7 muestra el ciclo de vida de un *datawarehouse* o *depósito de datos*. El principio de este depósito se basa en los datos existentes o heredados de sistemas transaccionales, éstos son la principal fuente de información, de allí que sea importante comprender sus características. Una vez identificadas las fuentes de datos en la empresa, es necesario desarrollar un plan de diseño para el datawarehouse, para lo cual se requiere definir la estructura del negocio y sus procesos. La forma en que se modele el negocio es fundamental, ya que se necesita que sea dinámico para que la empresa pueda ajustarse a las cambiantes necesidades del mercado.

El siguiente paso es uno de los más difíciles y el que muchas veces hace que se deseche la idea de un datawarehouse en una empresa, consiste en *"limpiar"* los datos de las fuentes existentes. Como se mencionó, los datos que se quieren almacenar se toman de diferentes sistemas de la compañía y la gran mayoría de las veces estos sistemas se encuentran almacenados en bases de datos diferentes y con lenguajes de programación también diferentes. El proceso de limpieza tiene como objetivo validar y corregir los datos existentes, de tal manera que se asegure la consistencia y calidad de los mismos.

Una vez que los datos están "limpios", se procede a mover los datos hacia el *datawarehouse*, el primer paso para ello consiste en *extraer* los datos de los sistemas existentes en la empresa, es importante tratar de derivar las reglas del negocio y definiciones que contienen esos datos dentro del sistema, luego los datos deben ser *transformados* de tal manera que se conviertan en un "blanco" para obtener información a partir de ellos, y como última etapa de este movimiento se debe *poblar* el datawarehouse.

[3] Realizado con base a información contenida en la referencia [FLAN98]

Estos datos que son replicados y almacenados en el datawarehouse utilizando alguno de los motores de bases de datos existentes: *Relacional*, como: Oracle, Informix, Sybase, *Multi-dimensional*, como: Arbor / Essbase, *Propietarios*, como: Red Brick, *Híbridos*, como: DB2/Arbor.

Se conoce como metadato la información que se usa para referirse a la información sobre los datos que han sido capturados y almacenados dentro del depósito, una buena administración de estos metadatos permitirá que éstos se reutilicen para generar nuevos datos.

Finalmente, una vez que se tiene el datawarehouse, se debe empezar a utilizar, accesando a los datos, haciendo preguntas, imprimiendo reportes. Además, usando las herramientas adecuadas, se puede mejorar la toma de decisiones, encontrar patrones, etc.

Una de las herramientas para obtener información de ese depósito es haciendo minería sobre los datos almacenados, utilizando para ello diversas técnicas, de las cuales se hablará posteriormente [FLAN98].

2.2.4 Técnicas de Minería de Datos

La tecnología de bases de datos se ha utilizado con gran éxito en el procesamiento tradicional de datos en los procesos de negocio. Actualmente ha crecido el interés de ampliar el dominio de esta tecnología y sacarle un mayor provecho. Una de la aplicaciones nuevas es la minería sobre esas bases de datos, uso que se considera, será de suma importancia para mejorar la toma de decisiones en las compañías. Ya que las organizaciones han venido acumulando gran cantidad de datos sobre clientes, transacciones, ventas, etc., sin darse cuenta del valor potencial de la información que se puede extraer de ellos [AGRA99] .

Desafortunadamente, las bases de datos que existen actualmente ofrecen pocas funcionalidades para el soporte de aplicaciones de minería, en otros casos, las técnicas utilizadas, tienen un comportamiento poco eficiente cuando las bases de datos son de tamaño considerable. Esto se ha convertido en una de las razones por las cuales todavía existen una gran cantidad de bases de datos inexploradas [AGRA99] .

El proceso de descubrimiento de conocimiento es un proceso que se realiza una vez que los datos han sido extraídos y almacenados en el datawarehouse, siendo esta una de las tareas más difíciles. Para obtener la información que se necesita se hace uso de diversas *técnicas*, las tareas que realizan estas técnicas se pueden dividir en dos categorías, las *descriptivas* que caracterizan las propiedades generales de los datos en la base de datos, y las *predictivas* que realizan inferencias sobre los datos reales para poder hacer pronósticos.

A continuación, se presentará una descripción de algunas técnicas y modelos que han sido desarrollados para enfrentarse a los problemas de minería. La técnica a usar dependerá de los datos que se quieran analizar.

Asociación

Considere un supermercado que posee una base de datos, que almacena las compras que realiza cada cliente. Podría resultar interesante para este supermercado saber que productos se compran juntos y el grado de confianza de estas conclusiones.

Para introducir el problema de descubrir reglas de asociación en una base de datos, se puede suponer que se tiene un conjunto de transacciones, donde cada una de ellas contiene una serie de ítemes. Una *regla de asociación* sobre estos es una expresión de la forma:

$$X \Rightarrow Y \quad \textit{donde X y Y son un conjunto de ítemes}$$

De manera intuitiva, se puede decir que una regla de asociación es una transacción en una base datos, la cual al contener X tiende a contener Y.

Un ejemplo de una regla de asociación es el siguiente:

"El 30% de los tiquetes de compra que contienen cerveza, también contienen pañales; y el 2% de todos los tiquetes de compras contienen ambos productos".

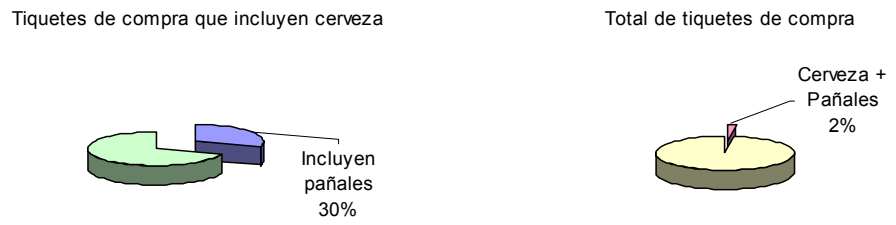


Figura 2-8. Representación grafica de una regla de asociación

En este caso el 30% representa el *nivel de confianza* de la regla y el 2% el *soporte* de la misma.

El problema, entonces, es encontrar todas las reglas de asociación que satisfagan un mínimo de confianza y soporte especificado por el usuario. [ARNI96]

Este tipo de técnica encuentra su aplicación en el descubrimiento de afinidades en compra de productos en supermercados, diseño de catálogos, segmentación de clientes y patrones de compras entre otros.

Algoritmo A priori

El problema de minería utilizando reglas de asociación, se puede descomponer en dos sub-problemas:

- **Encontrar todas las combinaciones de ítemes en una transacción que tengan el soporte mínimo esperado. Esto será la frecuencia de los ítemes.**
- **Usar la frecuencia de ítemes para generar las reglas deseadas.**

Expuesto de otra forma, suponga que se tienen los ítems ABCD, y AB es un conjunto de ítemes frecuentes (cumplen con el mínimo soporte esperado), una vez que se tiene esto, lo siguiente es determinar si la regla AB ⇒ CD se cumple con el grado de confiabilidad que se requiere (ver figura 2-9).

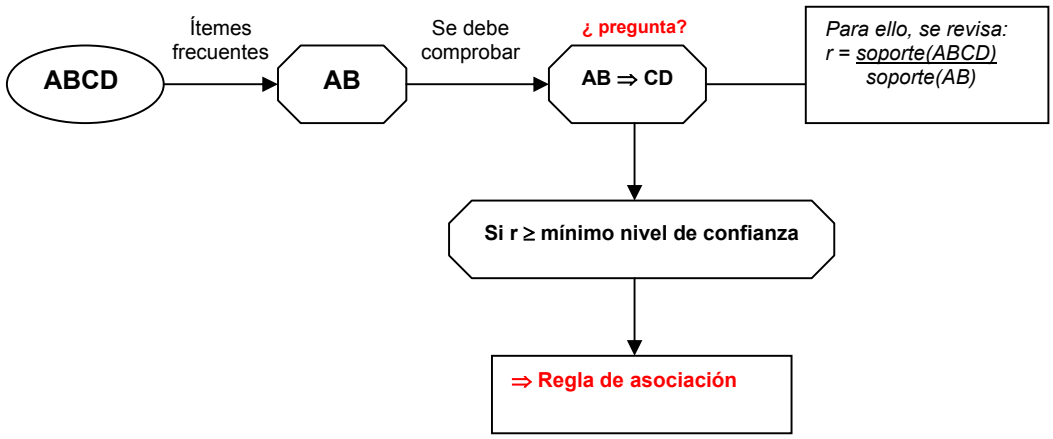


Figura 2-9. Representación gráfica de cómo se resuelve una regla de asociación

La compañía IBM, ha desarrollado un proyecto: Quest, el cual tiene como principal objetivo desarrollar tecnología que permita el tener disponibles la información necesaria para la toma de decisiones. En este proyecto de utiliza el algoritmo Apriori[4], para todas las frecuencias de ítemes en un conjunto de datos.

[4] Agrawal, R. And Srikant, R. 1994. Fast Algorithms for Mining Association Rules. *In Proc. of the 20th Int´l Conference on Very Large Databases.*

En la figura 2-10, se muestra el algoritmo. El primer paso de éste consiste en un simple conteo de la ocurrencia de cada ítem 1-itemsets (conjunto de ítems con un término); el siguiente paso se compone de dos fases, la primera genera los ítems candidatos C_k, usando la función a priori-gen(), la segunda realiza una función de filtro. Luego el algoritmo busca en la base de datos para cada transacción, cual de los candidatos contenidos en C_k están en la transacción.

```
procedure AprioriAlg()
begin
        L1 := { frequent 1-itemsets };
        For (k := 2; Lk-1 ≠ 0; k++) do {
                Ck := a priori-gen(Lk-1); //Nuevos candidatos
                forall transactions t in the dataset do {
                        forall candidates c ∈ Ck contained in t do
                                c.count ++;
                }
                Lk := { c ∈ Ck | c.count ≥ min-support }
        }
        Answer := ∪k Lk;
end
```

Figura 2-10. Algoritmo Apriori

Generalizaciones

Frecuentemente el uso de taxonomías sobre los ítemes es posible. En la figura 2-8 se muestra un ejemplo. Los usuarios frecuentemente están interesados en generar reglas que soporten diferentes niveles de taxonomías, sin embargo este no es un problema fácil de resolver.

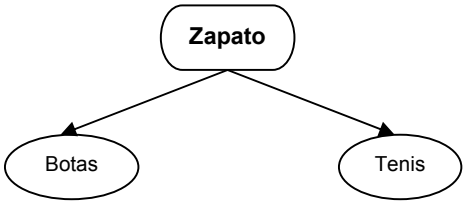


Figura 2-11. Ejemplo de taxonomía

Caracterización de datos

Los datos y objetos almacenados en las bases de datos usualmente contienen información detallada y en diferentes niveles.

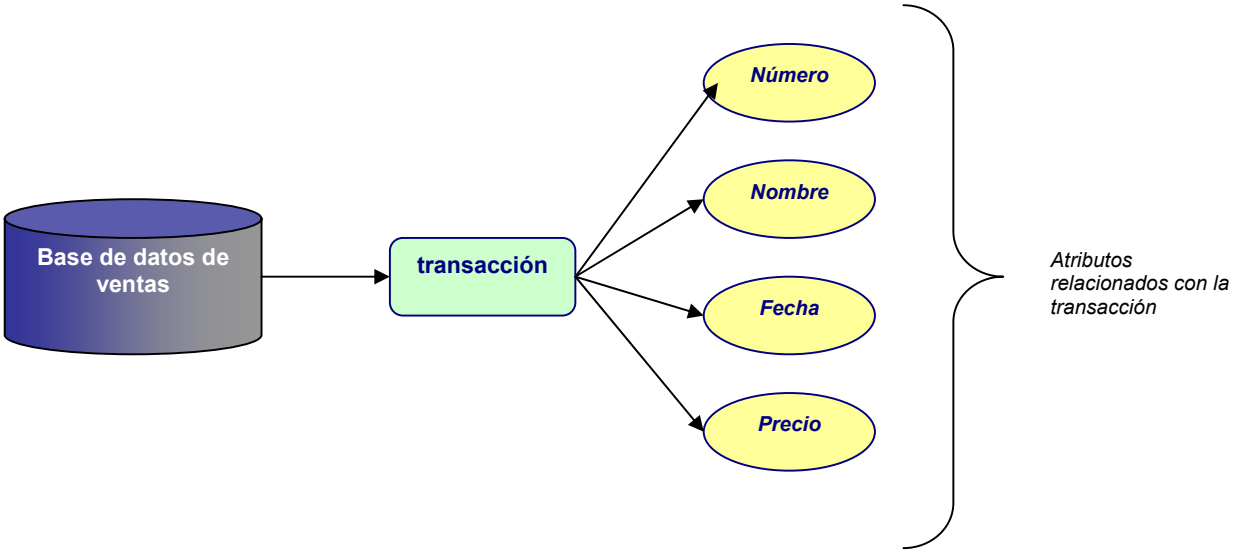


Figura 2-12. Ejemplo de información asociada a los datos

En la figura 2-12, se puede observar cómo un dato en una base de datos de ventas, puede tener asociados una serie de atributos. Mucha veces se desea presentar la información en forma sumarizada y general, y es aquí donde la técnica de minería de datos llamada caracterización es muy útil.

La caracterización o generalización, es un proceso que extrae un conjunto de datos relevantes de una base de datos desde un nivel bajo a un nivel más alto. La forma de realizar esta generalización , se puede hacer de dos formas:

- **Acceso a cubos de datos**
- **Inducción orientada a atributos**

Agrupamiento

La técnica de agrupamiento es una técnica descriptiva que busca identificar grupos homogéneos de objetos basados en los valores de sus atributos (dimensiones). Esta técnica ha sido estudiada intensamente en estadística, reconocimiento de patrones y aprendizaje de máquinas.

Las actuales técnicas de agrupamiento pueden clasificarse en dos amplias categorías: particionadas y jerárquicas. El agrupamiento por partes toma una partición de los objetos dentro del grupo de tal manera que los objetos en el grupo sean más parecidos entre ellos, que entre otros grupos. El conocido método *k-means and k-medoid* determina el grupo k representativo y asigna cada objeto al grupo que está representativamente más cercano, tal que la suma de los cuadrados de la distancia entre los objetos y sus representaciones sea mínima.

El agrupamiento jerárquico es una secuencia de particiones jerarquizadas, un conglomerado. Este empieza colocando cada objeto en su propio grupo y después ordena estos grupos atómicos en largos grupos hasta que todos los objetos estén en un solo grupo. Revertiendo luego el proceso y subdividiendo el grupo en pequeñas piezas. [AGRA96]

Requerimientos de los algoritmos de la técnica de agrupamiento

Efectivo tratamiento de alta dimensionalidad

Un objeto (dato almacenado) típicamente tiene una docena de atributos y el dominio para cada atributo puede ser extenso. Esto no significa buscar por grupo en cada espacio unidimensional como un promedio de la densidad de cada punto en cualquier parte del espacio unidimensional, ya que esto representa una confianza muy baja. Por lo tanto, funciones de distancia que utilicen todas las dimensiones de los datos pueden

ser no efectivas. Sin embargo, muchos grupos pueden existir en diferentes subespacios conformados por combinaciones diferentes de atributos.

Interpretación de los resultados

Las aplicaciones de minería de datos requieren de descripciones que sean fáciles de asimilar por parte del usuario final, la visualización y las explicaciones son de gran importancia. Es muy importante tener formas simples de representar la información, ya que la mayoría de las técnicas no trabajan bien en espacios de muchas dimensiones.

Escalabilidad y usabilidad

Las técnicas de agrupamiento deben ser rápidas y escalables con el número de dimensiones y el tamaño de los datos. Debe ser insensible al orden en el cual los datos son almacenados. [AGRA96]

Algoritmo CLIQUE[5]

Este algoritmo encuentra automáticamente subespacios con grupos de alta densidad; este produce idénticos resultados independientemente del orden en el cual fueron almacenados y presentados los datos. Genera descripciones de grupos en la forma de DNF y se esfuerza por generar descripciones mínimas para una fácil comprensión. Evaluaciones empíricas muestran que CLIQUE escala linealmente con el número de registros y tiene una buena escalabilidad relacionada con el número de dimensiones (atributos) de los datos, o con la mayor dimensión en la cual los grupos están almacenados.

Generalmente estamos interesados en identificar automáticamente subespacios en un conjunto de datos con diferentes dimensiones. Restringiendo nuestra búsqueda a solamente subespacios en el espacio original, es decir, sin crear nuevas dimensiones (por ejemplo combinaciones lineales de las dimensiones originales). Esta restricción es importante porque permite simplificar y comprender mejor los resultados finales. Cada una de las dimensiones originales tiene un significado real para el usuario, mientras que una combinación lineal de dimensiones puede ser difícil de interpretar. [FAYY96]

[5] De CLustering In QUEst, del proyecto de investigación de IBM Almaden

Utilizando una densidad basada en aproximaciones a grupos, donde un grupo es una región que tiene una mayor densidad de puntos que las regiones que la rodean. El problema es identificar automáticamente estas proyecciones de los datos en un conjunto de atributos.

La técnica de agrupamiento CLIQUE consiste en aplicar los siguientes pasos:

- Identificación de los subespacios que contienen grupos
- Identificación de los grupos
- Generación de la mínima descripción de los grupos

Descripción del problema

Sea $A = \{A_1, A_2, ..., A_d\}$ un conjunto de borde, dominios totalmente ordenados y $S = A_1 \times A_2 \times ... \times A_d$ un espacio numérico d-dimensional. En otras palabras $A_1 ..., A_d$ son las dimensiones (atributos) de S.

La entrada de datos consiste en un conjunto de puntos d-dimensionales $V = \{v_1, v_2, ..., v_m\}$ donde $v_i = \langle v_{i1}, v_{i2}, ..., v_{id} \rangle$.

Dividimos el espacio de datos en unidades rectangulares no traslapadas. Las unidades son obtenidas mediante divisiones de cada dimensión en el total de intervalos de igual longitud, el cual es un parámetro de entrada.
Cada unidad "u" resulta de la intersección de un intervalo con un atributo. Esta intersección es de la forma $\{u_1, ..., u_d\}$ donde $u_i = [l_i, h_i)$ es un intervalo abierto a la derecha dentro de la partición A_i.

Decimos entonces que un punto $v = \langle v_1, ..., v_d \rangle$ está contenido en una unidad $u = \{u_1, ..., u_d\}$ si $l_i \le v_i < h_i$ para todos los u_i. La selección de una unidad es definida para ser una fracción del total de puntos de datos contenidos en la unidad.
Se llama a una unidad *u dense* si la selección (u) es mayor que τ, donde la *density threshold* τ es otro parámetro de entrada.

De manera similar se definen las unidades en el espacio d-dimensional original.

Un grupo es el máximo conjunto de unidades densas conectadas en k-dimensiones. Dos unidades k-dimensionales se encuentran conectadas si tienen una cara en común. Una región en una dimensión k es un conjunto de k-dimensiones de ejes rectangulares paralelos.

Una región R está contenida en un grupo C si $R \cap C = R$. [AGRA96]

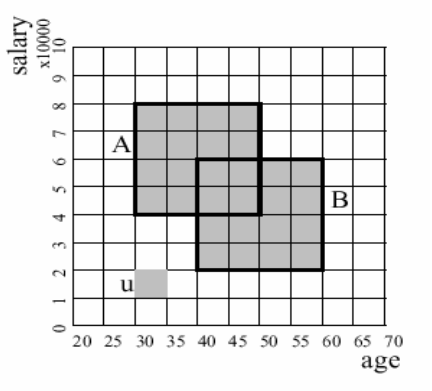


Figura 2-13. Ilustración de las definiciones del ejemplo

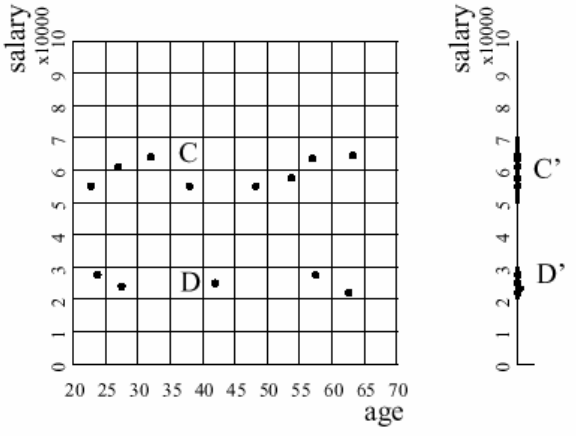


Figura 2-14. Identificación de los grupos en los subespacios del espacio de datos original

En la figura 2-14 se muestra un ejemplo de un espacio bidimensional (edad, salario) los cuales ha sido dividido mediante una cuadrícula de 10 x 10. Una unidad es la

intersección de intervalos; por ejemplo, la unidad $u = (30 \le edad < 35) \Lambda (1 \le salario < 2)$. Una región es la unión rectangular de unidades. A y B son ambas regiones, $A = (30 \le edad < 50) \Lambda (4 \le salario < 8)$ y $B = (40 \le edad < 60) \Lambda (2 \le salario < 6)$. Asumiendo que la densidad de estas unidades ha sido sombreada, $A \cup B$ es un grupo. Note que A es la máxima región contenida en el grupo, mientras que $A \cap B$ es la mínima región en el grupo. La mínima descripción de este grupo es la siguiente expresión DNF: [AGRA96]

$$((30 \le edad < 50) \Lambda (4 \le salario < 8)) \vee ((40 \le edad < 60) \Lambda (2 \le salario < 6))$$

En la figura 2-13, asumiendo τ = 20%, hay dos grupos en el espacio de la dimensión salario: C´= 5 $\le$ salario < 7 y D´ = 2 $\le$ salario < 3. No hay grupos en el espacio de edad, porque no hay densidad de unidades en ese espacio.

Clasificación de datos

Es otra técnica de minería de datos que permite establecer propiedades comunes en conjuntos de objetos en una base de datos y clasificarlos en diferentes clases, de acuerdo a un modelo de clasificación.

Para construir el modelo de clasificación, una base de datos E de ejemplo, se toma como un conjunto de datos de prueba, donde cada tupla contiene el mismo conjunto de características que las tuplas en la base de datos W y adicionalmente cada tupla contiene una identidad o clase (etiqueta) asociada a ella. El objetivo de la clasificación es primero analizar la base de datos de prueba y desarrollar una descripción precisa o modelo para cada una de las clases usando las características disponibles en los datos. Luego las descripciones de las clases se usan para clasificar la futura prueba en la base de datos W o para desarrollar una mejor descripción (llamadas ***reglas de clasificación***) para cada clase en la base de datos.

Los datos pueden clasificarse por medio de: estadísticas, regresión lineal, análisis discriminante lineal, redes neuronales, conjuntos aproximados o

lógica difusa. La clasificación resultante se puede expresar como un árbol de decisión.

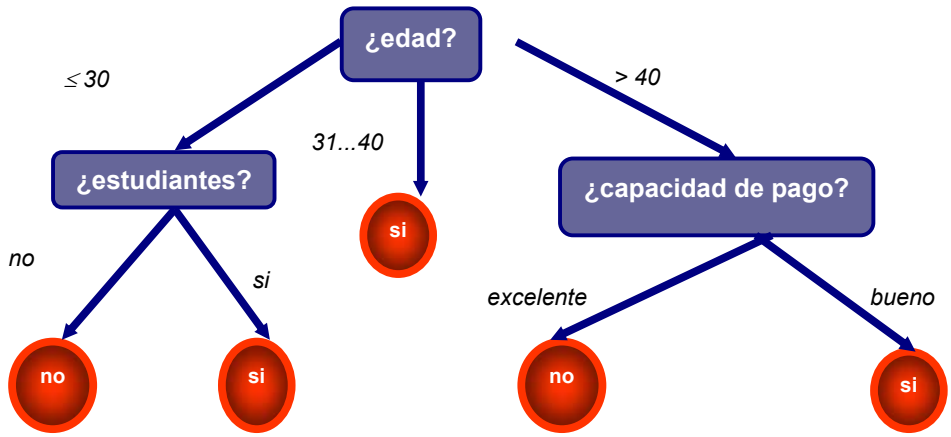


Figura 2-15. Ejemplo de una regla de clasificación, mostrada como un árbol de decisión para la compra de un computador[6]

La clasificación tiene aplicación en diagnósticos médicos, predicciones y selección de mercados entre otros.

Por ejemplo, se desean clasificar ítemes, según tres respuestas: *adecuada*, *poco representativa*, *no responde* y contar con un modelo para cada clase basado en características de los ítemes: *precio*, *marca*, *lugar de confección*, *tipo* y *categoría*. La clasificación resultante podría distinguir ampliamente las clases entre sí, presentando un cuadro organizado del conjunto de datos; también se puede expresar como un árbol de decisión. Este podría identificar precio como un factor relevante de las tres clases. El árbol puede revelar que, después del *precio*, otras características que podrían ayudar a hacer más distinciones son *marca* y *lugar de confección*. Dicho árbol puede ayudar a comprender el impacto de una determinada campaña de ventas y diseñar, así, una campaña futura más efectiva.

Cubos de datos

[6] Gráfico preparado por el Dr. Carlos Gónzalez para el curso Dtawarehouse, Data Mining y OLAP. ITCR: 2001

Los cubos de datos, se relacionan con términos como: *bases de datos multidimensionales*, *materialización de vistas* y *OLAP* (procesamiento analítico en línea). La idea general de esta aproximación es materializar ciertos cómputos caros en los que frecuentemente se incurre, sobre todo aquellos que implican funciones de agregación, como: conteos, sumas, promedio, máximo, etc., y almacenar estas vistas en una base de datos multidimensional para soporte en la toma de decisiones, descubrimiento de conocimiento y otras aplicaciones. Las funciones de agregación se pueden pre-computar de acuerdo a grupos y subgrupos de atributos. Los valores en cada atributo se pueden agrupar por su jerarquía o su estructura; por ejemplo fecha se puede agrupar por días, meses, semanas, años.

La generalización y especificación se puede desarrollar sobre una base de datos multidimensional por medio de *"roll-up"* o *"drill-down"*. Las operaciones de roll-up reducen el número de dimensiones en un cubo de datos o generaliza los valores de los atributos desde el nivel más bajo hacia el nivel más alto, las operaciones de drill-down trabajan a la inversa.

En la siguiente figura se muestra un esquema con la información de ventas de una compañía en un cubo de información, con 8 dimensiones.

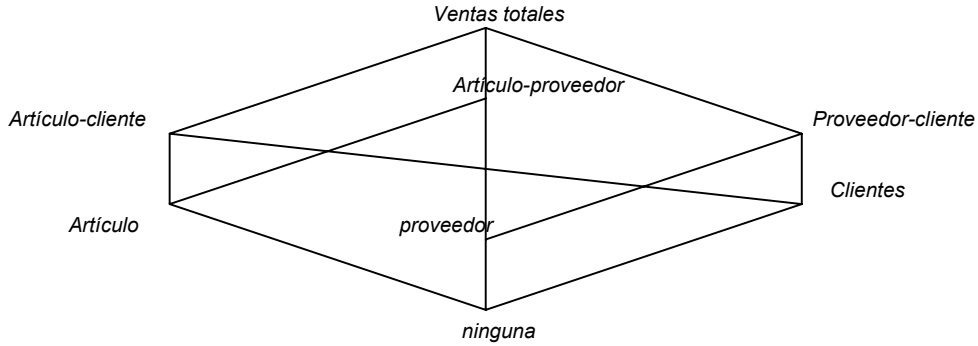


Figura 2-16. 8 vistas de un cubo de datos para información de ventas[7]

Existen tres formas de implementar los cubos:

- **Materialización física de todos los datos en el cubo**
- **No materializar nada**
- **Materializar solamente una parte del cubo**

Inducción orientada a atributos

Esta aproximación toma consultas de minería de datos expresadas en un lenguaje similar al *SQL* y guarda los grupos de datos relevantes en la base de datos. La generalización se hace entonces sobre un conjunto de datos relevantes utilizando diferentes técnicas, como árboles de decisión, extracción de atributos, funciones de agregación, entre otras.

La generalización de datos se expresa en una forma de relaciones generales, donde cada uno de los operadores o transformaciones puede transformar los datos generalizados en diferentes tipos de conocimiento o presentarlos en formas diferentes. Por ejemplo, las operaciones de *drill – down* o *roll – up* pueden presentar vistas de los datos con múltiples niveles de abstracción, la relación generalizada puede ser presentada por medio de tablas de sumarización, gráficas, curvas, etc. De aquí se puede extraer reglas características las cuales sumarizan, las características de los datos generalizados.

Una aplicación de este tipo de minería, se presenta en la Figura 2-14., la cual presenta las características generales de la base de datos de donaciones para investigación del Natural Science and Engineering

[7] Tomado de Data Mining: An Overview from a Database Perspective [CHEN96]

Research Council of Canada (NSERC). Aquí se muestran los resultados de ejecutar consultas de minería de datos sobre las donaciones recibidas por dos provincias vecinas al oeste de Canadá, en la disciplina de ciencias de la computación; las provincias comparadas son: British Columbia y Alberta. La tabla de la figura 8, se compone de varias columnas: clase, la cual indica la provincia; disciplina, que indica la disciplina de investigación; la categoría de la donación; cantidad por categoría; porcentaje de soporte, el cual representa el número de las donaciones de investigación en la categoría analizada vrs el total de donaciones en la respectiva provincia; y el porcentaje de comparación, el cual representa el número de donaciones de una categoría específica en una provincia vrs la otra. Por ejemplo, la primera fila de la tabla indica que para donaciones de infraestructura en la disciplina de computación, en el rango de 40Ks a 60Ks, British Columbia tomó el 2% del total de donaciones en computación. Por su parte, Alberta, tomó el 1.72%; sin embargo si se realiza una comparación entre ambas provincias, se observa que British Columbia tomó el 66.67%, mientras que Alberta tomó solamente el 33.33%. [CHEN96].

Class	Discipline	Grant_Category	Amount_Categoy	Support %	Comparision %
BC	Computer	Infrastructure Grant	40Ks – 60Ks	2.00	66.67
Alberta				1.72	33.33
BC	Computer	Other	20Ks – 40Ks	2.00	66.67
Alberta				1.72	33.33
BC	Computer	Other	60Ks	2.00	50.00
Alberta				3.45	50.00
BC	Computer	Research Grant: Individual	0Ks – 20Ks	38.00	63.33
Alberta				37.93	36.67
BC	Computer	Research Grant: Individual	20Ks – 40Ks	28.00	56.00
Alberta				37.93	44.00
BC	Computer	Research Grant: Individual	40Ks – 60Ks	6.00	75.00
Alberta				3.45	25.00
BC	Computer	Research Grant: Individual	60Ks	3.00	100.00
Alberta				0.00	0.00
BC	Computer	Scholarship	0Ks – 20Ks	19.00	76.00
Alberta				10.34	24.00

Figura 2-17. Ejemplo: Comparación de donaciones para investigación[8]

[8] Tomado de Data Mining: An Overview from a Database Perspective [CHEN96]

2.3 Indicadores para la toma de decisiones

Las exigencias actuales hacen que las instituciones requieran de herramientas que las diferencien del resto y les permitan mejorar su función. Una de esas herramientas son las denominadas herramientas de gestión.

Pero, los recursos con que cuentan las organizaciones para construir sus herramientas de gestión se limita a los informes elaborados por el departamento de sistemas, basados en ciertos requerimientos que los usuarios definen a través de su percepción y experiencia en el negocio y que luego son interpretados según su visión y análisis de los conceptos referenciados.

Por otra parte, muchas veces no se tiene conocimiento sobre herramientas dinámicas ni de la construcción de adecuados indicadores de gestión que les permita a los altos funcionarios (juntas directivas, etc.) indagar en las áreas del negocio, realizando exploración, análisis y localización de tendencias de la información desde las perspectivas manejadas dentro y hacia fuera de la organización.

Se puede decir, entonces, que son necesarias herramientas que ayuden a medir el cómo se están haciendo las cosas. Surge por lo tanto el uso de Indicadores de Desarrollo o Indicadores de Gestión, los cuales se pueden definir como un grupo de herramientas que ayudan a responder la pregunta ¿Conozco qué estoy haciendo?. Una definición más formal puede ser: Un indicador de desarrollo define de manera numérica el grado en que los objetivos planteados se han realizado, puede ser expresado como un porcentaje, un índice, una tasa o cualquier otra forma de comparación que permita regular intervalos o comparar más de un criterio. [BULL98]

Los indicadores se usan en diferentes instituciones y con fines distintos, por ejemplo, si uno de los objetivos de una corporación es entrenar personas para obtener empleo, un indicador podría ser qué porcentaje de las personas entrenadas han obtenido empleo.

Para desarrollar los indicadores es muy importante identificar claramente los valores y la filosofía de la organización, los usuarios y sus necesidades, los objetivos y el proceso

que se debe realizar [BULL98]. Para definir los indicadores existen una serie de premisas, las cuales se detallan a continuación [VASS02]:

- Definición operacional. Se debe dejar explícito el algoritmo o procedimiento requerido para llegar a la expresión matemática con la cual se representará el indicador.
- Establecer un acuerdo o pacto con el usuario en cuanto a las características del producto o servicio que se va a presentar y las expectativas del usuario.
- Características:
    - Los indicadores tienen como objetivo la evaluación de un producto, servicio o gestión con base en los objetivos planteados.
    - Define con claridad el comportamiento de ese producto, servicio o gestión.
    - Son un elemento fundamental en el proceso de toma de decisiones.
    - Sirve como parámetro para mejorar las expectativas del cliente.

Además existen ciertos elementos determinantes que configuran un indicador de gestión, los cuales deben ser expresamente incorporados a la función de la organización y mantenerse vigentes [VASS02] :

- Denominación: debe contemplar únicamente la característica, el evento o el hecho que se quiere controlar y se expresa en cantidad, tasa, porción o porcentaje.
- Patrón de comparación: previamente se establecen los criterios de análisis y medidas junto con los patrones contra los cuales se compara la medición.
- Interpretación: consiste en precisar como se leerá el resultado de lo que ha sido medido o expresado cuantitativamente. Así mismo, establecer de que manera podría ser graficado para su seguimiento.
- Periodicidad: se debe establecer cuántas evaluaciones se harán dentro del período de presentación del servicio y en que momento.
- Datos requeridos: para poder efectuar el cálculo, es necesario definir la fuente de la información, quién genera y procesa los datos.

2.4 Minería de datos con SQL Server 2000

La Minería de Datos o el descubrimiento de conocimiento permite a las compañías encontrar rasgos característicos en sus almacenes de datos que les permiten diseñar estrategias para obtener mayores beneficios.

SQL Server 7.0 de Microsoft en 1998, empezó a posesionarse del desarrollo de sistemas de apoyo en la toma de decisiones y de Business Intelligence (Inteligencia de negocios). SQL Server 7.0 cuenta con herramientas OLAP bastante accesibles, flexibles y funcionales, para la manipulación y consulta de cubos multidimensionales. Pero en SQL Server 2000 se han mejorado estas herramientas introduciendo algoritmos de data mining, llamados “Analysis Services”[SQLS00].

2.4.1 Evolución hacia el Data Mining

La minería de datos ha evolucionado de la mano con los “datawarehouse”; las personas han comprendido que los datos almacenados pueden mostrarles el camino por el cual debe seguir su empresa. Este planteamiento dio lugar al desarrollo de una nueva generación de sistemas informáticos como los EIS (Executive Information System), los cuales abordaban tareas de información y de análisis.

A medida que fue desarrollándose la tecnología de los datawarehousing, los sistemas EIS y de apoyo en la toma de decisiones, se dio paso a los conceptos más generales de Business Intelligence y Data Mining.

La Inteligencia de Negocios implica la organización de los datos en distintas dimensiones potenciales de análisis, de modo que se pueden mostrar y establecer referencias cruzadas entre cualesquiera de las vistas de datos (ejemplo: resultados de ventas) extraídos de cualesquiera otras dimensiones potenciales (ejemplos: región o línea de producto). La posibilidad de pasar de una dimensión a otra permite descender al detalle o distanciarse para ofrecer una visión más general. La posibilidad de mostrar variaciones en los datos de distintas dimensiones permite generar informes multidimensionales en tiempo real. Este modo general de entender la manipulación de

datos recibió el nombre de OLAP (OnLine Analytical Processing o Procesamiento Analítico en Línea), es decir, procesamiento de datos con fines analíticos en lugar de operativos. El término “on line” hace referencia al hecho de que los datos analíticos siempre se encuentran disponibles. OLAP garantiza que los datos contenidos en los almacenes de datos se encuentren siempre disponibles en un formato que permita su uso en tareas analíticas de apoyo en la toma de decisiones, además, se caracteriza por el preprocesamiento indexado y almacenamiento de datos en distintas representaciones dimensionales que le permiten generar con rapidez las distintas vistas dimensionales que requiere la Inteligencia de Negocios.

Las herramientas OLAP que se utilizan en la Inteligencia de Negocios no siempre detectan todas las pautas y dependencias que existen en los datos. Los cubos OLAP están indicados para la exploración de cantidades limitadas de datos y conllevan importantes variaciones en función de dimensiones empresariales críticas y conocidas. Pero cuando cambian las dimensiones como resultado de cambios en la empresa o de que se exploren situaciones nuevas, la tecnología del data mining es un complemento potente y flexible de OLAP, ya que las soluciones son ideales para la criba de cientos de dimensiones de análisis y de combinaciones asociadas que compiten entre sí y que son potencialmente útiles.

2.4.2 Desarrollo de Data Mining con SQL Server 2000

Los algoritmos de data mining incluidos en SQL Server 2000, tienen como objetivo la búsqueda de la “estructura que hay en los datos”, donde las estructuras se hacen patentes a través de las pautas, que son relaciones o correlaciones entre datos.

Las consultas de data mining son distintas a las consultas tradicionales del mismo modo que los modelos de data mining son distintos de las tablas de bases de datos tradicionales. En las consultas se especifica la pregunta a la que se desea dar respuesta y el procesador de consultas de data mining envía a la estación de consulta los resultados mediante un modelo estructural que responde a la pregunta realizada.

A continuación se muestra un ejemplo de cómo se crea un modelo de data mining que permite predecir o clasificar la edad en función de otros atributos del conjunto de datos como el sexo, el nombre del producto, el tipo de producto o la cantidad. El cliente ejecuta una sentencia CREATE, similar a la CREATE TABLE. La especificación OLE DB para data mining incluye una descripción completa del lenguaje que se utiliza para la creación y manipulación de modelos de data mining (www.microsoft.com/data/oledb).

Consulta que crea un modelo de data mining

```
CREATE MINING MODEL [Predicción de edad]
(
[ID cliente] LONG KEY,
[Sexo] TEXT DISCRETE,
[Edad] DOUBLE DISCRETIZED() PREDICT,
[Compras del producto] TABLE
(
[Nombre del producto] TEXT KEY,
[Cantidad] DOUBLE NORMAL CONTINUOUS,
[Tipo de producto] TEXT DISCRETE RELATED TO [Nombre del producto]
)
)
USING [Arboles de decision]
```

Figura 2-18. Ejemplo: Consulta que crea un modelo de data mining[9]

Cuando se crea una estructura de modelo de data mining, se almacena como parte de una jerarquía de objetos en el directorio << Servicios de análisis >>. Las pautas o estructura de los datos, se almacenan de forma resumida con las dimensiones, pautas y relaciones con el fin de preservar el poder de predicción o de clasificación de las datos con independencia de lo que les ocurra a los datos originales del nivel de filas en que se base el modelo.

Los servicios de análisis de SQL Server 2000, han seguido tres estrategias: la del autoservicio, la de la integración de OLAP y data mining, y la de UDA (Universal Data Access o Acceso Universal a Datos)

El estándar OLE DB para data mining incluye la idea de un mecanismo universal de acceso a datos que permite compartir datos y resultados de data mining entre entornos heterogéneos con múltiples aplicaciones. Se caracteriza por el acceso heterogéneo a

[9] Tomado de [SQLS00]

datos, por medio de almacenamiento de consultas multidimensionales y de explotación compartido y por una interfaz común para consultas de OLAP y consultas de data mining.

Los modelos de data mining se pueden desarrollar a partir de fuentes relacionales (tablas estándar) o dimensionales (estructuras cúbicas). El << Analysis Manager >> (Administrador de análisis) facilita la interacción con los modelos de data mining.

2.4.3 OLE DB para data mining

OLE DB para data mining utiliza los algoritmos de data mining más conocidos, con este las aplicaciones de data mining pueden acceder a cualquier fuente de datos tabular a través de un proveedor de OLE DB y pueden efectuarse análisis de data mining directamente en bases de datos relacionales. OLE DB para minería de datos utiliza los siguientes conceptos y características:

Modelo de data mining

El modelo de data mining es igual a una tabla relacional salvo por el hecho de que contiene columnas especiales que permiten detectar las pautas y relaciones típicas que pone de manifiesto esta tecnología, estas columnas también permiten realizar predicciones. El modelo de data mining es un instrumento esencial que no sólo crea el modelo de predicciones, sino que también hace las predicciones. A diferencia de las tablas relacionales estándar, que contienen datos sin procesar, los modelos de data mining almacenan las pautas halladas mediante el algoritmo de data mining.

OLE DB para data mining es una ampliación de OLE DB que permite tratar los modelos de data mining como un tipo esencial de tabla. Cuando se introducen los datos en la tabla, un algoritmo de data mining los procesa y el procesador de consulta guarda el modelo de data mining resultante en lugar de los datos propiamente dichos. Tan pronto como se ha guardado, el modelo de data mining se puede examinar, refinar o usar para hacer predicciones.

2.4.4 El proceso de Data Mining

Los datos que se utilizan mediante la tecnología de minería de datos son un conjunto de tablas. Al conjunto de datos que compone una sola entidad (por ejemplo: un cliente) se le denomina <<caso>> y al conjunto de casos asociados se le denomina <<conjunto de casos>>. OLE DB para data mining utiliza tablas anidadas, es decir, <<tablas almacenadas a su vez en otras tablas>>, definidas mediante Data Shaping Service (Servicio que permite a las aplicaciones crear relaciones entre claves, campos y conjunto de filas y que es compatible con el lenguaje Shape).

Una característica que presenta la implementación de data mining en SQL Server 2000 es la facilidad de despliegue, además uno de los objetivos principales que persigue es integrar la funcionalidad de data mining directamente en la base de datos, de modo que un modelo de data mining sea un objeto de base de datos en la misma medida que lo es una tabla de datos. [SQLS00]

Servicios de análisis

La tecnología de data mining se puede aplicar a una gran cantidad de tareas diferentes que se pueden clasificar en tres tipos básicos: los modelos de resultado, los modelos de cluster y los modelos de afinidad. Los modelos de resultado permiten predecir o clasificar un resultado en función de uno o más campos (o variables) del conjunto de datos. Los modelos de cluster, a veces denominados «segmentación», permiten agrupar casos similares que comparten los valores de muchos campos de un conjunto de datos. Los modelos de afinidad (incluido el análisis de asociación, secuencia y desviación) y el modelado de dependencias permiten ver la relación o secuencia que existe entre un campo y otro. Los Servicios de análisis de SQL Server 2000 incluyen dos algoritmos básicos de data mining para la realización de tareas de clasificación y clustering: los árboles de decisión y los análisis de clusters.

Modelos de resultados con árboles de decisión

El modelado de resultados utiliza un conjunto de variables de entrada para la predicción o clasificación del valor de una variable de objetivo, o de respuesta, que es el resultado. La variable de objetivo puede ser categórica (con valores discretos como «Respondió/No respondió») o continua (con valores que expresan, por ejemplo, la cantidad en colones invertida en una compra). Cuando la variable de objetivo es categórica, recibe el nombre de «tarea de clasificación»: un modelo que muestra cuáles son las combinaciones de las variables de entrada que se pueden utilizar para clasificar el resultado de manera fiable. Cuando la variable de objetivo es continua, recibe el nombre de «modelo de regresión». La regresión es el tipo más común de análisis que permite predecir los valores de una variable objetivo continua a partir de los valores combinados de las variables de entrada. Por razones de simplicidad, Microsoft utiliza el término «clasificación» para referirse tanto a la «clasificación» propiamente dicha como a los «árboles de regresión». Si esto diera lugar a confusión, sería bueno recordar que los árboles de decisión permiten predecir o clasificar tanto resultados discretos como continuos.

Los árboles de decisión son una técnica que se utiliza con frecuencia para la realización de tareas de modelado de predicciones en las que existe un campo de resultado para el que se desean buscar pautas. Los árboles de decisión son muy fáciles de utilizar, generan gráficos de fácil lectura y funcionan bien tanto con datos categóricos como continuos. La Figura 2-19 muestra cómo se pueden organizar los datos para medir el nivel de respuesta que obtiene una invitación a un congreso sobre tecnologías de la información. Aun cuando el conjunto de datos que hemos utilizado es reducido, resultaría bastante difícil determinar a simple vista los atributos (columnas) que permiten predecir la probabilidad de que los sujetos acepten (es decir, respondan a) la invitación, si es que existe alguno. Imaginemos, por ejemplo, que intentamos determinar cuál es el factor que influye en la probabilidad de respuesta en una base de datos que contiene más de 10.000 registros. ¿Será el puesto de trabajo, el sexo, el número de empleados o el volumen de ventas? Si ya resulta difícil ver las relaciones predictivas que existen entre dos variables, resultará imposible determinar las combinaciones de relaciones predictivas que generan una sólida clasificación predictiva de la probabilidad de respuesta. [SQLS00]

Ejemplo que muestra la respuesta a una invitación a un congreso.					
Nº cliente	**Puesto**	**Sexo**	**TamañoEmpresa**	**VolumenVentas**	**Respuesta**
1	Científico jefe	M	Pequeño	> 200 millones ptas.	No respondió
2	Desarrollador	H	Pequeño	> 200 millones ptas.	No respondió
3	Programador de TI	H	Pequeño	> 200 millones ptas.	No respondió
4	Programador de TI	H	Pequeño	> 200 millones ptas.	No respondió
5	Programador de TI	H	Pequeño	> 200 millones ptas.	No respondió
6	Jefe de producto	H	Pequeño	< 200 millones ptas.	No respondió
7	Jefe de producto	H	Pequeño	> 200 millones ptas.	No respondió
8	Jefe de producto	H	Pequeño	> 200 millones ptas.	No respondió
9	Jefe de producto	H	Pequeño	> 200 millones ptas.	No respondió
10	Jefe de producto	M	Grande	< 200 millones ptas.	No respondió
11	Jefe de producto	M	Grande	< 200 millones ptas.	No respondió
12	Jefe de producto	M	Grande	< 200 millones ptas.	No respondió
13	Jefe de producto	M	Pequeño	< 200 millones ptas.	No respondió
14	Jefe de producto	H	Grande	> 200 millones ptas.	
15	Jefe de producto	H	Pequeño	> 200 millones ptas.	No respondió

Figura 2-19. Ejemplo análisis invitación a un congreso[10]

La Figura 2-20 muestra un árbol de decisión que revela la estructura predictiva de los datos: el tamaño de la empresa, expresado mediante el número de empleados, parece ser el predictor dominante de la asistencia al congreso. La tasa global de asistencia (el número de personas que respondieron a la invitación) es del 40%. El porcentaje de invitados de grandes empresas que asistieron al congreso fue del 75%, mientras que el porcentaje de invitados de pequeñas empresas que asistieron fue sólo del 27%. En otras palabras, es aproximadamente tres veces más probable que asistan al congreso los empleados de las grandes empresas que los de las empresas pequeñas. Pero esta tendencia que se aprecia en las pequeñas empresas se invierte tan pronto como se tiene en cuenta el volumen de ventas, ya que en el caso de las dos pequeñas empresas cuyos volúmenes de ventas eran inferiores a los 200 millones de pesetas, el nivel de asistencia fue del cien por ciento. Esta cifra revela que, en nuestro ejemplo, todos los empleados cuyas empresas tenían unos ingresos anuales inferiores a los 200 millones de pesetas asistieron al congreso. Le recordamos que sólo se trata de un ejemplo y que

[10] Tomado de [SQLS00]

nunca basaríamos unos resultados en un número tan reducido de registros a no ser que, previamente, hubiéramos realizado suficientes pruebas con otros conjuntos de datos para verificar que la pauta se repitiera de modo fiable en la población objeto del estudio. [SQLS00]

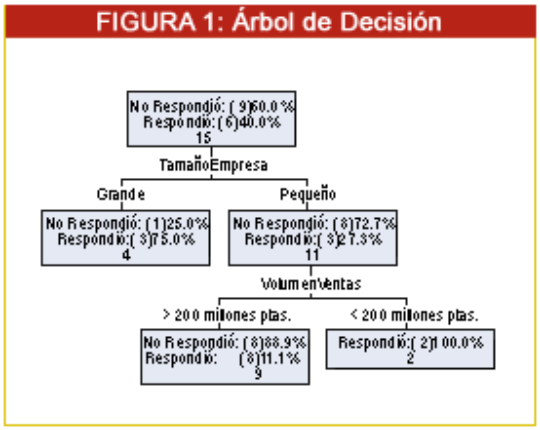


Figura 2-20. Ejemplo árbol de decisión[11]

Los árboles de decisión funcionan recopilando el conjunto total de datos (que aparece normalmente representado en la parte superior de la figura como el nodo raíz u origen del árbol de decisión) y particionando los registros o casos que existan en el nodo raíz para formar las ramas. Las ramas aparecen dispuestas de modo que forman un árbol invertido y los nodos que hay en sus extremos reciben, normalmente, el nombre de «hojas». En un estudio real de mercado, existirían muchos más atributos (columnas) por cada asistente potencial al congreso, así como muchos más asistentes potenciales. A medida que aumenta la escala del problema, se hace más difícil evaluar las características predictivas de forma manual, por lo que resulta imprescindible recurrir a técnicas automatizadas como las que ofrecen las sofisticadas herramientas de OLAP. Aun así, aunque en menor medida, sigue resultando difícil evaluar el poder predictivo de múltiples atributos,. Lo que se persigue mediante data mining y los árboles de decisión es poder evaluar el poder predictivo combinado de múltiples atributos. La Figura 2-19 da buena idea de cómo funcionan estas tareas de evaluación. Podemos ver que, en

[11] Tomado de [SQLS00]

conjunto, el 40% de los miembros del conjunto de datos respondieron a la invitación. Sin embargo, si sólo se toman en cuenta las empresas pequeñas (Tamaño- Empresa: Pequeño) que poseen grandes volúmenes de ventas (VolumenVentas: > 200 millones ptas.), la tasa de respuesta cae hasta, aproximadamente, el 11%. Este ejemplo muestra que se puede medir la capacidad para evaluar la caída en la tasa de respuesta a medida que se introducen nuevos atributos en el análisis y, a continuación, utilizar dicha medida para generar un indicador del poder predictivo del modelo. Los árboles de decisión son muy escalables, ya que admiten campos que contienen atributos con muchos valores y muchos registros de datos. Debido a esta gran escalabilidad, los árboles de decisión resultan extremadamente útiles para una amplia gama de tareas de clasificación y modelado predictivo. [SQLS00]

Segmentación (análisis de clusters)

La segmentación es el proceso de agrupación o clustering de casos que comparten un mismo conjunto de atributos. Los árboles de decisión también permiten encontrar segmentos, pero en función de una variable de resultado determinada, como por ejemplo, la asistencia a un congreso. Por ello, los valores, expresados mediante cadenas o códigos numéricos, de una rama del árbol de decisión forman un cluster cuyos casos (una hoja del árbol de decisión) comparten el atributo de la rama en cuestión. En la Figura 2-20, por ejemplo, podemos ver que las empresas pequeñas que tienen altos volúmenes de ventas forman un segmento cuya tasa de respuesta es inferior a la que presentan los restantes segmentos del árbol de decisión. El árbol de decisión forma una rama que muestra la similitud que existe entre los casos de la hoja (cluster), y los resultados de las dos opciones del nodo que, en este ejemplo, son «Respondió» y «No respondió». Si no existe ninguna variable de resultado o si lo que se desea es ver cómo se agrupan las observaciones en función de los valores compartidos de múltiples variables de resultado, el análisis de cluster es la técnica a elegir.

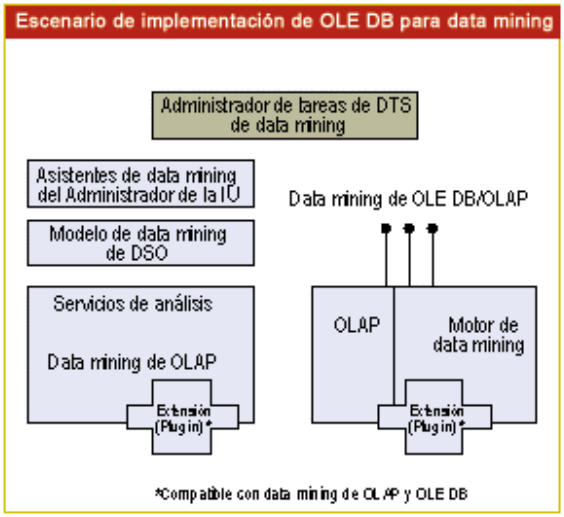


Figura 2-21. Escenario de implantación de OLE DB para data mining

Los análisis de clusters forman grupos de casos lo más homogéneos posible en lo que respecta a varios atributos compartidos (tales como la altura, el peso y la edad) y lo más diferentes posible de otros clusters a su vez homogéneos. Por ejemplo, un análisis de clusters podría identificar todos los casos de gran estatura, de mucho peso y de menor edad de un cluster y todos los casos de escasa estatura, de poco peso y de mayor edad de otro cluster. Los casos que muestran pautas de adquisición o gasto similares forman segmentos de mercado fácilmente identificables a los que se pueden dirigir distintos productos. Por lo que se refiere a la interacción personalizada, distintos clusters pueden mostrar claros indicios que aconsejen tratamientos diferentes.

Con el paso del tiempo, se han ido desarrollando varias técnicas para la realización de tareas de análisis de clusters; una de las más antiguas es el análisis de clusters «K-means». En dicho análisis de clusters, el usuario asigna varias medias que servirán de receptáculos o clusters donde incluir las observaciones del conjunto de datos. Los casos se asignan, entonces, a cada uno de los receptáculos en función de las similitudes que existan entre ellos. «Servicios de análisis» utiliza el método de análisis de clusters denominado del «vecino más cercano» (nearest-neighbor), con K-means asignadas al azar.

CAPITULO 3. Metodología

Para realizar este trabajo de investigación se siguieron los siguientes pasos:

3.1. Entrevistas con personal del sector construcción del país

Esto con el fin de establecer necesidades de información y orientar de esta manera la búsqueda de información.

- Profesores universitarios, de la Escuela de Ingeniería en Construcción del ITCR.
- Desarrolladores de proyectos.
- Funcionarios del Ministerio de Vivienda, del INVU y del BANVHI.
- Otros usuarios.

En las visitas y entrevistas se obtuvo la siguiente información:

Instituto Nacional de Estadística y Censo:

La visita fue atendida por la señora Marita Begueri, coordinadora de la Encuesta de Hogares y el Censo Nacional.

La señora Begueri se mostró muy interesada en el proyecto y dispuesta a suministrar información sin costo, siempre y cuando fuera utilizada por universidad pública.

En cuanto a la posibilidad de establecer un convenio con el ITCR, para que en un futuro se pueda realizar un proyecto en esta área, para ello, primero se debe evaluar que tipo de valor agregado obtendrían de la información final que podría generar.

Aclaró además, que la información que se encuentra en la oficina del INEC posee un mayor nivel de segregación que la disponible en Internet.

Ministerio de Salud

Los representantes del Ministerio de Salud, el Ing. Edgar Morales, Director de la Dirección de Informática y la Dra. Azalia Espinoza de la Dirección de Servicios de Salud, se mostraron interesados en el proyecto, aunque un poco “celosos” de la información que manejan y estarían a la espera de la creación de alguna figura relacionada con alguna universidad pública para el manejo de la información.

Cámara Costarricense de la Construcción

De las instituciones visitadas, fue la que expresó mayor interés en el interés en el proyecto y el Lic. Randall Murillo, Gerente de la Cámara, se mostró muy dispuesto a colaborar.

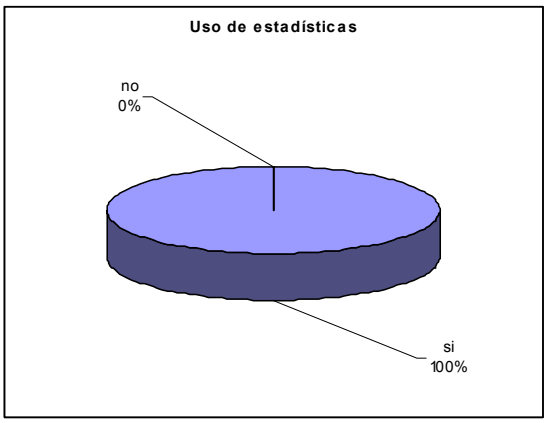
Producto de este trabajo, se elaboró un convenio marco entre la cámara y el ITCR, el cual ya fue firmado. Este es un primer paso para la obtención de datos y luego la generación de información.

Ministerio de la Vivienda y Asentamientos Humanos

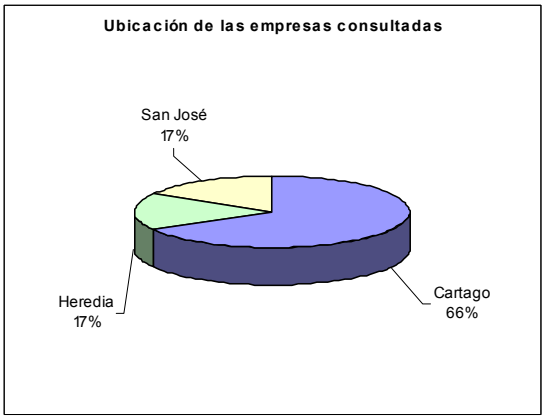
El Arq. Díaz del MIVAH se mostró muy interesado en el proyecto, ya que de manera informal el actual Ministro de Vivienda quiere implementar una serie de indicadores para la toma de decisiones, para lo cual ha venido coordinando con el INEC en la elaboración de alguna información específica. Por lo tanto el proyecto tendría una buena aceptación en el Ministerio y estarían dispuestos a establecer un convenio con el ITCR para el desarrollo de estos indicadores.

Para determinar el tipo de información, la forma en que ésta debería mostrarse, el interés y la necesidad de información, se utilizó una encuesta, que se aplicó a diversas instituciones y arrojó los siguientes resultados:

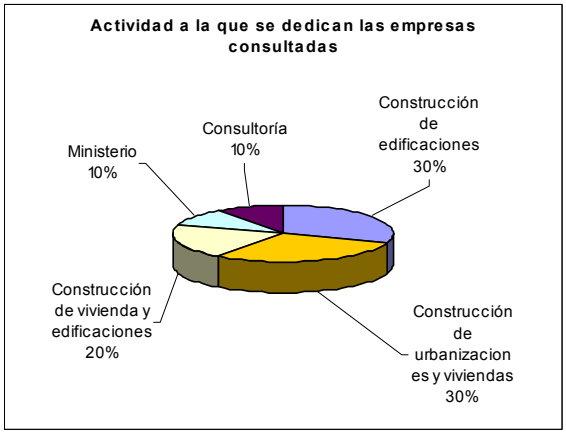
Pregunta 1. *Actividad a la que se dedican las empresas o instituciones consultadas:*



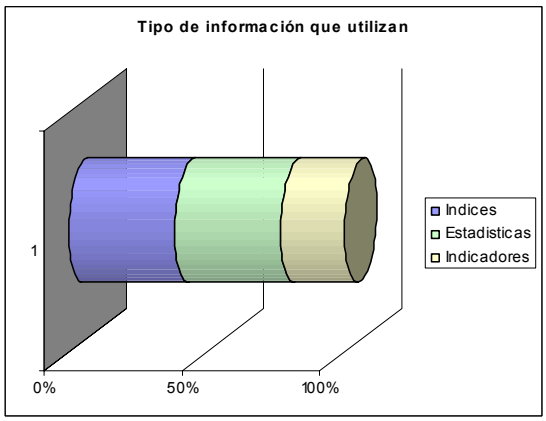
Pregunta 2. *Ubicación de las empresas consultadas:*



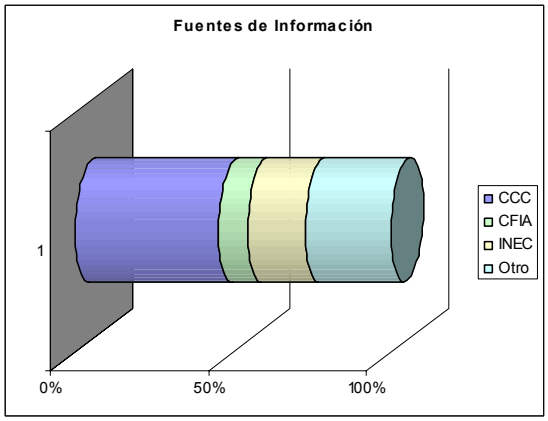
Pregunta 3. *Uso de estadísticas por parte de las empresas para la toma de decisiones:*



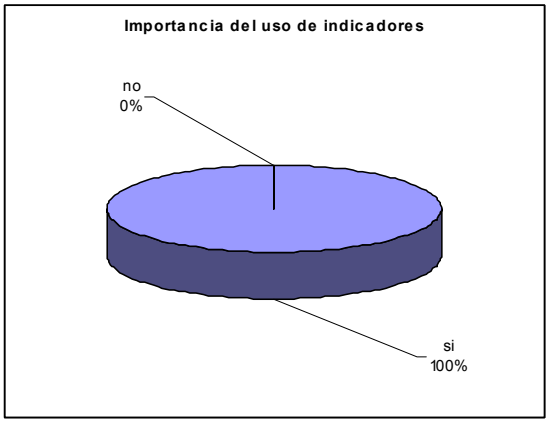
Pregunta 4. *Tipo de información que utilizan las empresas:*



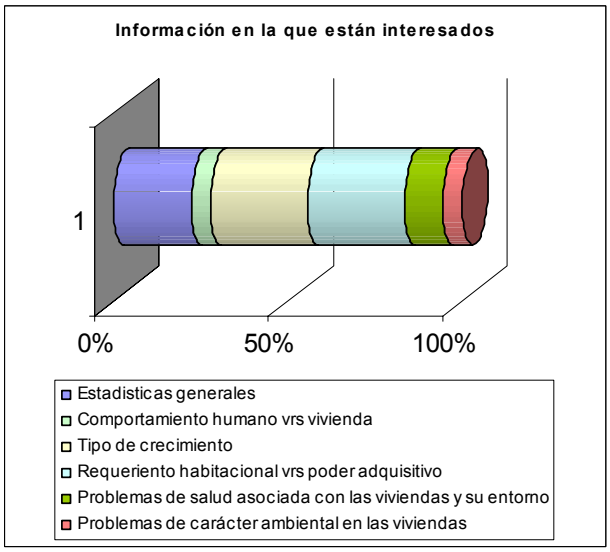
Pregunta 5. *Fuentes de información de las empresas:*



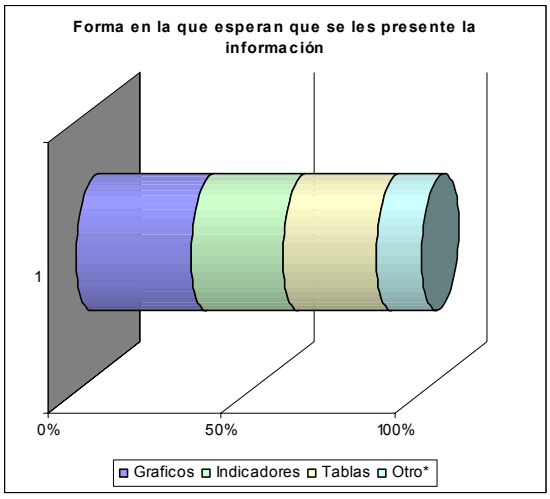
Pregunta 6. *Importancia del uso de Indicadores:*



Pregunta 7. *Información en la que están interesadas las empresas:*



Pregunta 8. *Forma en que se desea que se presente la información:*



Con base en estos datos, se estableció la información por analizar y cruzar, la cual se detalla en el capítulo siguiente, bajo el título necesidades del sector construcción.

3.2. Análisis de las fuentes de datos

Con esta visita se pretendió analizar alguna de la información disponible, orientada a las prioridades establecidas anteriormente.

Se visitaron diferentes instituciones y se analizó la información disponible en cada una de ellas, se elaboraron y realizaron encuestas. Específicamente se visitaron el INEC, el Ministerio de Salud, la Cámara Costarricense de la Construcción y FUPROVI.

El detalle de la información disponible, se muestra en el capítulo 2, bajo el título fuentes de datos.

Es muy importante mencionar que se encontró un problema, los datos son almacenados en motores de bases de datos diferentes y con distintos niveles de agregación, lo cual implica que la fase de limpieza de datos, requerirá un esfuerzo adicional. De allí que para efectos de este trabajo, se hará una carga demostrativa de algunos datos, de forma manual.

3.3. Planteamiento del tipo de información que se desea generar

Con base en las fuentes de datos y las entrevistas realizadas, se establecieron una serie de áreas con la información que podría analizarse y la cual sería la base para la creación de los indicadores. Ver capítulo 4.

3.5. Creación de un prototipo de depósito de datos

Esta etapa es preliminar a la construcción de un cubo de información y se utilizó en este caso el programa Access 2000, para crear un prototipo de depósito de datos sobre el cual extraer la información. En este caso el prototipo se centró, como se mencionó anteriormente, en la provincia de Cartago y se alimentó manualmente con la información disponible. No se creó ningún programa para alimentarlo. Ver capítulo 4.

3.6. Creación de un cubo de datos

Con base en el prototipo de depósito de datos, se creó un cubo de datos (OLAP utilizando SQL Server 2000, esto con el fin de establecer posteriormente los indicadores. Ver capítulo 4.

3.7 Creación de Indicadores

Una vez creado el cubo de información, utilizando el prototipo de información almacenada, se obtuvieron algunos indicadores relevantes para la toma de decisiones del sector vivienda. Es importante señalar que estos indicadores se centraron en una provincia, pero su aplicación es general y para ello se necesita alimentar el depósito de datos adecuadamente. Ver capítulo 4.

CAPITULO 4. Indicadores para el sector vivienda

4.1 Necesidades de información del sector construcción

Para establecer la información que requiere el sector construcción, se consultaron varios actores del mismo, de tal manera que indicaran el uso de información en la toma de decisiones, que tipo de información se utiliza, fuentes de información y que tipo de información necesita para realizar su trabajo.

Para ello se entrevistó a personas en las siguientes instituciones:

- Cámara Costarricense de la Construcción.
- Ministerio de Vivienda y Asentamientos Humanos.
- Escuela de Ingeniería en Construcción y Centro de Investigaciones en Vivienda y Construcción.
- Municipalidades.
- Empresas constructoras.

De estas entrevistas se pudo comprobar que la mayoría de información que se utiliza proviene del INEC, pero cada institución la procesa de diferente manera de acuerdo a sus necesidades.

En general alguna de la información que se desea cruzar se puede resumir de la siguiente manera.

- Estadísticas generales sobre vivienda
    a. Cantidad de viviendas construidas anualmente
    b. % de viviendas de interés social (con bono)

c. % de viviendas de clase media
d. % de viviendas de clase alta
e. Cantidad de viviendas en área rural
f. Cantidad de viviendas en área urbana
g. Tipo de materiales usados
h. Acceso a electricidad
i. Acceso a agua potable
j. Cuál es el déficit de vivienda, cual nivel social está más afectado?

- Comportamiento humano vrs vivienda
    a. Relación entre violencia y tipo de vivienda, identificando sistema constructivo y hacinamiento. Los datos se podrían tomar de : Estadísticas de violencia (MSP) y Estadísticas de construcción (CCC).

- Riesgo vrs Vivienda
    a. Identificar el factor de riesgo que tienen las viviendas actualmente. Los datos se podrían tomar de Base de datos de la CNE y las Municipalidades.

- Tipo de crecimiento
    a. Establecer tendencias de crecimiento, hacia dónde está tendiendo el crecimiento y que tipo de sistemas constructivos se están utilizando. Fuentes de datos: CCC y Municipalidades.

- Requerimiento habitacional vrs poder adquisitivo
    a. Identificar el déficit cualitativo y cuantitativo y relacionarlo con la capacidad de adquirir o mejorar las viviendas que tienen los usuarios. Fuentes de datos: INEC.

- Problemas de salud asociadas con las viviendas y sus entornos
    a. Problemas de origen sanitario (calidad del agua, del aire, instalaciones sanitarias, etc.)
    b. Problemas asociados a los materiales de construcción
    c. Problemas asociados al entorno
        i. Contaminación sonora
        ii. Contaminación eléctrica
        iii. Ubicación de fábricas contaminantes cerca

d. Problemas asociados al diseño: falta de iluminación, ventilación, tamaño de la vivienda.

- ❖ Sostenibilidad de las viviendas
    a. Identificar que tan amigables con el ambiente son las viviendas, identificando aspectos como: uso de recursos, acceso a los sistemas de infraestructura, tratamiento de desechos. Fuente: INEC y Municipalidades.

- ❖ Desarrollo económico vrs construcción
    a. Proyecciones del sector construcción
    b. Requerimientos de infraestructura del sector turismo
    c. Accesibilidad al crédito

4.2 Creación del cubo de datos

Para la elaboración de los indicadores para la toma de decisiones en el sector vivienda se utilizará un cubo de datos multidimensional el cual se creará utilizando la metodología planteada en la referencia [HUSE00].

Los cubos de datos son los objetos principales de un sistema OLAP, son un conjunto de datos que se construyen usualmente a partir de un subconjunto de Datawarehouse y se organizan y abstraen en una estructura multidimensional definida por dimensiones y medidas [GONZ01].

La metodología para la elaboración incluye la identificación de los hechos y las relaciones, la estructuración del esquema E/R, la derivación del gráfico multidimensional y la conversión a un modelo multidimensional.

4.1.1 Modelo Conceptual

- ❖ Definición del contexto

Aquí es muy importante definir el objetivo que se pretende alcanzar o la necesidad que se quiere satisfacer, además de qué aspectos son necesarios para alcanzarlos. Al final de esta etapa se obtendrá la especificación de requerimientos.

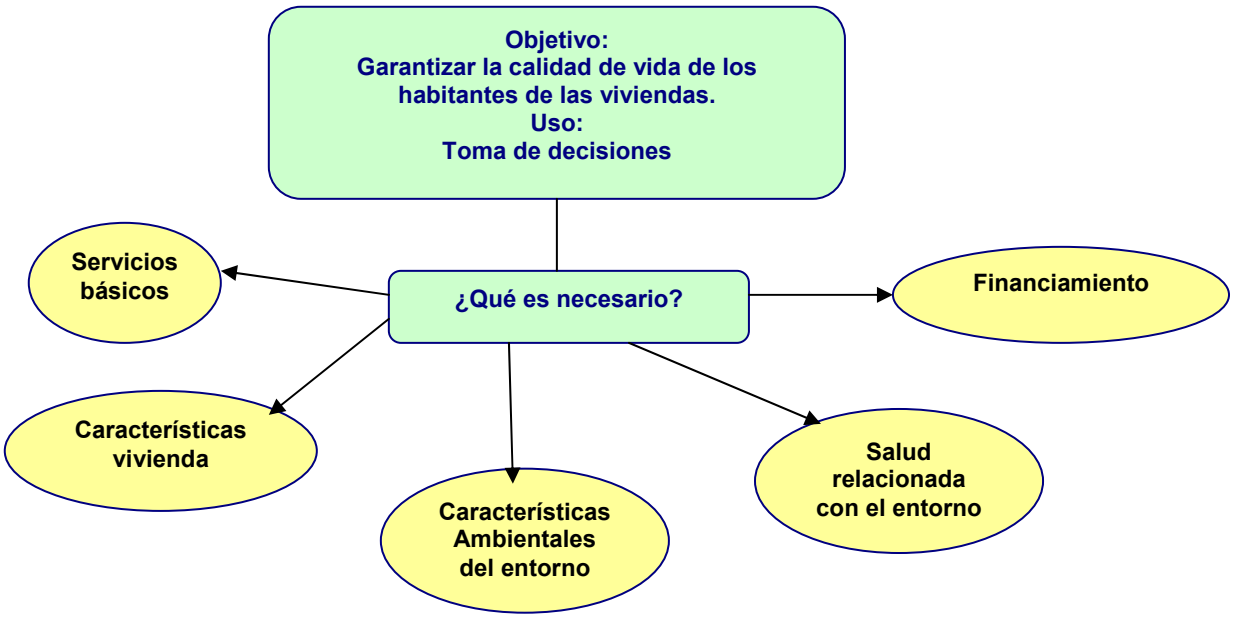


Figura 4-1. Esquema del planteamiento del sistema

La especificación de los requerimientos se hará con base en el esquema de la figura 4-1 y consiste en definir los atributos necesarios con una descripción informal.

Un atributo describe información adicional relacionada con el nivel de dimensión [HUSE00].

Atributo	Descripción	M[12]	D[13]	O[14]
serviciosbasicos	Evaluación de los servicios básicos con se cuenta	no	si	no
D_electricidad	Saber si se cuenta o no con disponibilidad del servicio eléctrico	si	no	no
D_agua	Disponibilidad de acueducto y servicio de agua potable	si	no	no
C_agua	Calidad del agua que se recibe	si	no	si
Q_agua	Compañía que brinda el servicio	si	no	si
D_aguasnegras	Disponibilidad de evacuación de aguas negras	si	no	no
T_aguasnegras	Disponibilidad de algún tratamiento para las aguas negras	si	no	no
TP_aguasnegras	Tipo de tratamiento para las aguas negras	si	no	si
D_aguaspluviales	Disponibilidad de evacuación de aguas pluviales	si	no	no
D_transporte	Disponibilidad de transporte	si	no	no
D_telefono	Disponibilidad de servicio telefónico	si	no	no
D_industrias	Cercanía con industrias	si	no	si
D_salud	Cercanía con centros de salud	si	no	si
caracteristicasvivienda	Describir las características de las viviendas y su relación con la cantidad de habitantes	no	si	no
T_vivienda	Tamaño de la vivienda	si	no	no
N_personas	Número de personas que habitan determinada vivienda	si	no	no
D_materiales	Materiales predominantes en la construcción de la vivienda	si	no	si
ambiente	Características ambientales	no	si	no
D_clima	Descripción del clima de la zona	si	no	si
D_zona	Descripción del tipo de zona, si existe o no riesgo	si	no	si
D_topografia	Descripción de la topografía del terreno	si	no	si
salud	Características relacionadas con la salud de los habitantes de las viviendas	no	si	no
P_enfermedades	Enfermedades más comunes en determinada zona	si	no	no
D_Contaminacion	Existencia de contaminación y de qué tipo	si	no	si
financiamiento	Características del posible financiamiento y su disponibilidad	no	si	no
D_financiamiento	Accesibilidad al crédito	si	no	si
D_nivelsocioeconomico	Nivel socioeconómico	si	no	no
T_financiamiento	Tipo de financiamiento	si	no	si
solución	Características de la solución propuesta	no	si	no
D_construccionnueva	Construcción nueva a evaluar	si	no	no
D_remodelacion	Remodelación a evaluar	si	no	no

[12] Atributo usado como medida.
[13] Atributo dimensional.
[14] Atributo opcional.

Figura 4-2. Definición de atributos

Luego se procede a realizar el modelo conceptual determinando las dependencias funcionales desde los niveles de dimensión hasta las medidas.

El modelo conceptual tiene como objetivo producir un esquema multidimensional gráfico y consta de tres fases: definición conceptual de medidas, diseño dimensional jerárquico y definición de restricciones.

Antes de iniciar la definición conceptual de medidas, es importante tener claro que una medida es un conjunto de valores que se basan en una columna en la tabla de hechos y usualmente es numérico. Las medidas son los valores centrales que son agregados y analizados [GONZ01].

Teniendo un conjunto de medidas de la forma $M = \{m_i, ..., m_k\}$, para la definición conceptual se debe:

- Determinar las dependencias funcionales desde los niveles de dimensión hasta las medidas, (FD_s).
  - Determinar la llave para cada medida, $D_i \subseteq D$ para cada medida m_i.
  - Definir el conjunto de llaves, $F_{key} \rightarrow todas\ las\ FD_s\ de\ la\ forma\ D_i \rightarrow m_i$.
  - Expresar las (FD_s) de la forma $D_i \rightarrow m_i \in F_{key}$.
  - Para cada nivel de dimensión terminal definir la dimensión correspondiente.

Con base en las consideraciones anteriores se tiene:

- Definición de llaves:

$$FD(D\_electricidad) \rightarrow electricidad \in F_{key}$$
$$FD(D\_agua, C\_agua, Q\_agua) \rightarrow agua \in F_{key}$$
$$FD(D\_aguasnegras, T\_aguasnegras, TP\_aguasnegras) \rightarrow aguasnegras \in F_{key}$$
$$FD(D\_aguaspluviales) \rightarrow aguaspluviales \in F_{key}$$
$$FD(D\_transporte) \rightarrow transporte \in F_{key}$$

$FD(D\_telefono) \rightarrow telefono \in F_{key}$

$FD(D\_industrias, D\_salud) \rightarrow cercania \in F_{key}$

$FD(T\_vivienda, D\_materiales, N\_personas) \rightarrow vivienda \in F_{key}$

$FD(D\_clima, D\_zona, D\_topografia) \rightarrow ambiente \in F_{key}$

$FD(P\_enfermedades, D\_contacion) \rightarrow salud \in F_{key}$

$FD(D\_financiamiento, D\_nivelsocioeconomico, T\_financiamiento) \rightarrow financiamiento \in F_{key}$

$FD(D\_construccionnueva, D\_remodelacion) \rightarrow solucion \in F_{key}$

- Dependencias funcionales entre los niveles de dimensión y las medidas

Esquema de hechos	Medidas	Dimensión
Calidad de vida	D_electricidad	Servicios básicos
	D_agua	
	C_agua	
	Q_agua	
	D_aguasnegras	
	T_aguasnegras	
	TP_aguasnegras	
	D_aguaspluviales	
	D_transporte	
	D_telefono	
	D_industrias	
	D_salud	
	T_vivienda	Vivienda
	N_personas	
	D_materiales	
	D_clima	Ambiente
	D_zona	
	D_topografia	
	P_enfermedades	Salud
	D_Contaminacion	
	D_financiamiento	Financiamiento
	D_nivelsocioeconomico	
	T_financiamiento	
	D_construccionnueva	Solución
	D_remodelacion	

Figura 4-3. Dependencias funcionales

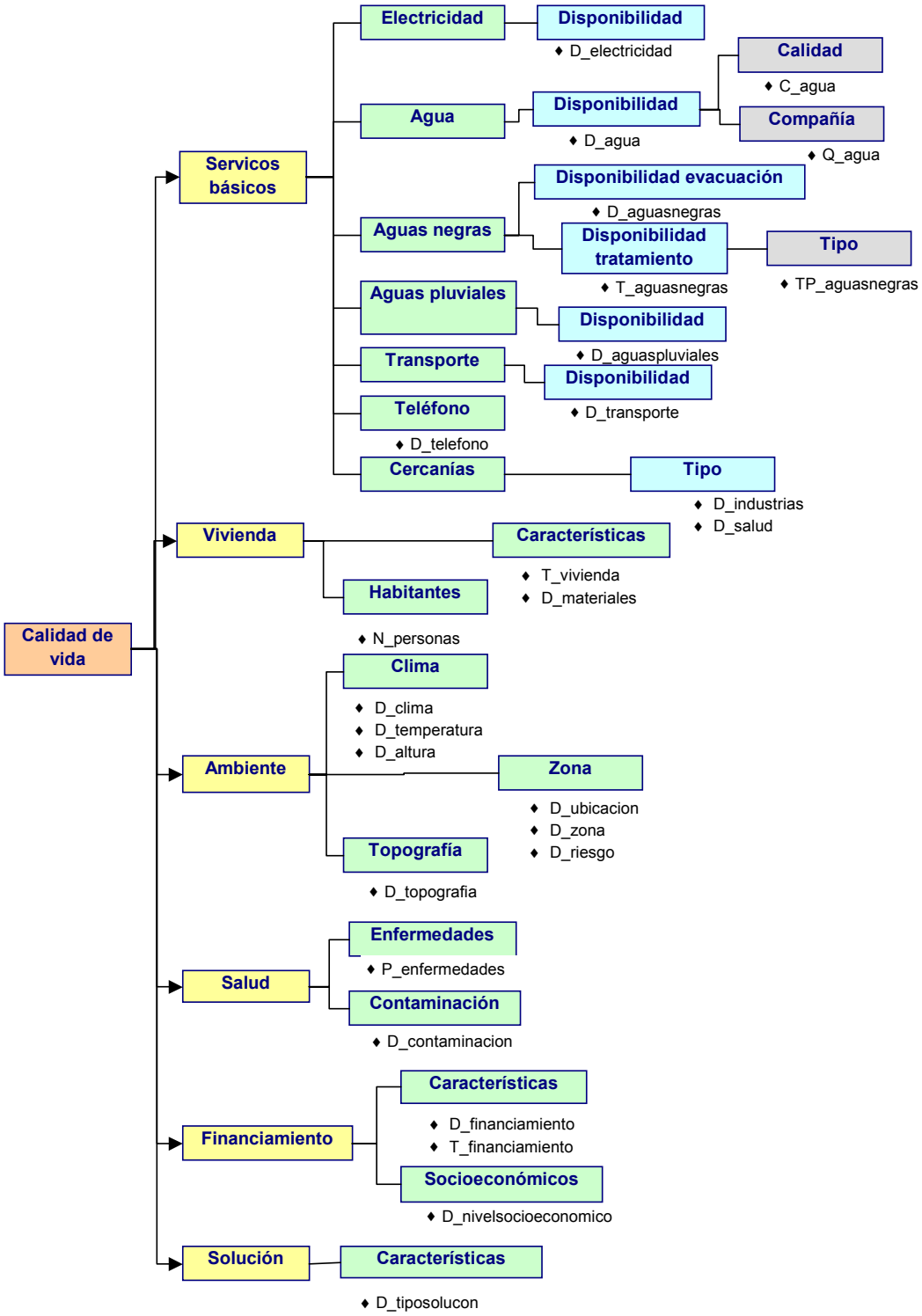


Figura 4-4. Diseño dimensional jerárquico

4.1.2 Uso de SQL Server 2000 – Analysis Server para el desarrollo del cubo

Para iniciar la creación de un cubo de datos utilizando las herramientas de Analysis Server, se debe contar primero con un depósito de datos sobre el cual se construirá el cubo. En el caso específico de este trabajo, se planteará un prototipo de datawarehouse diseñado con base en todas las consideraciones anteriores y se utilizará ACCESS 2000 para implementar ese prototipo.

Datawarehouse prototipo

Con base en el análisis anterior se plantea una base de datos en ACCESS 2000 la cual cuenta con 13 tablas más una tabla principal que será la tabla de hechos.

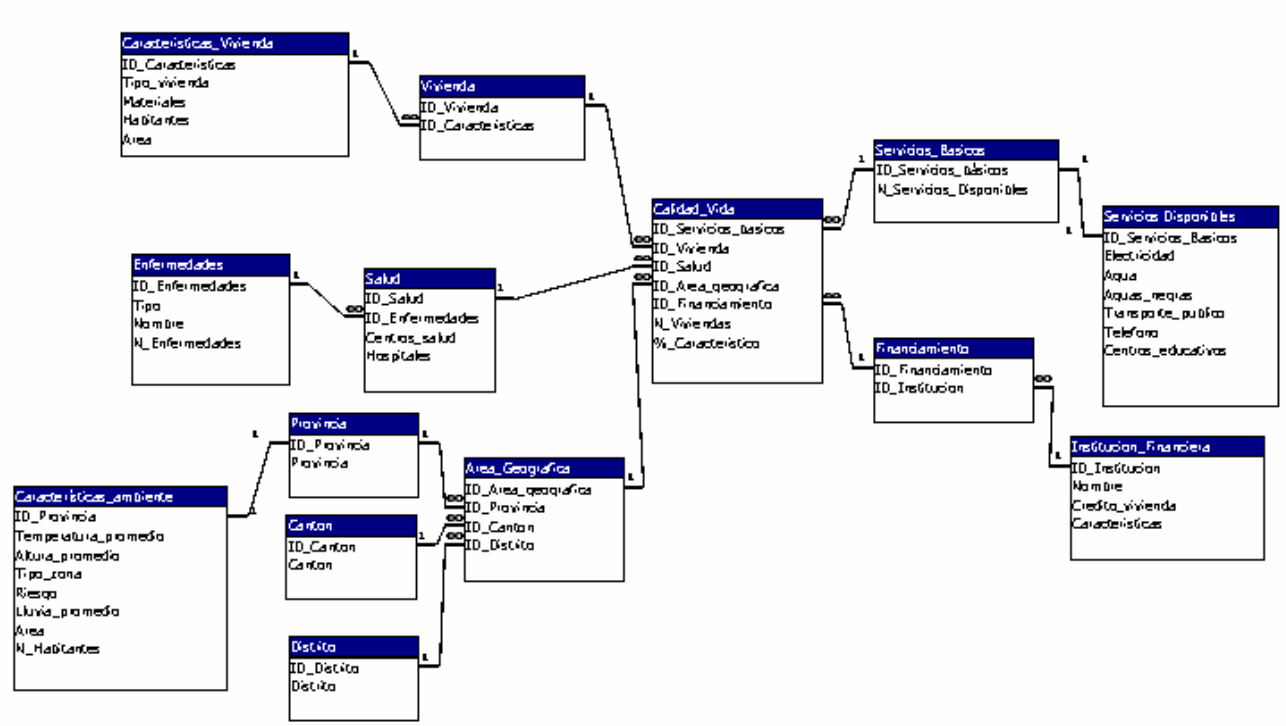


Figura 4-5. Relaciones existente en la base de datos prototipo[15]

[15] Se adjunta prototipo de base de datos “Prueba_DW_Indicadores”, creada en Access 2000.

Una vez creado este prototipo, se debe crear la base de datos usando las herramientas de Analysis Server del SQL Server 2000, utilizando el Analysis Manager. El nombre que se utilizó para esta base de datos fue "Prueba_DW_Indicadores".

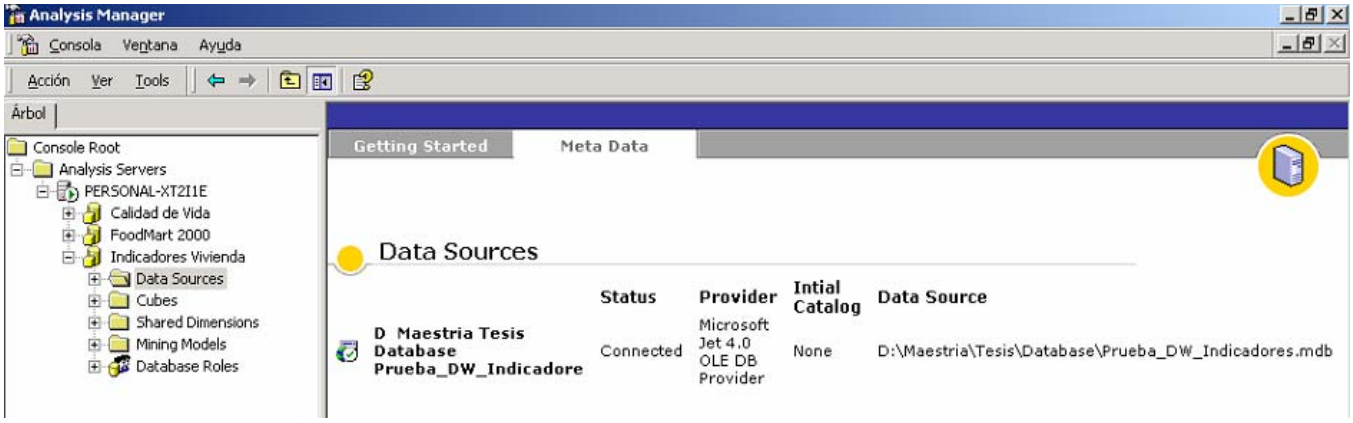


Figura 4-6. Ejemplo de pantalla una vez creada la base de datos

El siguiente paso es el diseño del cubo de datos sobre esta base de datos prototipo, para el caso específico de este trabajo, se realizará un diseño conceptual, ya que no se alimentará la base de datos en su totalidad, pues el objetivo es establecer un grupo de indicadores, su creación podrá hacerse posteriormente por las personas o instituciones interesadas.

Diseño del cubo de datos

Este diseño se basará en todo el análisis realizado anteriormente y con base en el depósito de datos prototipo elaborado en ACCESS 2000 y denominado "Prueba_DW_Indicadores".

Selección de la tabla de hechos y las medidas que se utilizarán:

La tabla de hechos que se utilizará es la denominada "Calidad Vida", esta contiene las llaves necesarias para establecer las relaciones con el resto de tablas, además se escogieron como atributos de medida: ***N_Viviendas*** el cual nos indicará el número de viviendas que se encuentran en una determinada condición y ***%_Caracteristico*** el cual nos indicará que porcentaje de viviendas o zonas del país se encuentran con determinadas características.

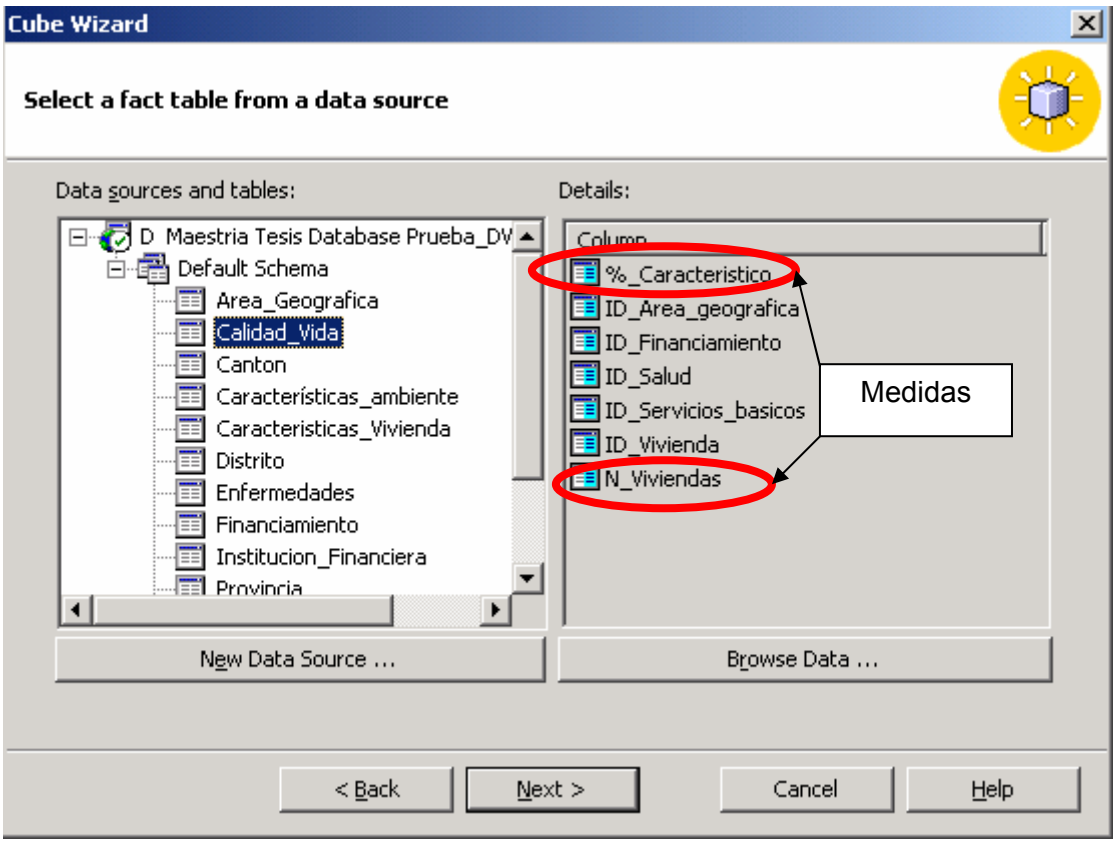


Figura 4-7. Ejemplo de pantalla creación del cubo, tabla de hechos

Una vez definida la tabla de hechos se deben definir las dimensiones, estás corresponderán a las definidas anteriormente son algunos cambios de nomenclatura.

En este cubo de definirán 5 dimensiones: Area geográfica, Salud, Vivienda, Servicios básicos y Financiamiento. Las dimensiones se crearán utilizando el snowflake schema (esquema de copo de nieve), éste se utiliza para tablas relacionales multidimensionales. Se deben seleccionar una o más columnas de la tablas relacionadas, cada columna especifica un nivel de dimensión.

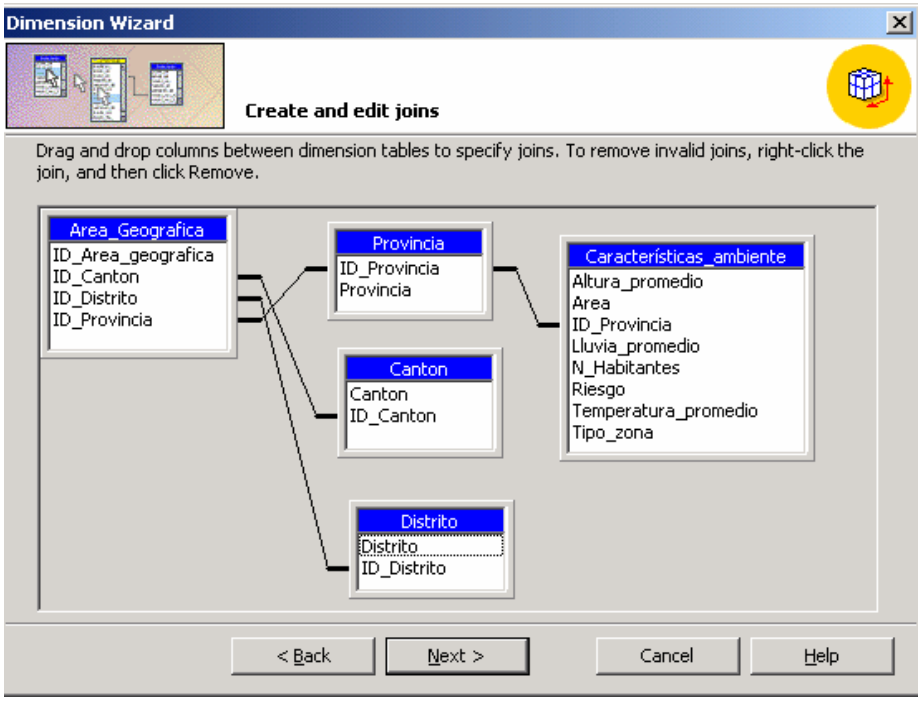


Figura 4-8. Ejemplo de pantalla definición de dimensiones, dimensión área geográfica

Una vez definida la dimensión y las relaciones (debido al esquema que se está utilizando), se debe definir el nivel de jerarquía en cada dimensión.

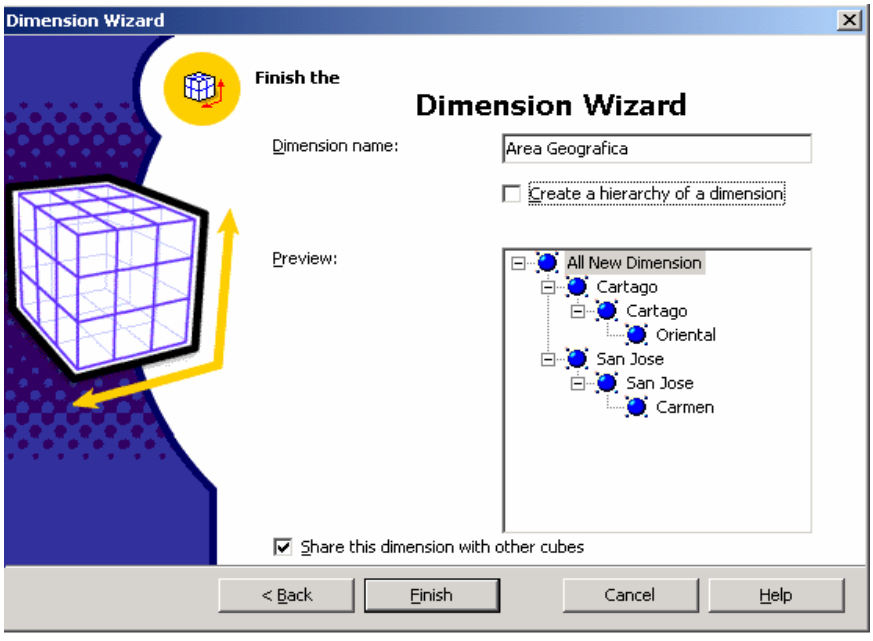


Figura 4-9. Ejemplo de pantalla definición de jerarquías, dimensión área geográfica

Cuando se definen las dimensiones y sus niveles de jerarquía, se tiene definido el cubo, en este caso el cubo se llama “Indicadores Vivienda”.

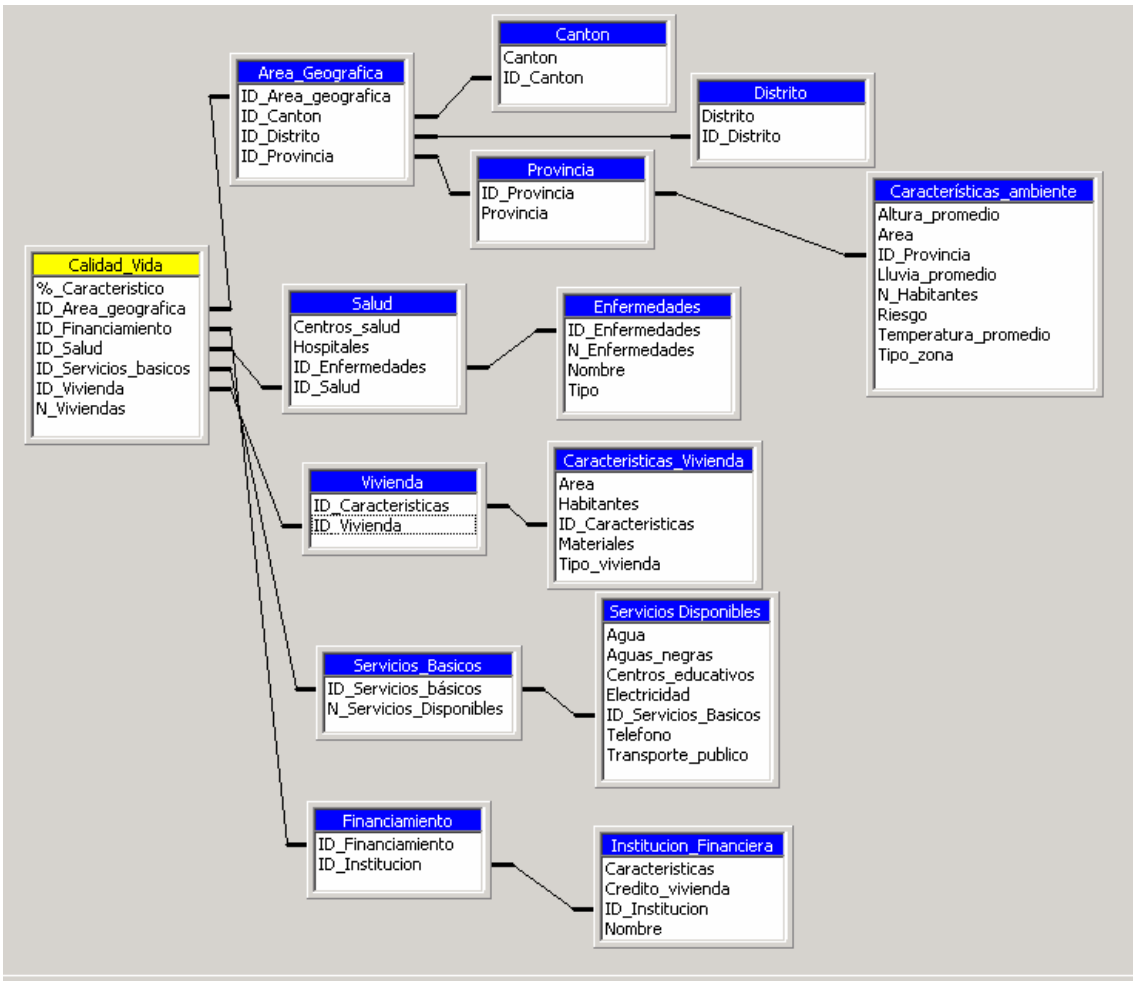


Figura 4-10. Ejemplo de pantalla definición de cubo, Indicadores Vivienda

Se tiene de esta manera el cubo definido, con un nombre y guardado. Por lo tanto se tiene definido en un depósito de datos OLAP, pero aún no se han creado los archivos, ya que primero debe procesarse.

Procesamiento del cubo

Cuando se procesa un cubo, el Analysis Services combina primero los mapas de dimensiones desde todas las dimensiones usadas en el cubo dentro de un mapa multidimensional del cubo.

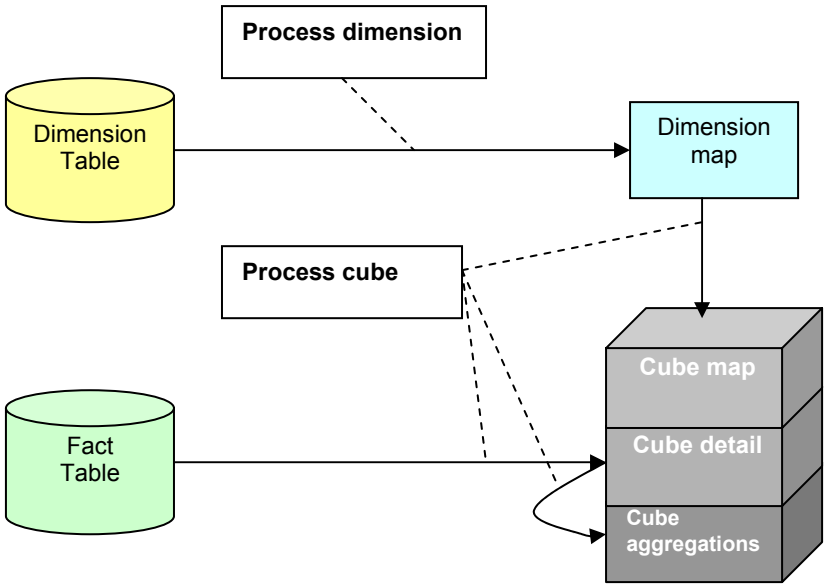


Figura 4-11. Procesamiento de un cubo[16]

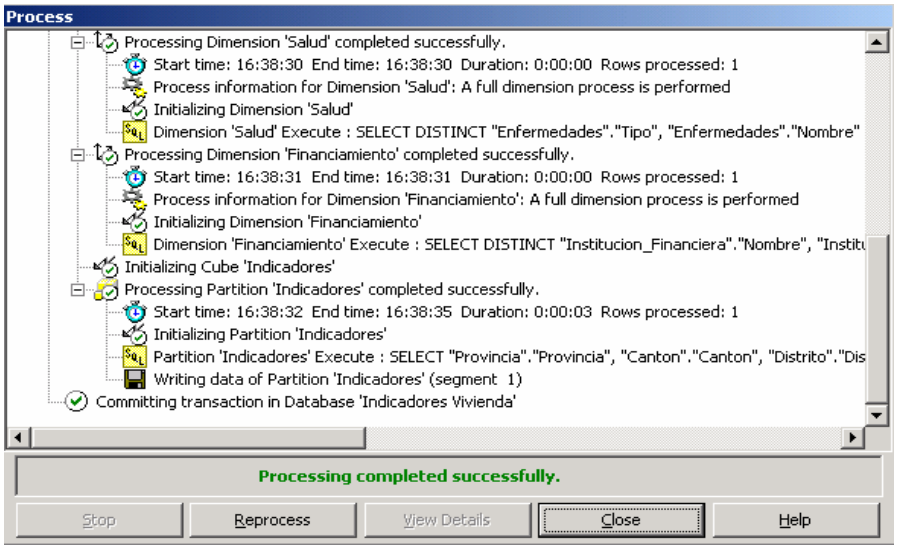


Figura 4-12. Ejemplo pantalla de procesamiento del cubo

Una vez creado y procesado el cubo se puede iniciar el proceso de búsqueda y análisis de la información, lo cual será la base para el establecimiento de los indicadores.

[16] Tomado de Microsoft SQL Server 2000, Analysis Server. OLAP Train, Redd Jacobson

4.2 Establecimiento de indicadores

Para el establecimiento de los indicadores, se procedió a alimentar el depósito de datos con algunos ejemplos, específicamente se utilizaron los datos de la provincia de Cartago. Es importante recalcar que aunque se utilizaron sólo estos datos las relaciones que se obtengan y los indicadores que se especifiquen pueden ser usados de forma general una vez que se cuente con el depósito de datos totalmente lleno.

Se puede analizar el cubo y visualizar los datos contenidos en él utilizando el browser del Analysis Manager. Por medio de esta herramienta se visualizará el cubo creado, al cual se le extraerá la información que se requiera.

La técnica utilizada es la de árbol de decisión, aunque también se puede hacer utilizando análisis de clustering.

La extracción de la información se hará a través de una pantalla similar a la mostrada en la figura 4-13, se identifican las dimensiones, las medidas y se extraerán posteriormente los indicadores.

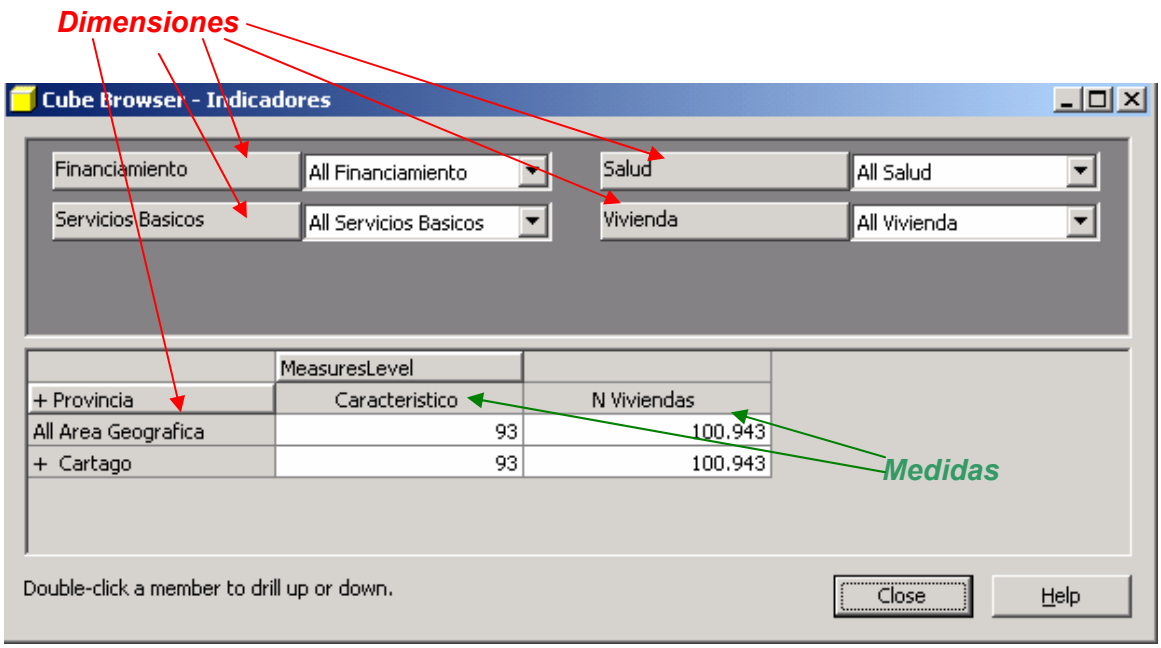


Figura 4-13. Visualización del cubo "Indicadores"

Para el establecimiento de los indicadores se tomará en cuenta lo señalado en el capítulo 2, referente a características de los indicadores; pero es importante señalar que estos se obtendrán utilizando la técnica de árbol de decisión, además se orientarán a las necesidades señaladas por el sector construcción.

Indicadores:

- Indicador Vivienda y Salud:

Este indicador me establecerá que relación existe entre una vivienda y la salud de las personas que la habitan.

Se comparan las características de la vivienda: área, materiales; los habitantes promedio y las enfermedades más comunes reportadas. Este indicador proporciona un número de viviendas con las características planteadas y un porcentaje del total de viviendas, lo cual se puede realizar por área geográfica. Obsérvese el siguiente ejemplo:

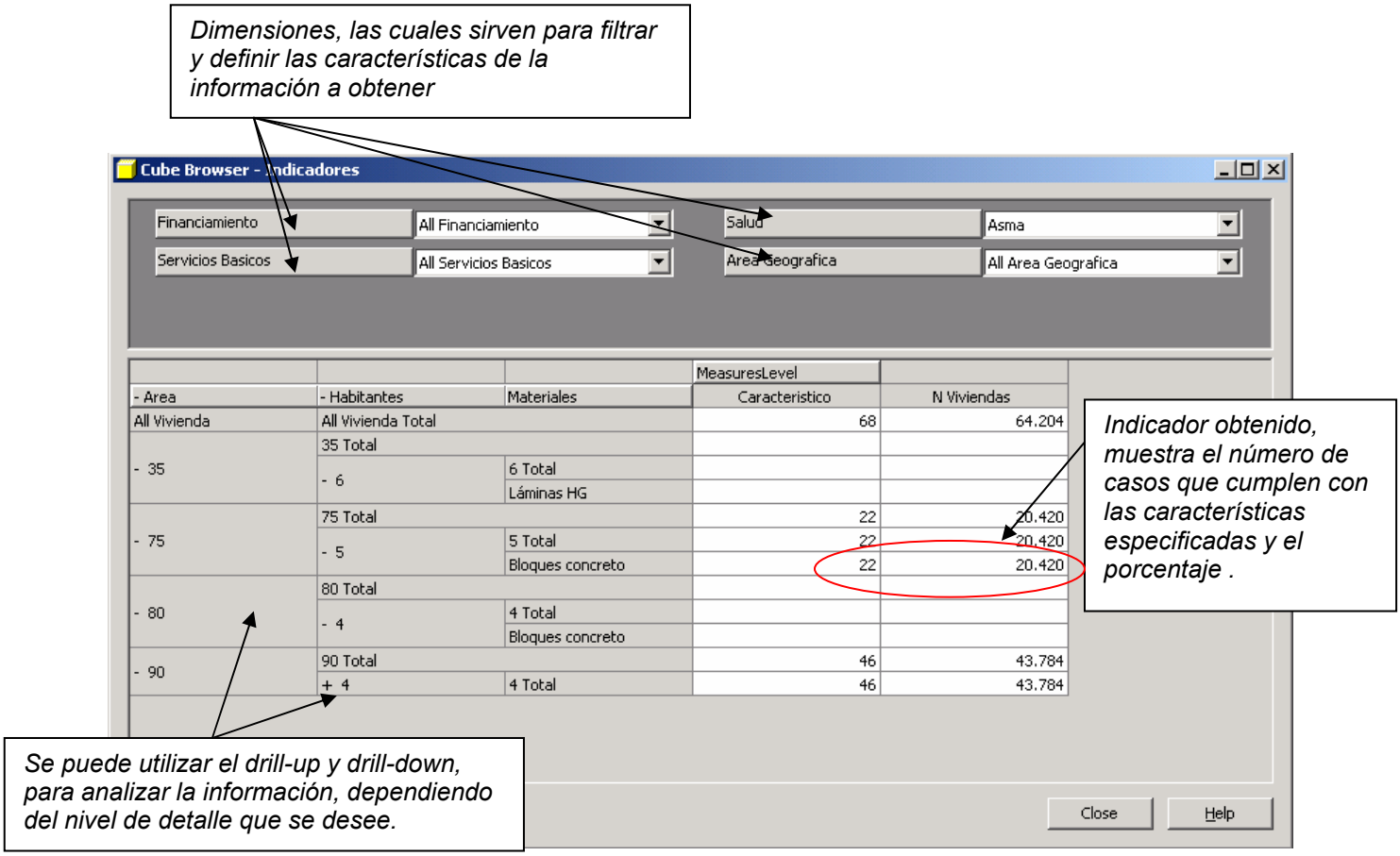


Figura 4-14. Ejemplo, indicador Vivienda y Salud

En este caso se están analizando los casos de asma, en toda en área geográfica (este ejemplo se enfoca en Cartago), con cualquier tipo de acceso a financiamiento y que cuenta con todos los servicios básicos.

La información obtenida aquí es de suma importancia, ya que puede ser la clave del porqué se están presentando algunas enfermedades, con cierta regularidad, y si estas tienen relación con el hábitat de las personas. Además, se puede introducir la variable tiempo a la información analizada, para ver su comportamiento a través del tiempo. El resultado final podría reflejarse en políticas en la construcción de viviendas, en el uso de determinado materiales, incluso relacionados con la zona geográfica.

❖ Indicador Accesibilidad al crédito

Este indicador establecerá que tanto acceso al crédito para vivienda tienen determinadas familias.

Aquí se propone analizar por zona geográfica y por características de vivienda la disponibilidad en determinada zona de crédito accesible y las características de ese crédito, por ejemplo, el tipo de institución que lo brinda, el tipo interés (fijo, fluctuante).

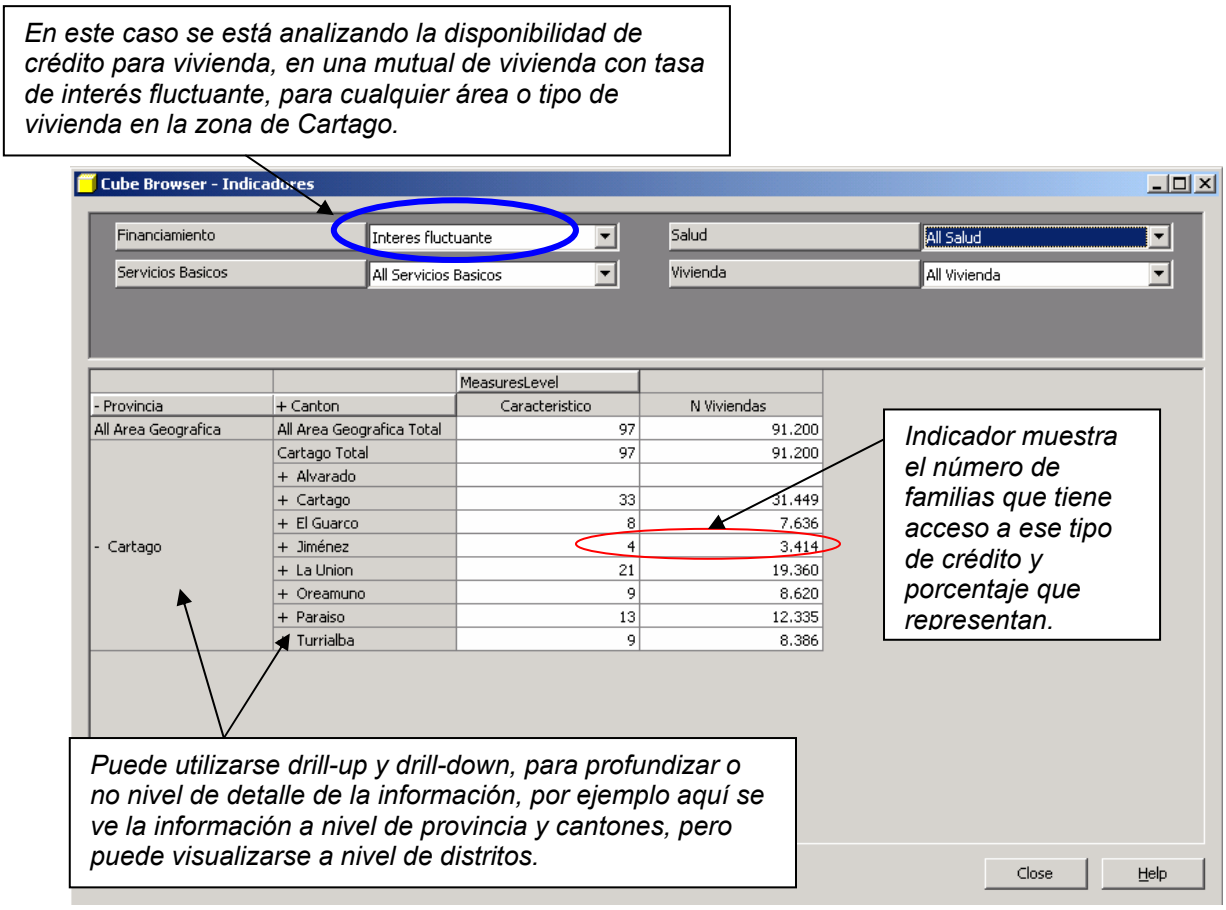


Figura 4-15. Ejemplo, indicador Accesibilidad al Crédito

Este tipo de información, también, es muy importante y su uso sería una excelente herramienta para la toma de decisiones del gobierno, ya que podría orientar sus políticas en este campo a satisfacer áreas con menores posibilidades de crédito, las cuales no necesariamente son las familias de menos recursos.

❖ Vulnerabilidad

Este indicador refleja la vulnerabilidad al riesgo natural (amenazas naturales) en una determinada zona.

Se analiza en este, no sólo la zona y su riesgo, sino las características de las viviendas y la relación de su sistema constructivo con el entorno. Para la elaboración de este indicador fue necesario la creación de una nueva dimensión denominada vulnerabilidad. Es importante destacar la facilidad con que se puede modificar el cubo creado y agregar otra dimensión en caso de ser necesario, lo cual se convierte en una ventaja.

En este caso se está analizando la vulnerabilidad (riesgo alto, moderado o bajo), dependiendo de la zona geográfica y las características de vivienda, principalmente.

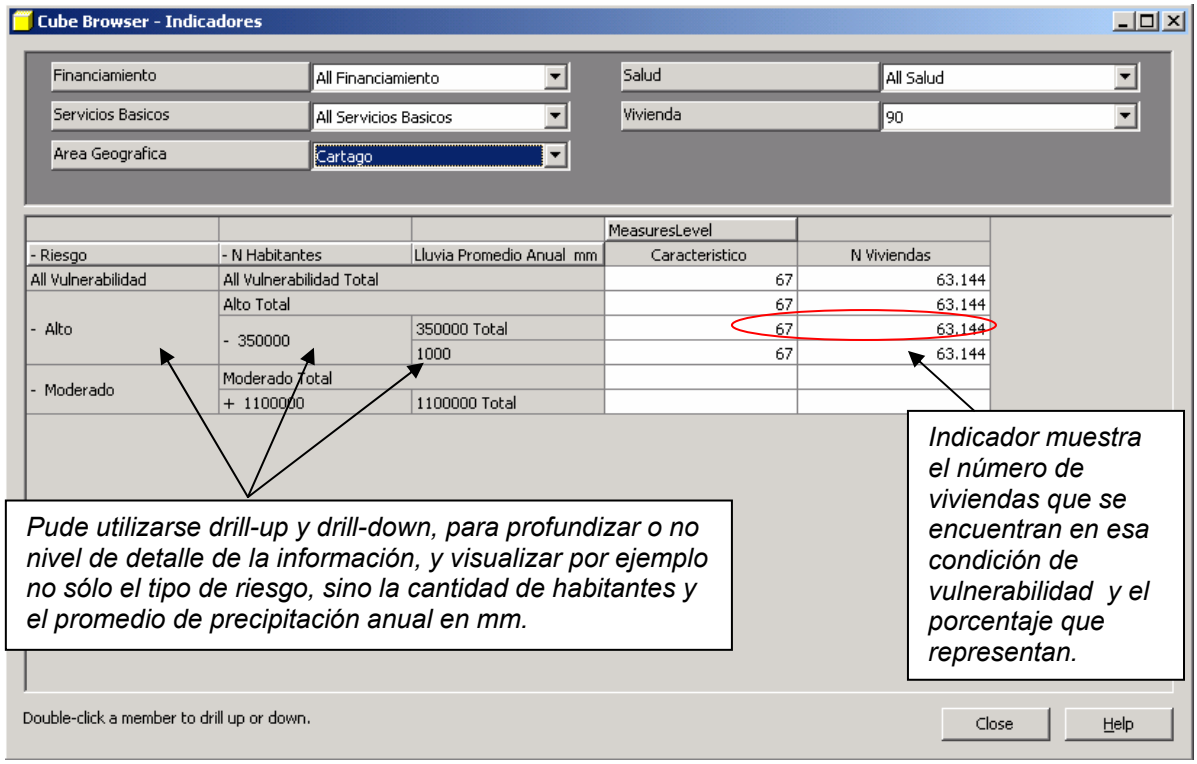


Figura 4-16. Ejemplo, indicador Vulnerabilidad

Esta información es de suma importancia, para establecer políticas de crecimiento urbano, reubicación y otorgamiento de servicios y permisos de construcción.

- Indicador Disponibilidad de Servicios Básicos y Salud

Este indicador representa la relación entre la disponibilidad de servicios básicos y los problemas de salud de la población.

Aquí se relacionan aspectos como área geográfica, disponibilidad de servicios básicos (más adelante si se mejoran los datos, se podrían relacionar la calidad de los servicios) y la incidencia de algunas enfermedades.

En este caso se está analizando la relación entre la disponibilidad de servicios básicos y la incidencia de algunas enfermedades, este análisis se hará por área geográfica.

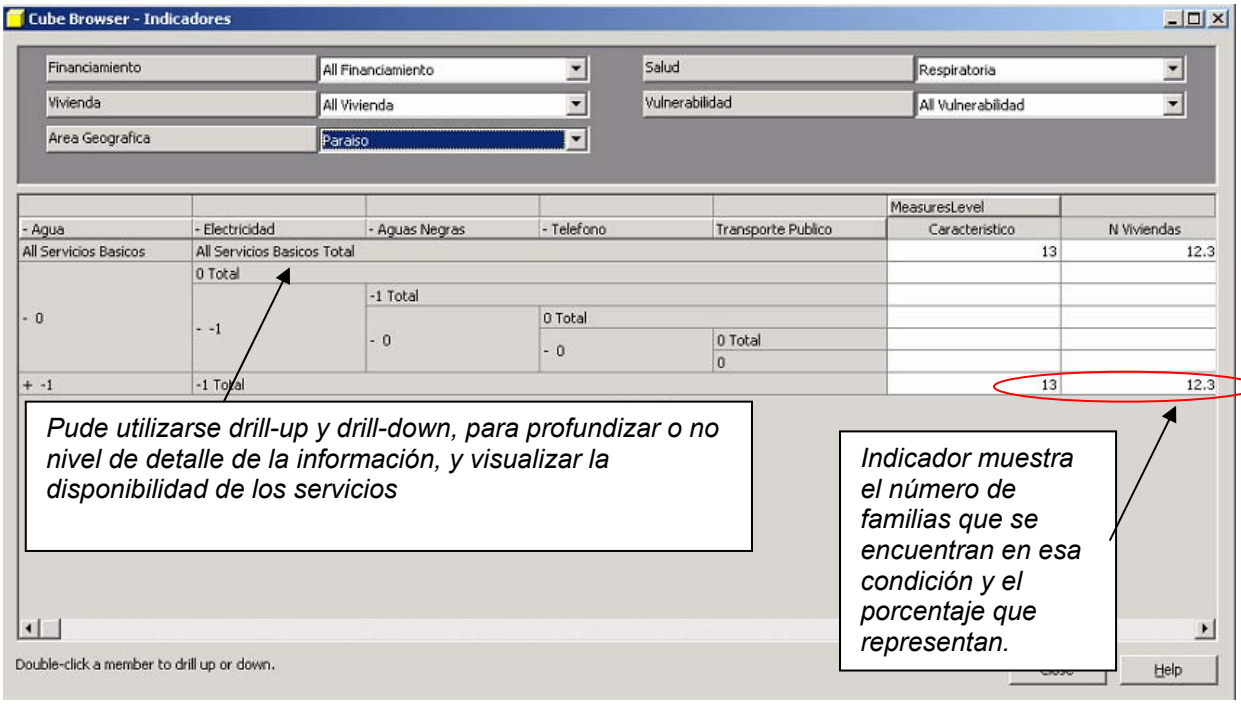


Figura 4-17. Ejemplo, indicador Servicios Básicos y Salud

Este tipo de información puede resultar muy provechosa para orientar las políticas de inversión en servicios públicos, tanto a nivel de instituciones del Estado, como a nivel municipal, además, también puede orientar la toma de decisiones en el sector salud.

CAPITULO 5. Conclusiones y Recomendaciones

5.1 Conclusiones

Una vez realizada la investigación se concluye que existe una gran cantidad de datos almacenados en diferentes instituciones a nivel nacional que podrían generar información para la toma de decisiones, pero que no está siendo utilizada adecuadamente en la toma de decisiones en el sector vivienda (interés de este trabajo). Uno de los problemas para utilizarla es la forma en que se encuentra almacenada ya que responden a diferentes motores de bases de datos y con niveles distintos de agregación. Se analizaron tres importantes fuentes de datos a nivel nacional, el Instituto Nacional de Estadística y Censos (INEC), el Ministerio de Salud y la Cámara Costarricense de la Construcción. En las dos primeras la información generada es pública y está disponible para su uso por parte de instituciones de carácter público. El Ministerio de Salud, en general, fue el más reacio a otorgar información para este trabajo, pero estaría de acuerdo en hacerlo para un proyecto de investigación a nivel de una institución pública, por ejemplo una universidad. La institución que posee mayor información es el INEC, y es la que se encuentra mejor organizada y con un marco legal que ampara su labor.

La minería de datos es una herramienta de la ciencia de la computación, específicamente en el área de Sistemas de Información, que puede aplicarse a cualquier otra ciencia o área de investigación. Una de esas áreas de gran importancia es la relación entre vivienda y salud, ya que mucho se ha dicho al respecto, pero ha faltado una verdadera investigación que demuestre con números la relación existente. La minería de datos puede ser esa llave, tal y como lo fue para descubrir la causa del cólera.

El uso del Analysis Server de SQL Server 2000, es una herramienta de gran ayuda, se puede tener un depósito de datos en otros motor que no sea SQL Server y aún así se puede utilizar para analizar los datos y extraer la información. En este caso la técnica utilizada fue la de árbol de decisión, en este caso particular funcionó bastante bien por el tipo de información que se quería analizar.

Se creó un cubo de información (OLAP) utilizando SQL Sever 2000. Este cubo se llamó Indicadores y se construyó sobre una base de datos prototipo llamada “Prueba_DW_Indicadores” creada en Access 2000. La construcción de este cubo se basó, principalmente, en cinco dimensiones: Area Geográfica, Financiamiento, Salud, Vivienda y Servicios Básicos; se utilizaron como medidas el número de viviendas con determinada característica y el porcentaje.

La etapa más difícil a la hora de construir un cubo de información que sirva como base a la toma de decisiones es la elaboración del depósito de datos, este debe ser un resumen de la información disponible en las bases de datos existentes, se resume en este trabajo un esquema para la construcción de este datawarehouse, no se construye el depósito definitivo, ya que era parte de los objetivos, pero si se construye un prototipo para probar el uso y utilidad de la minería de datos.

La cantidad, tipo y características de los indicadores que se pueden crear es bastante amplia, ya que dependerá de la utilización de las dimensiones y la propiedad de drill-up y drill-down que se utilice, en este caso se establecieron cuatro indicadores, a saber: Vivienda y Salud, Accesibilidad al Crédito, Vulnerabilidad y Disponibilidad de Servicios Básicos y Salud.

Es importante resaltar la versatilidad que presenta el Analysis Server de SQL Server 2000, para la creación de nuevas dimensiones en el cubo de datos y su actualización, lo cual lo convierte en una herramienta muy valiosa para generar información de valor agregado para la toma de decisiones.

El aprovechamiento de esta tesis dependerá de darle seguimiento por parte de alguna institución de importancia en el sector vivienda, la cual puede ser el Instituto Tecnológico de Costa Rica, a través del Centro de Investigaciones en Vivienda y Construcción y el Centro de Investigaciones en Computación.

5.2 Recomendaciones

Se sugiere continuar investigando en esta área estableciendo un proyecto conjunto entre el Centro de Investigaciones en Vivienda y Construcción (CIVCO), de la Escuela de Ingeniería en Construcción y el Centro de Investigaciones en Computación (CIC), del Departamento de Computación, ambos del Instituto Tecnológico de Costa Rica; para crear una oficina que pueda brindar información de valor agregado, que sea la base para la toma de decisiones en el sector vivienda. Para ello se deben establecer convenios con instituciones tales como el Ministerio de Vivienda y Asentamientos Humanos, el INEC, el Ministerio de Salud, la Cámara Costarricense de la Construcción, entre otros. Y según se puede ver, las instituciones están muy interesadas.

Continuar alimentando el depósito de datos. Para ésto se deberían crear algunas herramientas para la limpieza de los datos en forma automática, lo cual es una de la tareas más difíciles en la construcción de un datawarehouse.

Divulgar este informe a las instituciones interesadas para que mejoren en alguna medida la forma de tomar decisiones, de tal manera que estas sean basadas en información. Esta es una primera investigación aplicando los conceptos hacia el sector construcción, pera la aplicabilidad de éstos puede y debe ir más allá. El fomentar que las instituciones o personas que tienen en sus manos la toma de decisiones cuente con información para poder hacerlo, es una de las metas a alcanzar co esta y futuras investigaciones en este campo el Instituto tecnológico de Costa Rica debe procurarlo a través de la investigación y extensión.

Propiciar la interrelación de la computación con otras áreas y resaltar su aplicabilidad y utilidad en la resolución de problemas en otras disciplinas. Como se observó en este trabajo, el utilizar herramientas y técnicas de la computación para enfrentar y tratar de solucionar un grave problema nacional, tiene aplicación no sólo en el caso de la vivienda, sino que podría utilizarse en muchas otras problemáticas.

También sería conveniente, que se continúe investigando en cómo mejorar la accesibildad a la información generada, de tal manera que un usuario común pueda visualizarla y accederla de una manera fácil y amigable. Recordando que lo que se desea con este tipo de herramientas es ayudar en la toma de decisiones, las cuales no van a estar, en la mayoría de los casos, en manos de expertos en computación.

Bibliografía

[AGRA96] Agrawal, Rakesh; Gerhrke, Johannes; Gunopulos, Dimitrios & Raghavan, Prabhakar. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. IBM Almaden Research Center. 1996.

[AGRA99] Agrawal, Rakesh; Imielinski, Tomasz & Swami, Arun. Database Mining: A Performance Perspective. 1999.

[BULL98] Bullen, Paul. Performance Indicators. 1998. http://www.mapl.com.au/A1A.htm

[CHEN96] Chen, Ming-Syan; Han, Jiawel & Yu, Philip S. Data Mining: An Overview from a Database Perspective.1996.

[DEHA01] Dehaspe, Luc & Toiven, Hannu. Data Mining and Kwoledge Discovery. Department of Computer Science, Katholieke Universitiet. 2001.

[FAYY96] Fayyad, U.M; Piatetsky-Shapiro, G.; Smyth, P. & Uthurusamy, R. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press. 1996.

[FLAN98] Flanagan, Thomas & Safdei, Elias. DNA and Datawarehouse

[GONZ01] González Alvarado, Carlos. Cubos en SQL Server 2000. Material para el curso Administración y Optimización de Bases de Datos. Instituto Tecnológico de Costa Rica. 2001.

[HUSE00] Hüsemann, Bodo; Lechtenbörger, Jens and Vossen, Gottfried. Conceptual Data Warehouse Design. Universität Münster, Germany. 2000.

[INEC01] Instituto Nacional de Estadística y Censos. IX Censo Nacional de Población y V de Vivienda del 2000: Resultados generales. INEC. San José, Costa Rica. 2001.

[JOHN01] Johnston, Steven. Data in the time of Cholera. 2001

[MSCR00] Ministerio de Salud de Costa Rica. Memoria Anual. Ministerio de Salud. 2000.

[PNDU01] Plan Nacional de Desarrollo Urbano, Area temática: Vivienda y Asentamientos Humanos. Informe final del Taller. Junio 2002.

[SQLS00] SQL Server Magazine. Data Mining en SQL Server 2000. http://www.w2000mag.com/sqlmag/atrasados/04_mar01/articulos/portada_1.htm

[VASS02] Vass, Valor Añadido Soluciones y Servicios. Construcción de indicadores de gestión y herramientas OLAP para pequeñas y medianas empresas. http://www.vass-consult.com/BI-Indicadores%20de%20gestion.doc