UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
E.A.P DE INGENIERÍA DE SISTEMAS

Una metodología para sectorizar pacientes en el consumo de medicamentos aplicando Datamart y Datamining en un Hospital Nacional

TESIS Para optar el título profesional de: INGENIERO DE SISTEMAS.

AUTOR:
Iván Gildo Tapia Rivas

LIMA – PERÚ 2006

DEDICATORIA

Este trabajo lo dedico a mis padres, y en especial a mi madre, quien me apoyó, como en otras ocasiones, durante todo este anhelado trabajo. A mi esposa y a mi hijo, quienes con su apoyo y presencia, me incentivaron hacia la culminación del mismo.

AGRADECIMIENTO

A la Universidad Nacional Mayor de San Marcos – Facultad de Ingeniería de Sistemas e Informática por haberme brindado la oportunidad de incrementar mis conocimientos, a los Señores catedráticos por sus orientaciones y sabios consejos, que me encaminaron hacia la superación y culminación de mis estudios.

RESUMEN

La Minería de Datos (Data Mining) es la búsqueda de patrones interesantes y de regularidades importantes en grandes bases de datos. La minería de datos inteligente utiliza métodos de aprendizaje automático para descubrir y enumerar patrones presentes en los datos.

Una forma para describir los atributos de una entidad de una base de datos es utilizar algoritmos de segmentación o clasificación.

El presente trabajo, propone un método para el análisis de datos, para evaluar la forma con la que se consumen los medicamentos en un hospital peruano, poder identificar algunas realidades o características no observables que producirían desabastecimiento o insatisfacción del paciente, y para que sirva como una herramienta en la toma de decisión sobre el abastecimiento de medicamentos en el hospital.

En esta investigación, se utilizan técnicas para la Extracción, Transformación y Carga de datos, y para la construcción de un Datamart, para finalmente un algoritmo de minería de datos adecuado para el tipo de información que se encuentra contenida.

Palabras Clave: Minería de Datos, Datamining, aprendizaje automático, Datamart, Inteligencia de Negocio, Algoritmo K-Means, Algoritmo de Clasificación.

ABSTRACT

The Mining of Data (Mining Data) is the search of interesting patterns and important regularities in great data bases. The intelligent mining of data uses methods of automatic learning to discover and to enumerate present patterns in the data.

A form to describe the attributes of an organization of a data base is to use algorithms of segmentation or classification.

The present work, proposes a method for the analysis of data, to evaluate the form with which the medicines in a Peruvian hospital are consumed, to be able to identify some non-observable realities or characteristics which they would produce shortage of supplies or dissatisfaction of the patient, and so that it serves as a tool in the decision making on the medicine supplying in the hospital.

In this investigation, techniques for the Extraction, Transformation and Load of data are used, and for the construction of a Datamart, finally an algorithm of mining of data adapted for the type of information that is contained.

Key words: Mining of Data, Datamining, automatic learning, Datamart, Intelligence of Business, K-Means Algorithm, Classification Algorithm.

INDICE

CAPÍTULO I: PLANTEAMIENTO METODOLÓGICO

CAPÍTULO II: MARCO TEÓRICO

CAPÍTULO III: MODELADO Y CONSTRUCCIÓN DE UN DATAMART COMO FUENTE DE INFORMACIÓN PARA UN ALGORITMO DE MINERÍA DE DATOS EN HOSPITAL NACIONAL

CAPÍTULO IV: RESULTADOS, DISCUSIÓN E INTERPRETACIÓN DE LA INVESTIGACIÓN

INDICE DE TABLAS

INDICE DE GRAFICOS

INTRODUCCIÓN

Día con día, las empresas vienen creando diversos sistemas para poder resolver problemas específicos, ya sea por área, por sucursal, por unidad de negocio o en su total. A través del tiempo, las empresas necesitan que la información, contenida en diversos almacenes de datos, de distinta arquitectura y diseño, sea usada para consultas simples, o complejas.

Por otro lado, la necesidad de tener la seguridad de conocer todos los procesos dentro de un negocio es cada vez más importante, y puede ser el único diferenciador para que las empresas subsistan o perezcan.

La Minería de Datos o Datamining, es un método eficiente para contribuir en acortar esta brecha, y brindar fuentes de información mas eficientes a los agentes decisores.

Este trabajo se divide en 4 partes, cada una de las cuales se describe a continuación, a manera resumida:

En el **Capitulo I** se realiza una explicación de los Antecedentes del problema, la justificación, y se define los objetivos generales y específicos de la tesis. Por último se define la hipótesis a formular. Aquí también se definen las variables dependientes e independientes.

En el **Capítulo II** se define el Marco teórico, donde se fundamenta las actuales situaciones por las que pasa toda empresa para poder obtener conocimiento.

Se empieza definiendo el antecedente, los conceptos con los que se ha contado en el proyecto, para poder ir introduciendo al lector con el trabajo realizado. Tratamos sobre temas como Datawarehouse (definición y arquitectura tanto física como lógica), Datamart, Almacenamiento OLAP (cubos, dimensiones, métricas, jerarquías y estrategias de almacenamiento).

Veremos también, conceptos sobre transformación y carga de datos, y por último tratamos conceptos de minería de datos y específicamente veremos el algoritmo K-means, que es el utilizado en el proyecto.

En el **Capítulo III** se realiza una visión general del proyecto, se define y explica los pasos que comprende la metodología, y se desarrollan los mismos. Aquí es donde doy a conocer la fuente de datos con los que contaremos en el proyecto, vemos como diseñamos y poblamos nuestro datamart. Por último, aplicamos el algoritmo k-

means sobre la muestra en estudio. Definimos los clusters obtenidos y definimos algunas conclusiones previas.

En el **Capítulo IV** se hace una breve explicación del hospital en estudio, para ubicar al lector en el contexto deseado. Se detallan los resultados, discusión e interpretación de la investigación. Para esto, se hace uso de los cuestionarios que se hicieron a la población, que sirve como fundamento de los resultados del proyecto.

Finalmente se presentan conclusiones y se plantean algunas recomendaciones que se pueden aplicar, que complementen el trabajo presentado.

CAPÍTULO I
PLANTEAMIENTO METODOLÓGICO

1.1 PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACIÓN.

Las organizaciones dedicadas a la atención de la salud, como muchas de sus pares en otras áreas de la economía, asisten a un proceso de creciente informatización. La mayor parte de las aplicaciones aún se vinculan con procesos netamente administrativo-contables, pero el grado de informatización de datos estrictamente médicos es cada vez mayor. Las Base de Datos Transaccionales propias de la organización médica en estudio no escapa a los problemas que afectan a las organizaciones de los otros sectores, y los analistas se enfrentan a los mismos problemas de “encarcelamiento” de los datos.

El control de medicamentos, y específicamente, su abastecimiento a tiempo, es uno de los problemas con más repercusión en los procesos del hospital en estudio.

Los responsables de la administración y provisión de los mismos, cuentan con su experiencia ganada con los años, y algunos reportes extraídos desde el Sistema de Gestión Hospitalaria, para saber cuándo reabastecer a las farmacias respectivas con más medicamentos para no perjudicar a la gran mayoría de la población que se atiende en el hospital. Es decir, el personal encargado, reabastece su inventario en base a banderas que indican si se llegó a un nivel mínimo para reabastecer.

Si bien es cierto, el Hospital cuenta con una gran cantidad de información contenida en su sistema central, ésta, se encuentra encerrada en los almacenes operacionales, y no se la considera como lo que es realmente, es decir, una fuente importante de conocimiento que puede ser útil para la institución y sus objetivos.

En entrevistas con los responsables del área de Farmacia, se supo que la información con la que cuentan, es a modo de reportes estadísticos, donde se visualizan números que informan el estado del stock de los medicamentos. Adicional a los reportes ya existentes, constantemente surgen necesidades para obtener información basándose en nuevos y diversos criterios, para lo que se recurre al área de sistemas, y se solicita los cambios en la emisión del reporte. Vemos aquí la dependencia que se tiene con el área de sistemas, y la inflexibilidad con la que se puede obtener información.

La indisponibilidad de información inmediata que tiene el Responsable del Área, de tener el conocimiento adecuado de su negocio. El tener que depender de procesos repetitivos para la obtención de información. El percibir al negocio en términos estadísticos, mas no con criterios analíticos para identificar conocimiento, son algunos de las debilidades actuales reconocidas.

En resumen. Vemos entonces, que no se tiene un modo de conocer, de antemano, porqué es que ciertos medicamentos son muy despachados, ni tampoco se conoce el tipo de pacientes que solicitan medicinas de un tipo y no de otro, o si los números que aparecen en los reportes, contienen algo más de información que solo valores estadísticos.

1.2 DELIMITACIÓN DEL PROBLEMA.

1.2.1 AMBITO INSTITUCIONAL.

Sector : Salud

Institución : Hospital Nacional "Guillermo Almenara Irigoyen"

1.2.2 AMBITO EMPRESARIAL.

Actividad : Servicio de Salud

1.2.3 AMBITO TERRITORIAL.

Ubicación : Lima - Perú

1.2.4 AMBITO TEMPORAL.

Es una investigación de actualidad (2006).

1.3 FORMULACIÓN DEL PROBLEMA.

1.3.1 PROBLEMA PRINCIPAL.

¿Qué efectos produce en el hospital, la falta de una metodología efectiva que permita sectorizar a los pacientes con respecto al consumo de medicamentos, para colaborar en una correcta toma de decisiones?

1.3.2 SUB–PROBLEMAS.

a. ¿Cómo es que el transformar la información correspondiente a un período de trabajo del sistema actual almacenado en tablas planas, hacia una Base de Datos Relacional contribuye a establecer objetivos importantes en la Toma de Decisiones?

b. ¿De qué manera el modelar, construir y cargar una Base de Datos Relacional hacia un Datamart para consultas OLAP ayuda a establecer estrategias para la Toma de Decisiones?

c. ¿Cómo es que la generación de pruebas de clasificación, utilizando el algoritmo de Minería de Datos: K-means para encontrar características similares en la información permite implantar procedimientos en la Toma de Decisiones?

1.4 OBJETIVOS.

1.4.1 OBJETIVO GENERAL.

Definir una metodología para el Modelado y Construcción de un Datamart como fuente de información para el algoritmo de Minería de Datos K-means, que ayude a la Toma de Decisiones para Sectorizar Pacientes en el Consumo de Medicamentos en un Hospital Nacional.

1.4.2 OBJETIVOS ESPECÍFICOS.

a. Demostrar como la transformación de la información correspondiente a un período de trabajo del sistema actual almacenado en tablas planas hacia una Base de Datos Relacional permite establecer objetivos a alcanzar en la Toma de Decisiones.

b. Conocer como el modelar, construir y cargar una Base de Datos Relacional a un Datamart coadyuva a establecer estrategias para la Toma de Decisiones.

c. Demostrar como el generar pruebas de clasificación, utilizando el algoritmo de Minería de Datos K-means para encontrar características

similares en la información permite identificar conocimiento en la Toma de Decisiones.

1.5 HIPÓTESIS.

1.5.1 HIPÓTESIS GENERAL.

El uso de Datamart y la aplicabilidad del algoritmo K-means permiten la mejora del proceso de toma de decisiones en el consumo de medicamentos en un hospital nacional.

1.6 VARIABLES.

1.6.1 VARIABLE DEPENDIENTE.

Mejora el proceso de Toma de Decisiones en el consumo de medicamentos en un hospital nacional (VD)

1.6.2 VARIABLE INDEPENDIENTE.

Datamart, Algoritmo de minería de datos K-means (VI)

1.7 DISEÑO Y TIPO DE INVESTIGACIÓN.

El presente estudio es de tipo Descriptivo. Se establece específicamente en el Hospital Nacional Guillermo Almenara Irigoyen; se identifica sus procedimientos, desempeño y la actitud o percepción de los pacientes sobre aspectos conformantes del bienestar.

Asimismo, se definen pasos a seguir para obtener resultados que apunten a conseguir una mejor toma de decisión en el abastecimiento de medicamentos.

1.8 POBLACIÓN Y MUESTRA.

1.8.1 DESCRIPCIÓN DE LA POBLACIÓN.

Para el presente estudio, se consideró como población a las personas que trabajan y son atendidas en el Hospital Nacional Guillermo Almenara Irigoyen.

1.8.2 TAMAÑO DE LA MUESTRA.

La muestra que se determinó fueron las personas que tienen que ver con el Área de Informática y Jefaturas del Departamento de Farmacia del Hospital Nacional Guillermo Almenara Irigoyen.

1.9 JUSTIFICACIÓN O RELEVANCIA DE LA INVESTIGACIÓN PROPUESTA.

- La justificación del trabajo, se centra en el hecho de que se propone una solución al requerimiento de un problema que actualmente tiene el área de farmacia del hospital Almenara. Identificar las características de los pacientes que acuden al hospital, y que hacen que el stock de medicamentos sufra picos de desabastecimiento constantemente, es motivo suficiente para tratar de conocer mejor el negocio de farmacia, y tener bases mas completas para poder requerir reabastecer de medicamentos en el área.

- Como consecuencia, se podrá identificar clases de pacientes que acuden al centro hospitalario, y poder probablemente, crear campañas orientadas al tipo de paciente, o auditar mejor a los médicos que suscriben determinados medicamentos para pacientes que podrían tomar otros. En fin, se convierte en un aporte importante, con futuras mejoras, que podrá servir como fuente de conocimiento para el área.

- Otra justificación está en el hecho de que es una opción con la que se demuestra que la aplicación de algoritmos de minería de datos simples, puede ser una opción viable para analizar información del Sistema de Gestión Hospitalaria.

- Es una opción también, de demostrar que se puede ir aminorando la dependencia del área de farmacia con la de sistemas, para obtener información en cualquier instante, produciendo así un mejor desempeño de ambas áreas.

CAPÍTULO II
MARCO TEÓRICO

2.1 ANTECEDENTES.

Como fruto de la búsqueda realizada por el autor, se ha encontrado algunas conceptualizaciones sobre la investigación, que han servido para la elaboración del presente trabajo y aparecen consignados en el marco conceptual.

Sin embargo, cabe indicar que hasta el presente no se han desarrollado trabajos sobre Datamart, Datamining y Toma de Decisiones en el ámbito del Consumo de Medicamentos y, específicamente, en el Hospital Nacional Guillermo Almenara Irigoyen.

Asimismo, con relación a las variables del tema, no se han encontrado investigaciones que hayan abordado estos temas aplicados a la problemática planteada, con lo cual consideramos que la presente investigación reúne las condiciones metodológicas suficientes para ser considerada inédita.

2.2 CONCEPTOS SOBRE DATAWAREHOUSE, DATAMART Y OLAP.

2.2.1 DATAWAREHOUSE.

Un DataWarehouse es un repositorio central o colección de datos en la cual se encuentra integrada la información de

la organización y que se usa como soporte para el proceso de toma de decisiones gerenciales.

El concepto de DataWarehouse comenzó a surgir cuando las organizaciones tuvieron la necesidad de usar los datos que cargaban a través de sus sistemas operacionales para planeamiento y toma de decisiones.

Para cumplir estos objetivos se necesitan efectuar consultas que sumarizan los datos, y que si se hacen sobre los sistemas operacionales reducen mucho la performance de las transacciones que se están haciendo al mismo tiempo. Fue entonces que se decidió separar los datos usados para reportes y toma de decisiones de los sistemas operacionales y así, diseñar y construir los llamados DataWarehouses para almacenar estos datos.

Las principales características que posee un DataWarehouse son:

• Es orientado a la información relevante de la organización: En un DataWarehouse la información se clasifica en base a los aspectos de interés para la empresa, es decir, se diseña para consultar eficientemente información relativa a las actividades básicas de la organización, como ventas, compras y producción, y no para soportar los procesos que se realizan en ella, como gestión de pedidos, facturación, etc.

• Es integrado: integra datos recogidos de diferentes sistemas operacionales de la organización y/o fuentes externas. Esta integración se hace estableciendo una consistencia en las convenciones para nombrar los datos, en

la definición de las claves, y en las medidas uniformes de los datos.

• Es variable en el tiempo: los datos son relativos a un periodo de tiempo y deben ser incrementados periódicamente. La información almacenada representa fotografías correspondientes a ciertos períodos de tiempo.

• Es no volátil: la información no se modifica después de que se inserta, solo se incrementa. El periodo cubierto por un DataWarehouse varía de 2 a 10 años. [PATRICIA ZVENGER]

Arquitectura Datawarehouse

Podemos dividirla en dos tipos:

- **Diseño Lógico**.

  De acuerdo a [PATRICIA ZVENGER], existen algunos requerimientos que debe cubrir un diseño lógico para un Datawarehouse.

  - Preparar el datawarehouse para soportar la recuperación de una gran cantidad de filas de datos en forma rápida.

  - La mayoría de los analistas de negocios van a querer ver datos totalizados. Estos datos en lo posible deben precalcularse y almacenarse de antemano para que esta recuperación sea rápida y eficiente. Es importante además discutir el nivel de granularidad y de detalle esperado por los analistas cuando hacen operaciones de DRILLDOWN.

- El diseño debe estar conducido por el acceso y por el uso, es decir, teniendo en cuenta qué tipo de reportes o resúmenes son los más frecuentes, y cuáles los más urgentes.
- Un diseño normalizado no es bueno, no solo por lo mencionado en la sección anterior, sino porque no resulta demasiado intuitivo para una persona de negocios, y podría volverse demasiado complejo.
- Todos los datos que se incluyan ya deben existir en las fuentes de datos operacionales, o ser derivables a partir de ellos. [PATRICIA ZVENGER]

Las dos técnicas de diseño más populares de almacenamiento lógico de un datawarehouse son las siguientes:

Esquema Estrella.

Este esquema está formado por un elemento central que consiste en una tabla llamada la Tabla de Hechos, que está conectada a varias Tablas de Dimensiones.

Las tablas de hechos contienen los valores precalculados que surgen de totalizar valores operacionales atómicos según las distintas dimensiones, tales como clientes, productos o períodos de tiempo.

Las tablas de hechos representan un evento crítico y cuantificable en el negocio, como ventas o costos. Su clave está compuesta por las claves primarias de

las tablas de dimensión relacionadas (las FOREIGN KEYS). Pueden existir varias tablas de hechos con información redundante, porque podrían contener distintos niveles de agregación de los mismos datos. Por ejemplo podría existir una tabla de hechos para las Ventas por Sucursal, Región y Fecha, otra para Ventas por Productos, Sucursal y Fecha, y otra para Ventas por Cliente, Región y Fecha.

En general las tablas de hechos tienen muchas filas y relativamente pocas columnas.

Las tablas de dimensión representan las diferentes perspectivas desde donde se ven y analizan los hechos de la tabla de hechos. A diferencia de las anteriores, su clave primaria está formada por un solo atributo, y su característica principal es que están denormalizadas. Esto significa que si la dimensión incluye una jerarquía, las columnas que la definen se almacenan en la misma tabla dando lugar a valores redundantes, lo cual es aceptable en este esquema.

En general suelen tener muchas columnas pero pocas filas. Siempre que sea posible, es conveniente compartir las tablas de dimensión entre distintas tablas de hechos.

Una de las dimensiones mas comunes es la que representa el tiempo, con atributos que describen periodos para años, cuatrimestres, periodos fiscales, y periodos contables.

Otras dimensiones comunes son las de clientes, productos, representantes de ventas, regiones, sucursales.

El esquema estrella es el más usado porque maneja bien la performance de consultas y reportes que incluyen años de datos históricos, y por su simplicidad en comparación con una base de datos normalizada.

En la siguiente figura vemos un ejemplo de esquema Estrella, donde la tabla de hechos es la tabla Ventas, y el resto son las tablas de dimensiones. [PATRICIA ZVENGER]

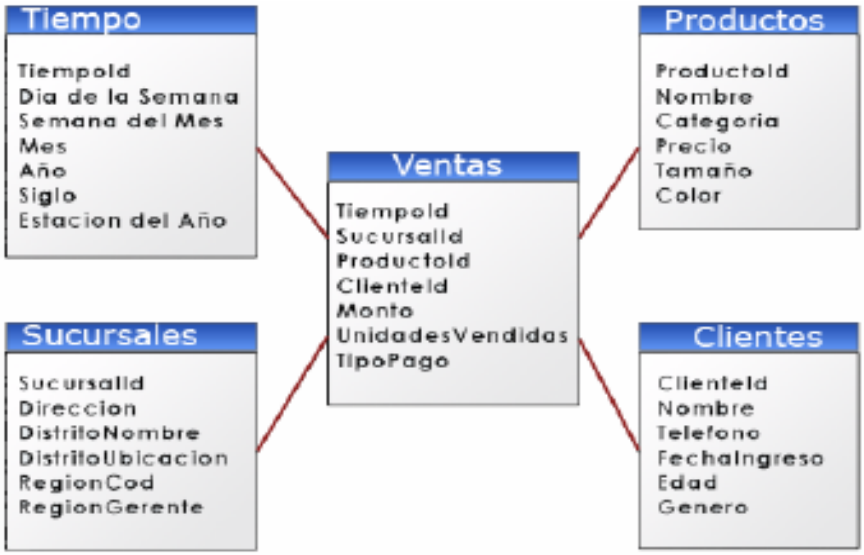


Figura 2.1. Esquema Estrella.

Esquema Copo de Nieve.

Es una variante del esquema estrella en el cual las tablas de dimensión están normalizadas, es decir, pueden incluir claves que apuntan a otras tablas de dimensión.

Las ventajas de esta normalización son la reducción del tamaño y redundancia en las tablas de dimensión, y un aumento de flexibilidad en la definición de dimensiones.

Sin embargo, el incremento en la cantidad de tablas hace que se necesiten más operaciones de unión para responder a las consultas, lo que empeora la performance, además del mantenimiento que requieren las tablas adicionales.

En la siguiente figura vemos un esquema similar al anterior, donde la tabla de dimensión Sucursal se expande en las tablas Distrito y Región. Ahora la tabla Sucursal contiene una columna clave DistritoId que apunta a la tabla Distrito, y esta a su vez tiene una columna RegionId que apunta a la tabla de dimensión Región. [PATRICIA ZVENGER]

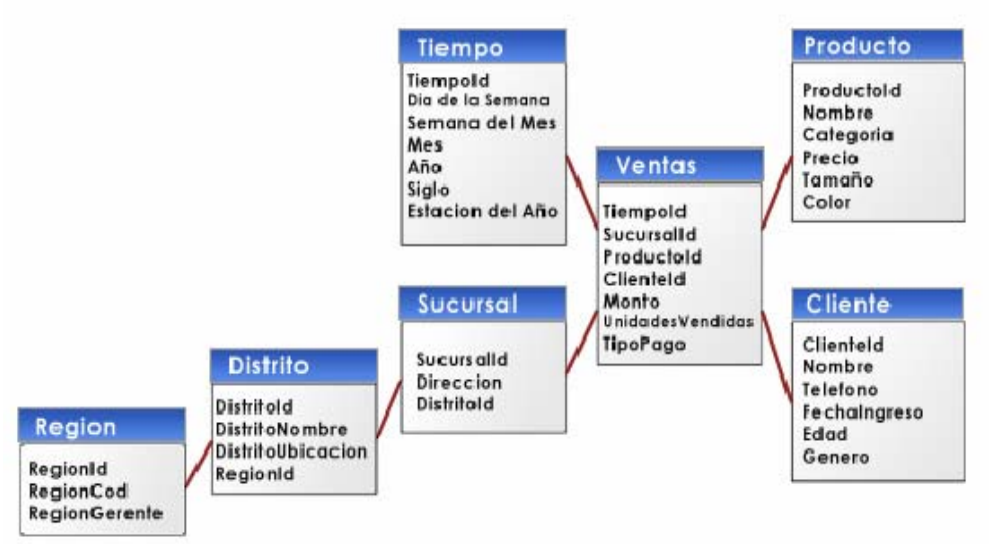


Figura 2.2. Esquema Copo de Nieve

- **Diseño Físico**.

Entre las decisiones de implementación que se deben tomar se incluyen el tamaño del espacio libre, el tamaño del buffer, el tamaño del bloque, y si se usa o no una técnica de compactación de la base de datos. Todas estas cuestiones afectarán la performance del DataWarehouse.

Algunos temas que impactan sobre el rendimiento del Datawarehouse son:

- Particionamiento.

  Generalmente cuando se hablan de base de datos enormes, donde las tablas de hechos ocupan varios cientos de gigabytes. El particionamiento permite que los datos de una tabla lógica, esté en varios datos físicos.

  El particionamiento es importante, pues permite realizar respaldos de porciones de una tabla, sin impactar en su accesibilidad. Por otro lado, permite guardar información mas frecuentemente accedidos, en dispositivos más rápidos. [PATRICIA ZVENGER]

- Clustering.

  Es una técnica útil, para el acceso secuencial de grandes cantidades de datos. Se obtiene definiendo un *índice de clustering* para una tabla, el cual determina el orden secuencial físico en el que se almacenan las filas en los conjuntos de datos.

  Esta técnica mejora drásticamente el acceso secuencial, y es la técnica mas usada para

procesamiento OLAP. Cuando las filas de la tabla no permanezcan almacenadas en el orden correspondiente a su *índice clustering*, situación conocida como fragmentación, la performance bajará y habrá que reorganizar la tabla. [PATRICIA ZVENGER]

- Indexado.

  Existen dos estrategias extremas de indexado: una es indexar todo, y la otra es no indexar nada, pero ninguna de las dos es conveniente. Las columnas que se elijan para indexar deben ser las que se usan más frecuentemente para recuperar las filas, y las que tienen una alta distribución de valores, no una baja como por ejemplo Código Postal.

  Una vez que se determinan las columnas a indexar, hay que determinar la estrategia de índice. La mayoría de las DBMSs proveen varios algoritmos, entre ellos B-tree, Hash, archivo Invertido, Sparse y Binario. Se debería optar por el más óptimo para el producto DBMSs que se está usando. [PATRICIA ZVENGER]

- Reorganizaciones.

  Las cargas incrementales de las bases de datos irán fragmentando las tablas, y esta fragmentación puede resultar en un decaimiento de la performance. La mayoría de las DBMSs proveen rutinas de reorganización para reclamar el espacio fragmentado y mover registros.

Las actividades básicas involucradas en la reorganización de una base de datos implican copiar la base de datos vieja en otro dispositivo, rebloquear las filas y recargarlas. Estas tareas no son triviales en un DataWarehouse, pero todos los DBMSs permiten reorganizar particiones, lo cual es otra buena razón para particionar las tablas. [PATRICIA ZVENGER]

- Backup y Recupero.

  Los DBMSs proveen utilidades para hacer backups completos y también incrementales. Muchas organizaciones tienen la errónea impresión de que los DataWarehouses siempre se pueden recrear a partir de las fuentes de datos originales. Sin embargo, además de que esta tarea puede llevar mucho tiempo porque hay que reejecutar los programas de extracción, transformación y carga, es posible que estos programas y los datos mismos ya no estén disponibles. [PATRICIA ZVENGER]

- Ejecución de las consultas en paralelo.

  Para mejorar la performance de una consulta es mejor dividirla en componentes que ejecuten concurrentemente. Algunos DBMSs ofrecen ejecución paralela en forma transparente, es decir, dividen la consulta por si solos. [PATRICIA ZVENGER]

2.2.2 DATAMART.

Las corporaciones de hoy se esfuerzan por conducir sus negocios hacia una base internacional.

Vemos compañías que surgieron en Estados Unidos y se expandieron a Europa, Asia y África. La expansión del negocio crea la necesidad de acceder a datos corporativos que están ubicados en diferentes puntos geográficos. Por ejemplo, un ejecutivo de ventas de una compañía con origen en Brasil que está situado en Chile puede necesitar acceso a la base de datos de la empresa para identificar los clientes potenciales que residen solo en Chile.

Este problema se soluciona creando versiones más pequeñas del DataWarehouse, los datamarts. Estas versiones se crean usando algún criterio particular, como por ejemplo el lugar geográfico. En el ejemplo anterior los datos de los clientes que residen en Chile se deben almacenar en el datamart de la sucursal en ese país.

La existencia de los datamarts crea nuevas formas de pensar cuando se diseñan los repositorios corporativos de datos. Algunas corporaciones reemplazan completamente el concepto de tener un DataWarehouse central, por varios datamarts más pequeños que se alimenten directamente de los sistemas operacionales.

Otras compañías usan datamarts para complementar sus DataWarehouses. Mueven datos desde el DataWarehouse hacia varios datamarts con el fin de permitir un análisis más eficiente. La separación de los datos se determina según

criterios como departamentos, áreas geográficas, periodos de tiempo, etc.

Finalmente, algunas organizaciones usan sus datamarts como el primer paso de almacenamiento de datos operacionales. Luego los datos de todos los datamarts se replican en un DataWarehouse corporativo central. [PATRICIA ZVENGER].

2.2.3 Almacenamiento OLAP.

OLAP se define como el análisis multidimensional e interactivo de la información de negocios a escala empresarial. El análisis multidimensional consiste en combinar distintas áreas de la organización, y así ubicar ciertos tipos de información que revelen el comportamiento del negocio. [PATRICIA ZVENGER]

Los usuarios de herramientas OLAP se mueven desde una perspectiva de negocio a otra, por ejemplo, pueden estar observando las ventas anuales por sucursal y pasar a ver las sucursales con más ganancias en los últimos tres meses, y además con la posibilidad de elegir entre diferentes niveles de detalle, como ventas por día, por semana o por cuatrimestre. Es esta exploración interactiva lo que distingue a OLAP de las herramientas simples de consulta y reportes. [PATRICIA ZVENGER]

El análisis multidimensional, permite a los analistas de negocios examinar sus indicadores clave o medidas, como ventas, costos, y ganancias, desde distintas perspectivas, como periodos de tiempo, productos, regiones. Estas perspectivas constituyen las dimensiones desde las que se explora la información.

La escala empresarial, se refiere a que OLAP trabaja con fuentes de datos corporativos, que contienen datos de toda la empresa.

Para proveer estas características, toda herramienta OLAP tiene tres principales características:

• Un modelo multidimensional de la información para el análisis interactivo.
• Un motor OLAP que procesa las consultas multidimensionales sobre los datos.
• Un mecanismo de almacenamiento para guardar los datos que se van a analizar. Este componente puede ser externo a la herramienta, como un RDBMS o un DataWarehouse.

La herramienta no solo permite flexibilidad en cuanto a la navegación por el modelo multidimensional de la información, sino que también es flexible en la definición de los reportes y aplicaciones que se construyen a partir de ella. [PATRICIA ZVENGER]

CUBOS MULTIDIMENSIONALES

En una base de datos multidimensional, el modelo de datos esta constituido por lo que se denomina un Cubo multidimensional o simplemente Cubo. En un cubo la información se representa por medio de matrices multidimensionales o cuadros de múltiples entradas, que nos permite realizar distintas combinaciones de sus elementos para visualizar los resultados desde distintas perspectivas y variando los niveles de detalle. Esta estructura es independiente del sistema transaccional de la organización, facilita y agiliza la consulta de información histórica ofreciendo la posibilidad de navegar y analizar los datos.

Aquí vemos como ejemplo un cubo multidimensional que contiene información de ventas discriminadas por periodos de tiempo, productos y zonas geográficas de la empresa.

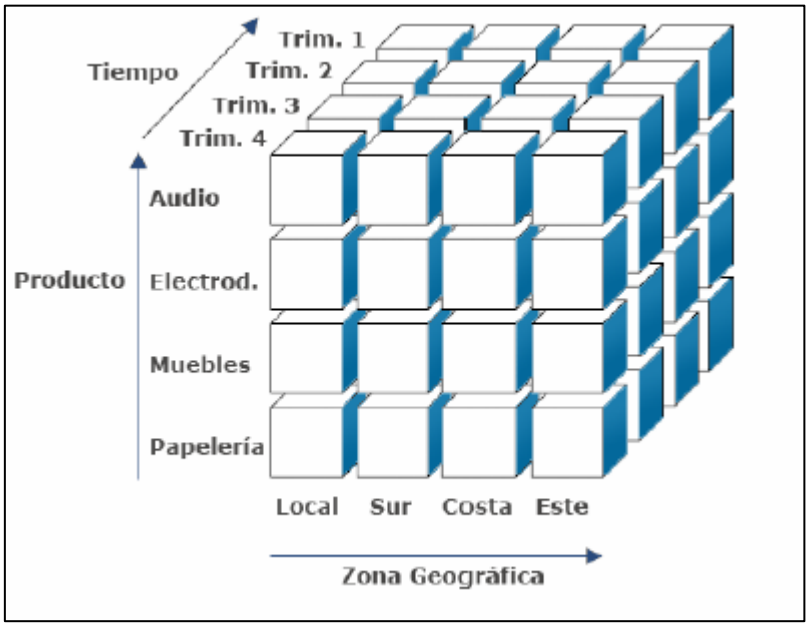


Figura 2.3. Cubo Multidimensional

Los ejes del cubo son las Dimensiones, y los valores que se presentan en la matriz, son las Medidas. [PATRICIA ZVENGER]

DIMENSIONES

Son objetos del negocio con los cuales se puede analizar la tendencia y el comportamiento del mismo. Las definiciones de las dimensiones se basan en políticas de la compañía o del mercado, e indican la manera en que la organización interpreta o clasifica su información para segmentar el análisis en sectores, facilitando la observación de los datos.

Para determinar las dimensiones requeridas para analizar los datos podemos hacer preguntas como: Cuándo, Dónde, Qué, Quién, Cuál, etc. [PATRICIA ZVENGER]

MEDIDAS O METRICAS

Son características cualitativas o cuantitativas de los objetos que se desean analizar en las empresas. Las medidas cuantitativas están dadas por valores o cifras porcentuales.
Por ejemplo, las ventas en dólares, cantidad de unidades en stock, cantidad de unidades de producto vendidas, horas trabajadas, el promedio de piezas producidas, el porcentaje de aceptación de un producto, el consumo de combustible de un vehículo, etc. [PATRICIA ZVENGER]

JERARQUIAS DE DIMENSIONES Y NIVELES

Generalmente las dimensiones se estructuran en jerarquías, y en cada jerarquía existen uno o mas niveles, los llamados Niveles de Agregación o simplemente Niveles. Toda dimensión tiene por lo menos una jerarquía con un único nivel. En la figura vemos un ejemplo de una dimensión de vendedores, que consiste de una única jerarquía, y tres niveles de agregación para agruparlos por ciudades y por regiones.

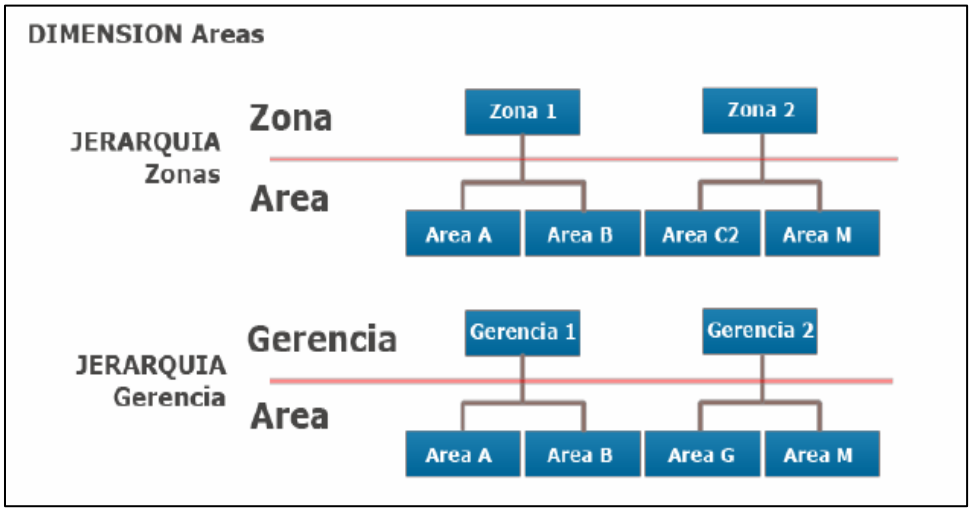


Figura 2.4. Dimensiones y Jerarquías.

En el grafico anterior, los niveles de Zonas y Gerencia no están relacionados entre si, a pesar de que ambos están relacionados con las Áreas. [PATRICIA ZVENGER]

2.2.4 ESTRATEGIAS DE ALMACENAMIENTO. (ROLAP, MOLAP, HOLAP)

Las bases de datos relacionales están optimizadas para obtener una performance óptima en consultas simples y frecuentes, pero no funcionan de manera ideal para las consultas multidimensionales y complejas de estas aplicaciones, ya que existen muchas de ellas que no se pueden expresar en una única consulta SQL, y seguramente se requerirán muchas operaciones de JOIN, lo cual reduce drásticamente el tiempo de respuesta de la consulta.

Para cubrir estas deficiencias surgieron tres estrategias de almacenamiento:

• Bases de datos multidimensionales especializadas, que proveen almacenamiento y recupero de datos optimizado para consultas OLAP.

• DataWarehouses, construidos sobre una tecnología relacional, pero la optimización se dirige al soporte de decisiones en lugar de a las operaciones transaccionales.

• Una tercera estrategia que consiste en la combinación de las dos anteriores.

Las herramientas OLAP que usan almacenamiento multidimensional son llamadas MOLAP, mientras que a las que almacenan los datos en bases relacionales se les llama herramientas ROLAP. Las herramientas que combinan los dos enfoques se conocen como OLAP Híbrido u HOLAP.

Cada alternativa tiene sus ventajas y desventajas. En lugar de discutir cual de las dos es mejor hay que definir un criterio para optar por una u otra, y evaluar el alcance de HOLAP, que en la práctica intenta combinar lo mejor de ambos mundos.

Algunas de las ventajas más importantes de cada enfoque son:

MOLAP

• Buena performance en las consultas, ya que el almacenamiento esta optimizado para el análisis multidimensional.

• La escalabilidad está limitada por la capacidad del Motor de Base de Datos y por el tiempo de carga de los datos.

• En general el análisis está limitado a los datos totalizados o sumarizados.

• El modelo multidimensional no es lo suficientemente flexible como para acomodarse a las necesidades constantemente cambiantes del negocio.

• La estructura que guarda los datos está incluida en la herramienta.

• Requiere una capa adicional de manejo de datos.

• No incluye soporte de paralelismo, replicación ni recuperación de datos.

• Puede requerir aprendizaje por ser una tecnología nueva en la organización. [PATRICIA ZVENGER]

ROLAP

• La performance de las consultas no es tan óptima como en MOLAP.

• Es capaz de manejar conjuntos de datos muy grandes, por encima de un terabyte.

• Además del análisis de información sumarizada, se pueden analizar datos detallados hasta el nivel de las transacciones.

• Es capaz de analizar los datos desde cualquier perspectiva en cualquier momento.

• La herramienta ROLAP requiere un DataWarehouse de donde extraer los datos para analizar.

• Las cuestiones técnicas del manejo de los datos está a cargo del Motor de Base de Datos.

• Incluye soporte para replicación, rollback y recuperación, y para acceso multiusuario. [PATRICIA ZVENGER]

2.3 CONCEPTUALIZACIONES SOBRE TRANSFORMACION Y CARGA DE DATOS.

2.3.1 MIGRACION DE DATOS: EXTRACCION, TRANSFORMACION Y CARGA

La migración de los datos desde las fuentes operacionales al DataWarehouse requiere la necesidad de procesos para extraer, transformar y cargar los datos, actividad que se conoce como ETL.

La mayoría de los datos de origen son los datos operacionales actuales, aunque parte de ellos pueden ser datos históricos archivados.

Si los requerimientos de datos incluyen algunos años de historia es necesario desarrollar tres conjuntos de programas ETL: una Carga Inicial, una Carga Histórica, y una Carga Incremental.

Carga Inicial

La carga inicial se asemeja mucho al proceso de conversión entre sistemas que se da en las organizaciones cuando pasan, por ejemplo, de sus viejos sistemas operacionales a un producto ERP.

Carga Histórica

Este proceso debe verse como una extensión de la carga inicial, pero la conversión aquí es un poco diferente porque los datos históricos son datos estáticos.

A diferencia de los datos operacionales, los datos estáticos ya se archivaron en dispositivos de almacenamiento offline. Es

común que con el transcurso del tiempo se eliminen elementos de datos que ya no sirven, se agreguen nuevos, se modifiquen los tipos de ciertos datos o los formatos de los registros, lo que implica que los datos históricos no necesariamente se puedan sincronizar con los datos operacionales. Por lo tanto los programas de conversión escritos para la carga inicial quizá no sean aplicables a la carga de datos históricos sin algunos cambios previos.

Carga Incremental

Una vez que el DataWarehouse está cargado con datos iniciales e históricos, hay que desarrollar otro proceso para la carga incremental, que se ejecutará mensual, semanal o diariamente. Existen dos formas de diseñar la carga incremental:

• *Extraer todos los registros*: Se extraen todos los registros operacionales, independientemente de los valores que hayan cambiado desde la última carga realizada.

En general esta opción no es viable debido al volumen de los datos, por eso la mayoría opta por la siguiente opción.

• *Extraer Deltas solamente*: Solo se extraen registros nuevos o registros que contengan valores que cambiaron desde la última carga realizada.

Diseñar programas ETL para extracciones delta es más fácil cuando las fuentes consisten en bases de datos relacionales y contamos con una columna “timestamp” para determinar los deltas. [PATRICIA ZVENGER]

Expliquemos ahora, lo que debe contemplar este proceso:

A. Extraer los Datos.

   Que consiste en determinar técnicas, para combinar eficiencia en el uso de la data de origen, así como detectar redundancias y datos y algún otro ruido. Además, hay que distinguir un dato que puede estar duplicado en distintas tablas.

B. Transformar Datos.

   Este proceso es el más crítico, debido a que debe controlar algunos factores:

   Claves primarias inconsistentes, valores inconsistentes, datos con diferentes formatos, valores erróneos, sinónimos y homónimos, Lógica embebida, Integración y Derivación, etc. descritos en [PATRICIA ZVENGER]

   En la figura vemos algunos ejemplos de transformación de datos: El primero referente a sexo, el segundo referente a unidades de medida, el tercero se refiere a estandarizar nombres, y por último, estandarizar formatos de fecha.

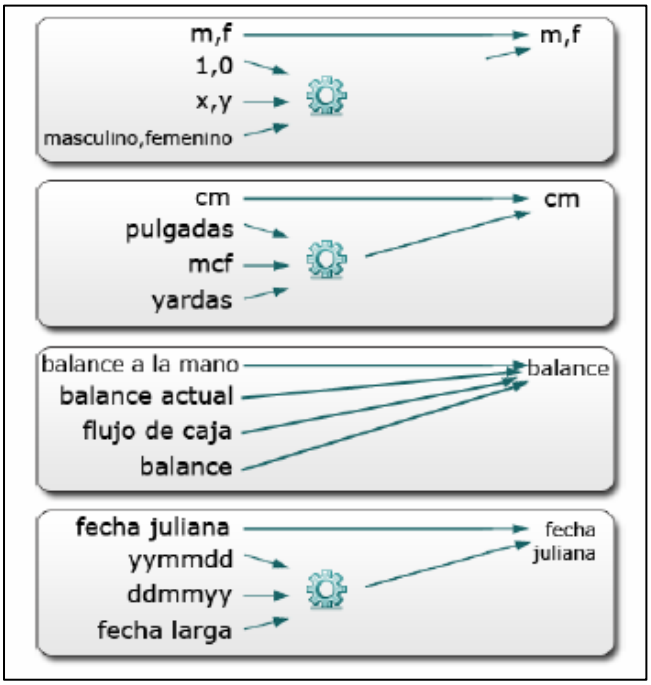


Figura 2.5. Ejemplos de Transformación

C. Cargar Datos.

Este paso, es el más simple, y seria el que completaría el proceso ETL. Aquí se tendría que tener cuidado, básicamente con los índices, y a la integridad referencial.

2.4 CONCEPTOS SOBRE MINERIA DE DATOS.

2.4.1 DATA MINING.

Data Mining, la extracción de información oculta y predecible de grandes bases de datos, es una tecnología para ayudar a las compañías a descubrir información relevante en sus bases de información. Las herramientas de Data Mining clasifican y predicen futuras tendencias y comportamientos. Los análisis

prospectivos automatizados ofrecidos por la automatización del Data Mining van más allá de los eventos pasados provistos por las herramientas usuales de sistemas de soporte de decisión.

Estas herramientas exploran las bases de datos en busca de patrones ocultos, encontrando información predecible que un experto no puede llegar a encontrar.

Muchas compañías ya colectan y refinan cantidades masivas de datos. Las técnicas de Data Mining pueden ser implementadas rápidamente en plataformas ya existentes de software y hardware para acrecentar el valor de las fuentes de información existentes y pueden ser integradas con nuevos productos y sistemas.

Los algoritmos de Data Mining utilizan técnicas que han existido por lo menos desde hace 10 años, pero que sólo han sido implementadas recientemente como herramientas maduras y confiables.

[MAGDALENA SERVENTE] otorga ciertas capacidades a la tecnología de Data Mining:

- **Descripción de clases**: Provee una clasificación (caracterización) concisa y resumida de un

conjunto de datos y los distingue (discriminación) unos de otros.

- Asociación: Es el descubrimiento de relaciones de asociación o correlación en un conjunto de datos.

- Clasificación: Analiza un conjunto de datos de entrenamiento cuya clasificación de clase se conoce y construye un modelo de objetos para cada clase. Puede representarse en árboles de decisión o reglas de clasificación.

- **Predicción**: Esta función de la minería predice los valores posibles de datos faltantes o la distribución de valores de ciertos atributos en un conjunto de objetos.

- **Clustering**: Identifica clusters en los datos, donde un cluster es una colección de datos “similares”. La similitud puede medirse mediante funciones de distancia, especificadas por los usuarios o por expertos. La Minería de Datos trata de encontrar clusters de buena calidad que sean escalables a grandes bases de datos y a datawarehouses multidimensionales.

- **Análisis de Series a través de Tiempo**: Analiza un gran conjunto de datos obtenidos con el correr del tiempo para encontrar en él regularidades y características interesantes, incluyendo la búsqueda de patrones secuenciales, periódicos, modas y desviaciones.

2.4.1.1 Algoritmo K-Means (K-Medias)

Uno de los algoritmos más utilizados para hacer clustering es el k-medias (kmeans), que se caracteriza por su sencillez. [MOLINA GARCIA]

1. En primer lugar se debe especificar por adelantado cuantos clusters se van a crear, éste es el parámetro ***k***, para lo cual se seleccionan ***k*** elementos aleatoriamente, que representarán el centro o media de cada cluster.
2. A continuación cada una de las instancias, ejemplos, es asignada al centro del cluster más cercano de acuerdo con la distancia Euclidea que le separa de él.
3. Para cada uno de los clusters así construidos se calcula el centroide de todas sus instancias y estos centroides son tomados como los nuevos centros de sus respectivos clusters.
4. Finalmente se repite el proceso completo con los nuevos centros de los clusters.
5. La iteración continúa hasta que se repite la asignación de los mismos ejemplos a los mismos clusters, ya que los puntos centrales de los clusters se han estabilizado y permanecerán invariables después de cada iteración. [MOLINA GARCIA]

Para obtener los centroides, se calcula la media o la moda según se trate de atributos numéricos o simbólicos.

El algoritmo lo ponemos para mejor entendimiento:

Algoritmo: Clustering k-means

Entrada: Un conjunto de N vectores de datos $X= \{x_1,\ldots,x_N\}$ en R^d y el número de clusters es K.

Salida: Una partición de vectores de datos dada por el vector identidad del cluster $Y= \{y_1,\ldots,y_N\}$, $y_n \in \{1,\ldots K\}$.

Pasos:

1. *Inicialización*: Inicializa los vectores centroides del cluster $\{\mu_1,\ldots,\mu_k\}$;
2. *Asignación de datos*: Para cada vector de datos x_n, el conjunto $y_n=\underset{k}{argmin}||x_n-\mu_k||^2$;
3. *Estimación del centroide*: Para un cluster K, sea $X_k=\{x_n|y_n=K\}$, el centroide se estima como $\mu_k=1/|x_k| \sum_{x\in X_k} x$;
4. Parar si Y no cambia, de otra forma regresar al paso 2.

Explicación:

- Se determinan K centros iniciales
- Se repite:
  - Crear los K grupos en base a los patrones más cercanos a cada centro.
  - Recalcular los K centros como los puntos medios de cada grupo creado.

Mientras los K centros tengan una variación apreciable en posición entre dos iteraciones.

Eventualmente, luego de algunas iteraciones los centros se estabilizan y con ellos la partición del espacio formado por los K grupos definidos por estos centros.

El Diagrama de Flujo del algoritmo podría ser el siguiente: [SANDRA CARTAGENOVA]

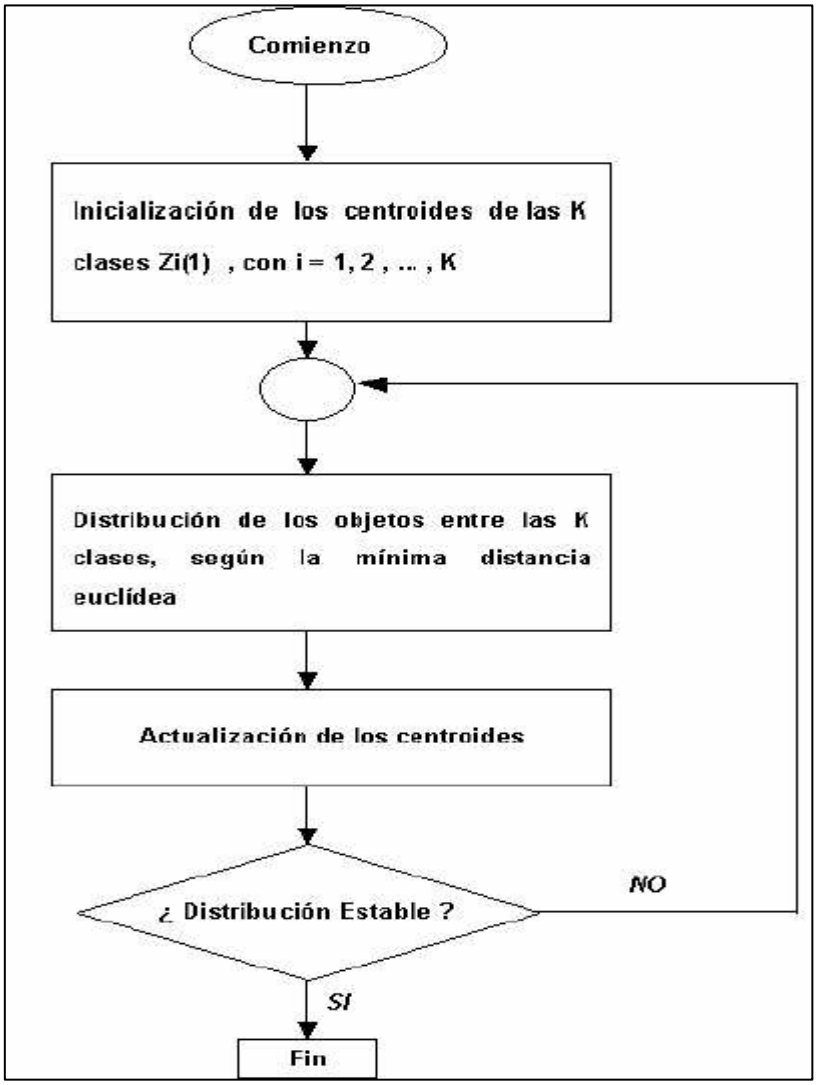


Figura 2.6. Algoritmo K-Means

A continuación, se muestran ejemplos de clustering con el algoritmo k-medias.

EJEMPLO 1: [MOLINA GARCIA]

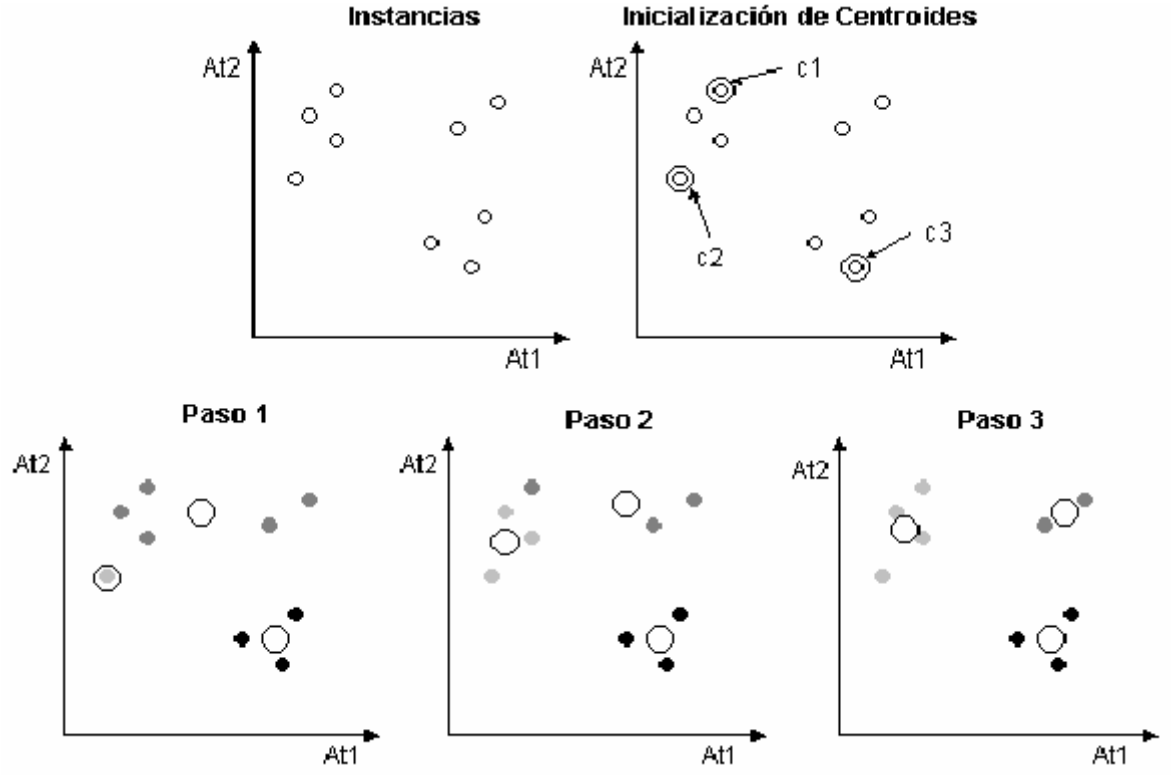


Figura 2.7. Ejemplo K-Means

En este caso se parte de un total de nueve ejemplos o instancias, se configura el algoritmo para que obtenga 3 clusters, y se inicializan aleatoriamente los centroides de los clusters a un ejemplo determinado. Una vez inicializados los datos, se comienza el bucle del algoritmo. En cada una de las gráficas inferiores se muestra un paso por el algoritmo. Cada uno de los ejemplos se representa con un tono de color diferente que indica la pertenencia del ejemplo a un cluster determinado, mientras que los centroides siguen mostrándose como círculos de mayor tamaño y sin

relleno. Por ultimo el proceso de clustering finaliza en el paso 3, ya que en la siguiente pasada del algoritmo (realmente haría cuatro pasadas, si se configurara así) ningún ejemplo cambiaría de cluster.

EJEMPLO 2: [LUIS GABRIEL]

En la primera columna se encuentra la posición del elemento y en la segunda su valor. Se han elegido inicialmente 2 centroides, ubicados en las posiciones 2 y 7. En la columna con etiqueta *dist 1* se ha registrado la distancia de cada objeto al primer centroide. De igual forma, en la siguiente columna se ha registrado la distancia de cada objeto al siguiente centroide. Luego se han escogido las distancias mínimas, y en la última columna de la tabla se realiza la asignación de elementos a cada uno de los grupos.

Número	*objeto*	*dist 1*	*dist 2*	*mínima d.*	*cluster*
1	9	1	2	1	1
2	***10***	0	3	0	1
3	4	6	3	3	2
4	5	5	2	2	2
5	9	1	2	1	1
6	3	7	4	4	2
7	**7**	3	0	0	2
8	*25*	15	18	15	1
9	8	2	1	1	2
10	0	10	7	7	2

Tabla 2.1. Ejemplo K-Means

Se recalculan los centros, como el promedio de las distancias dentro de cada conglomerado. Los nuevos centroides son: 4.25, 2.83.

Número	*objeto*	*dist 1*	*dist 2*	*mínima d.*	*cluster*
1	9	4.75	6.17	4.75	1
2	***10***	5.75	7.17	5.75	1
3	4	0.25	1.17	0.25	1
4	5	0.75	2.17	0.75	1
5	9	4.75	6.17	4.75	1
6	3	1.25	0.17	0.17	2
7	**7**	2.75	4.17	2.75	1
8	*25*	20.75	22.17	20.75	1
9	8	3.75	5.17	3.75	1
10	0	4.25	2.83	2.83	2

Tabla 2.2. Ejemplo K-Means

Ahora, en esta tabla, se calcula la distancia de cada elemento a los nuevos centros. Este proceso se repite iterativamente hasta un número de veces propuesto por el usuario o hasta que no varié la configuración dentro de los grupos.

CAPÍTULO III
MODELADO Y CONSTRUCCIÓN DE UN DATAMART COMO FUENTE DE INFORMACIÓN PARA UN ALGORITMO DE MINERÍA DE DATOS EN UN HOSPITAL NACIONAL

3.1 VISIÓN GENERAL DEL PLAN DE PROYECTO.

La mayoría de las organizaciones cuentan con datos de los sistemas de ingreso de transacciones, vitales para registrar las operaciones que sostienen a una empresa. A pesar de la riqueza de estos datos, no se puede recurrir a ellos con facilidad cuando necesitamos encontrar respuestas sobre el funcionamiento de la organización, como por ejemplo, en el caso del hospital:

- ¿Qué pacientes y con qué frecuencia acuden a atenderse? Con esta información se puede evaluar, con más certeza, si se está atendiendo a pacientes dudosamente recurrentes, o si verdaderamente se requiere tomar medidas para disminuir el número de atenciones.

- ¿Cuáles serían las diferentes alternativas de cambio en el consumo de medicamentos si se modifica una determinada variable? Ejemplo de esto podría ser:

  - ¿Qué pasaría si se cambia el horario de atención de las farmacias?
  - ¿Cuál puede ser la cantidad óptima de abastecimiento de medicamento, dado el presupuesto?
  - ¿En cuánto es posible mejorar la atención a largo plazo de pacientes identificables?

- ¿Cuál es la proyección histórica de determinado grupo de pacientes de acuerdo a sus diagnósticos previos? así, puede utilizarse esta información para ampliar o reducir la cantidad de insumos a esos grupos.

- ¿Cuáles son las causas de que los pacientes acudan tantas veces? así, se podrían tomar acciones proactivas tanto por grupo de pacientes como por tipo de enfermedades.

- ¿Cuánto se sobrepasó el último trimestre el consumo de algún medicamento? Es posible implementar una política de uso diferente para el trimestre siguiente para evitar que los pacientes carezcan de determinados medicamentos.

Si bien con los sistemas tradicionales se pueden preparar reportes ad-hoc para encontrar las respuestas a algunas de estas preguntas, se necesita mucho tiempo y recursos del departamento de sistemas para poder responderlas.

Además interfiere en el procesamiento de los sistemas transaccionales, aumentando los tiempos de respuestas de los mismos.

Los usuarios del sistema de donde se extrajo las fuentes de datos no son usuarios comunes, sino usuarios que toman

decisiones y planifican día a día, a mediano plazo o a largo plazo.

Tal como definimos en el capítulo I, nuestra población se compone de personal con algunos de los siguientes cargos:

- Gerente General
- Auditor del Área Medica.
- Médicos pertenecientes a Comisiones definidas.
- Operadores del Abastecimiento de medicamentos.

3.2 PROPUESTA METODOLÓGICA.

A continuación, se detallan los pasos que se siguieron en esta metodología para identificar sectores o grupos de pacientes con características similares en el consumo de medicamentos:

1. **Entender el problema existente en la información transaccional, analizándola y seleccionando los campos pertinentes de las tablas seleccionadas.**
2. **Limpiar los datos de la muestra seleccionada.**
3. **Diseñar el Esquema Dimensional del Datamart.**
4. **Llevar la muestra hacia un modelo dimensional.**
5. **Selección de Atributos para el análisis del algoritmo.**
6. **Aplicar el algoritmo k-means para el proceso de clasificación e interpretación.**

Los pasos anteriores, resumen los pasos que se realizarán en la investigación. El datamart a desarrollar se alimenta de los datos residentes en una base de datos relacional a desarrollar, cargada desde los archivos planos de los sistemas transaccionales existentes del hospital, por medio de diversos procesos de extracción, transformación y carga. Este nuevo repositorio, diseñado para brindar información dinámica, servirá como fuente de datos para la aplicación de un algoritmo de minería de datos de clasificación.

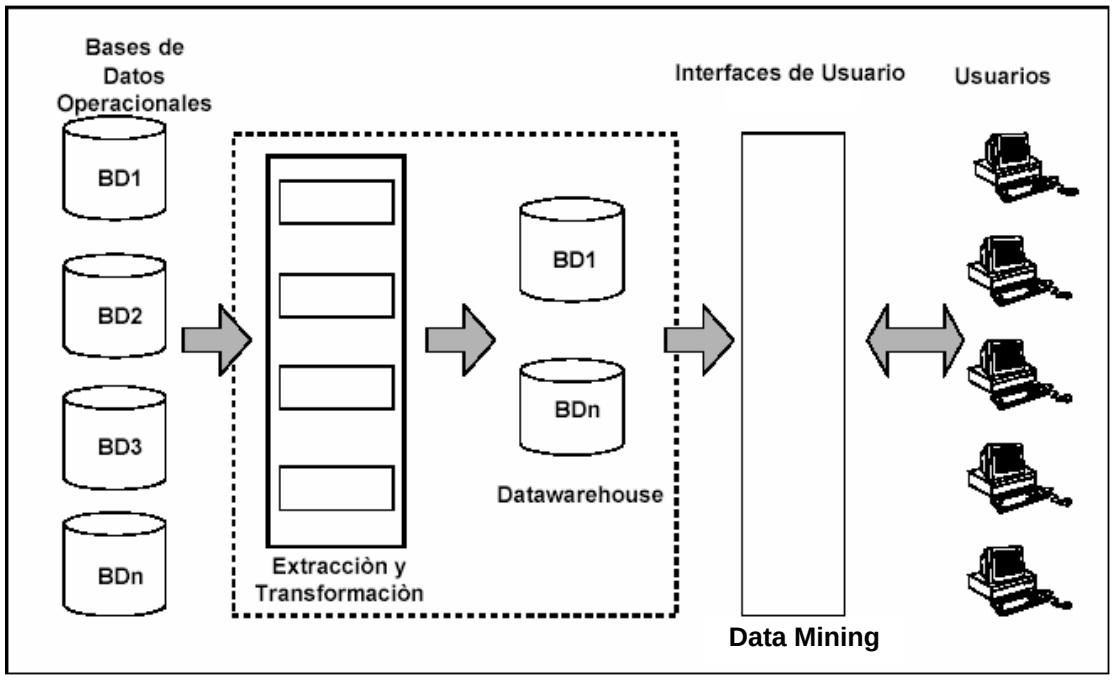


Figura 3.1. Pasos de la Metodología.

3.2.1 ENTENDER EL PROBLEMA EXISTENTE EN LA INFORMACIÓN TRANSACCIONAL, ANALIZARLA Y SELECCIONAR LOS CAMPOS NECESARIOS DE LAS TABLAS SELECCIONADAS

En esta sección, se debe ahondar en la información contenida en el sistema transaccional, empezando por limitar las consultas a las tablas relacionadas con las transacciones realizadas en el área de farmacia del hospital.

Las tablas seleccionadas tras la exploración de los datos son:

ADSERVIC: Contiene todas los servicios con los que cuenta el hospital.

ADTASEG: Contiene Todos los tipos de seguros que existen, y por los cuales los pacientes son admitidos en el Centro hospitalario (Obligatorios, Dependientes, Convivientes, Hijos, etc.)

DIAGNOS: Contiene información de los diagnósticos codificados y extraídos desde un documento estándar de aplicación general en toda la red asistencial del centro e estudio.

FM_PRECE: Contiene las presentaciones de los medicamentos (ampollas, jarabes, etc.)

MEDICAME: Contiene todos los medicamentos registrados en el hospital. Los mismos que son adquiridos por el área de abastecimiento del hospital, y quienes se encargan de distribuir los medicamentos a todas las farmacias.

MEDICO: Contiene información de los médicos registrados en el hospital.

ADHISCLI: Contiene información de cada paciente registrado en el hospital. Aquí se registra tanto información propia del sistema de farmacia de consulta externa, como de otras áreas.

POLICLIN: Contiene información de los policlínicos pertenecientes a la red EsSalud. Se registran policlínicos debido a que al hospital de estudio, llegan pacientes de distintos policlínicos.

TIEMPO: Contiene todas las fechas existentes en las informaciones transaccionales. De aquí se desprenderán las jerarquías de semanas, meses, trimestres, etc. que serán usados en el modelo multidimensional.

RECETA: Contiene las transacciones diarias correspondientes al área de farmacia de consulta externa. Se registran datos como *hora y fecha de registro*, código del *medicamento* solicitado, *medico* tratante, código del *paciente*, código del *servicio*, *cantidad* de medicamento solicitado (en unidades y soles).

Esta tabla contiene información correspondiente a las fechas de *2 de enero del 2004, hasta 11 de Agosto del 2004*.

La información seleccionada existe en archivos planos y forman parte de las tablas que serán modeladas en próximos pasos.

3.2.2 LIMPIAR LOS DATOS DE LA MUESTRA SELECCIONADA

Este paso es la etapa que tomó más tiempo, debido a que generalmente las transacciones realizadas en el hospital

de estudio, son hechas sin un control exhaustivo, más que el que brinda el propio usuario del sistema.

Es por esto, que hay que verificar en la mayoría de los datos, y ejecutando algunas consultas simples de Transact-SQL, si la información existente es valida o tiene, en algunos casos, datos inconsistentes con el tipo de dato almacenado.

Ejemplos de esto es la validación de fechas, validación de saldos (en donde el saldo no puede ser negativo), campos en blanco, campos de error (caracteres inválidos), etc.

3.2.3 DISEÑAR EL ESQUEMA DIMENSIONAL DEL DATAMART

Siguiendo las recomendaciones de [ANDRE-FELIPE] debemos diseñar un Esquema Físico o un proyecto físico de datos, el cual debe ser el más simple o próximo posible al esquema lógico.

- Definición de Dimensiones y Jerarquías.
  DIAGNOSTICO. Jerarquía: Ce_Cdiag
  MEDICAMENTO. Jerarquía: Md_Contro, Md_Codlog
  MEDICO. Jerarquía: Me_Dmed
  PERSONA. Jerarquía: Hi_Sexo, Hi_Estciv
  POLICLINICO. Jerarquía: Po_Npolic
  PRESENTACION. Jerarquía: Pr_Descrip
  SERVICIO. Jerarquía: Se_Cser
  TIEMPO. Jerarquía: Año, Trimestre, Mes, Dia
  TIPOSEGURO. Jerarquía: Ta_Dtase
- Definición de Tabla de Hechos.
  FACT_HOSPITAL.

- Creación de Claves

  Debemos ahora asignar Llaves primarias de cada dimensión:

DIMENSION	NOMBRE CLAVE	TIPO DATO
DIAGNOSTICO	codigo	entero
MEDICO	codigo	entero
PERSONA	codigo	entero
POLICLINICO	codigo	entero
PRESENTACION	pr_codigo1	entero
SERVICIO	codigo	entero
TIEMPO	codigo	entero
TIPOSEGURO	codigo	entero

- Creación de Índices

  Por cada dimensión se definieron un índice adicional al que contiene producto de sus claves, generalmente se crearon índices para los nombres o descripciones:

  DIAGNOSTICO: IX_Diagnostico.

  MEDICO: IX_Medico.

  PERSONA: IX_Persona.

  POLICLINICO: IX_Policlinico.

  PRESENTACION: IX_Presentacion.

  SERVICIO: IX_Servicio.

  TIEMPO: IX_Tiempo.

  TIPOSEGURO: IX_TipoSeguro.

FACT_HOSPITAL:
IX_fact_hospital,IX_fact_hospital_1,
IX_fact_hospital_2,IX_fact_hospital_3,
IX_fact_hospital_4,IX_fact_hospital_5,
IX_fact_hospital_6,IX_fact_hospital_7,
IX_fact_hospital_8.

3.2.4 LLEVAR LA MUESTRA HACIA UN MODELO DIMENSIONAL

En esta etapa, se necesita de herramientas de extracción, transformación y carga de datos, para poder pasar la información contenida en los repositorios transaccionales, hacia una base de datos dimensional.

En este caso, lo que se ha hecho primeramente, es:

Utilización tanto de sentencias “SQL-Transact”, así como sentencias FoxPro para generar las “dimensiones” y la “tabla de hechos” en un ambiente aun de Entidad-Relación. Dicha fuente de datos Entidad-Relación será la fuente de datos para el almacenamiento dimensional. La sintaxis utilizada es propia de FoxPro. El código utilizado para obtener estas “tablas-dimensiones” así como la “tabla de hechos” es el siguiente:

```
&& PROGRAMA DE CONVERSION DE TABLAS A DIMENSIONES
ESTANDARIZADAS
&& PARA ARMAR LA FACT_TABLE

SET DELE ON
SET DATE BRITI
SET CENT ON

Sele 1
Use fm_r0200_2004 inde fm_r0200_2004 Alias Rec  &&CLAVE: RP_CODLOG
Sele 2
USE Adservic inde Adservic Alias SER            &&CLAVE: SE_CSER
Sele 4
USE Adtaseg inde Adtaseg Alias SEG              &&CLAVE: TA_CTASE
Sele 5
USE Diagnos inde Diagnos Alias DIA              &&CLAVE: CE_CDIAG
Sele 6
USE Fm_prese inde Fm_prese Alias PRE            &&CLAVE: PR_CODIGO2
Sele 7
USE Medicame inde Medicame Alias MED            &&CLAVE: MD_CODLOG
Sele 8
USE Medico inde medico Alias MEDI               &&CLAVE: ME_NCMP
SELE 10
USE Policlin inde Policlin ALIAS POL            &&CLAVE: PO_CPOLIC
SELE 13
Use Adhiscli inde Adhiscli                      &&CLAVE: HI_AUTASE
sele 14
Use Tiempo inde Tiempo                          &&CLAVE: FECHA
SELE 11
USE FACT_HOSPITAL ALIAS FACT

SELE 1

GO TOP
RegProcesados=0

DO WHILE !EOF()
   if (rp_codlog!='' or rp_case!='' or rp_cmed!='' or rp_cdiag!='' or rp_fdespa={  /  /    })
    SELE 1
    SKIP
   endif
   RegProcesados=RegProcesados+1
```

```
          xcodlog = rp_codlog
xautoge = rp_case
xmedico = rp_cmed
xservic = rp_cser
xdiagno = rp_cdiag
xfecdes = rp_fdespa
fact_canti = rp_qmeddes
sele 14
      seek xfecdes
      IF FOUND()
            fact_fecha= codigo
      ENDIF
      sele 13
      seek xautoge
      IF FOUND()
            fact_persona = codigo
            xpolicl = Hi_cpolic
            xtiposeg = HI_TASE
      SELE 4
          SEEK xtiposeg
          IF FOUND()
            fact_Tiposeg = codigo
          ENDIF
            SELE 10
            SEEK xpolicl
            IF FOUND()
                  fact_policl = codigo
            ENDIF
      ELSE
            sele 1
            skip
            loop
      ENDIF
Sele 2
Seek xservic
IF FOUND()
   fact_servic = codigo
ENDIF
Sele 7
Seek xcodlog
IF FOUND()
   fact_medicam = codigo
   Precio = Md_Precom
   xtipomedi = ALLTR(md_presen)
ENDIF
```

```
    SELE 6
    SEEK xtipomedi
    IF FOUND()
      fact_Tipome = PR_Codigo1
    ENDIF
    Sele 8
    Seek xmedico
    IF FOUND()
      fact_medico = codigo
    ENDIF
    Sele 5
    Seek xdiagno
    IF FOUND()
      fact_diagno = codigo
    ENDIF
    wait wind nowai "Registros Procesados :" + str( RegProcesados ,6)
    SELE 11
    APPE BLANK
    REPL KEY_PERSON  WITH fact_persona
    REPL KEY_MEDICA  WITH fact_medicam
    REPL KEY_MEDICO  WITH fact_medico
    REPL KEY_DIAGNO  WITH fact_diagno
    REPL KEY_PRESEN  WITH fact_Tipome
    REPL KEY_SERVIC  WITH fact_servic
    REPL KEY_TIEMPO  WITH fact_fecha
    REPL KEY_POLICL  WITH fact_policl
    REPL KEY_TIPOSE  WITH fact_Tiposeg
    REPL CONSUMO_UD  WITH fact_canti
    REPL CONSUMO_CO  WITH Precio * fact_canti
    SELE 1
    SKIP
ENDD
```

El procedimiento realiza lo siguiente:

1. Después de limpiar los datos a utilizar, se modifican las estructuras de cada tabla, para añadir un campo, que equivaldrá a las llaves de cada tabla. Es por eso que en cada tabla se añadió la columna “CODIGO” como representación a la Primary Key de cada una de ellas. Cabe destacar que según [PATRICIA ZVENGER], el mejor

tipo de dato para un datamart o datawarehouse es el tipo “entero”, aunque a veces puede ser útil “carácter” de longitud fija. Este proceso ya lo hemos definido en el punto anterior.

2. El bucle nos indica que por cada fila recorrida en la tabla RECETA (que contiene todas las transacciones realizada en el sistema), le corresponde a cada campo, algún código correspondiente a la entidad analizada. Por ejemplo:

   Para el campo RP_CODLOG, existe un CODIGO en la Tabla MEDICINA; para el campo RP_CASE, corresponde un CODIGO en la tabla PERSONA; para el campo RP_CSER corresponde un CODIGO en la tabla SERVICIO, y así sucesivamente.

   Cada CODIGO capturado, se almacena en variables locales, para después insertar dichas variables en cada campo de la nueva tabla generada llamada **FACT_CONSUMO**, armando así una estructura parecida a una tabla de hechos de diseño estrella, pero a modo de archivos planos aun.

3. Al final del código generado, se copia también la **cantidad** de medicamentos solicitados en cada transacción, así como el **costo** de este requerimiento. Los campos **CONSUMO_UD** y **CONSUMO_CO** son las cantidades y los costos de los medicamentos.

Una vez realizado este paso, lo que se utiliza ahora, es alguna herramienta de extracción y carga de datos, desde las

tablas planas generadas a un ambiente de Entidad-Relación, es decir, a un motor de base de datos.

La misma es la que se muestra a continuación:

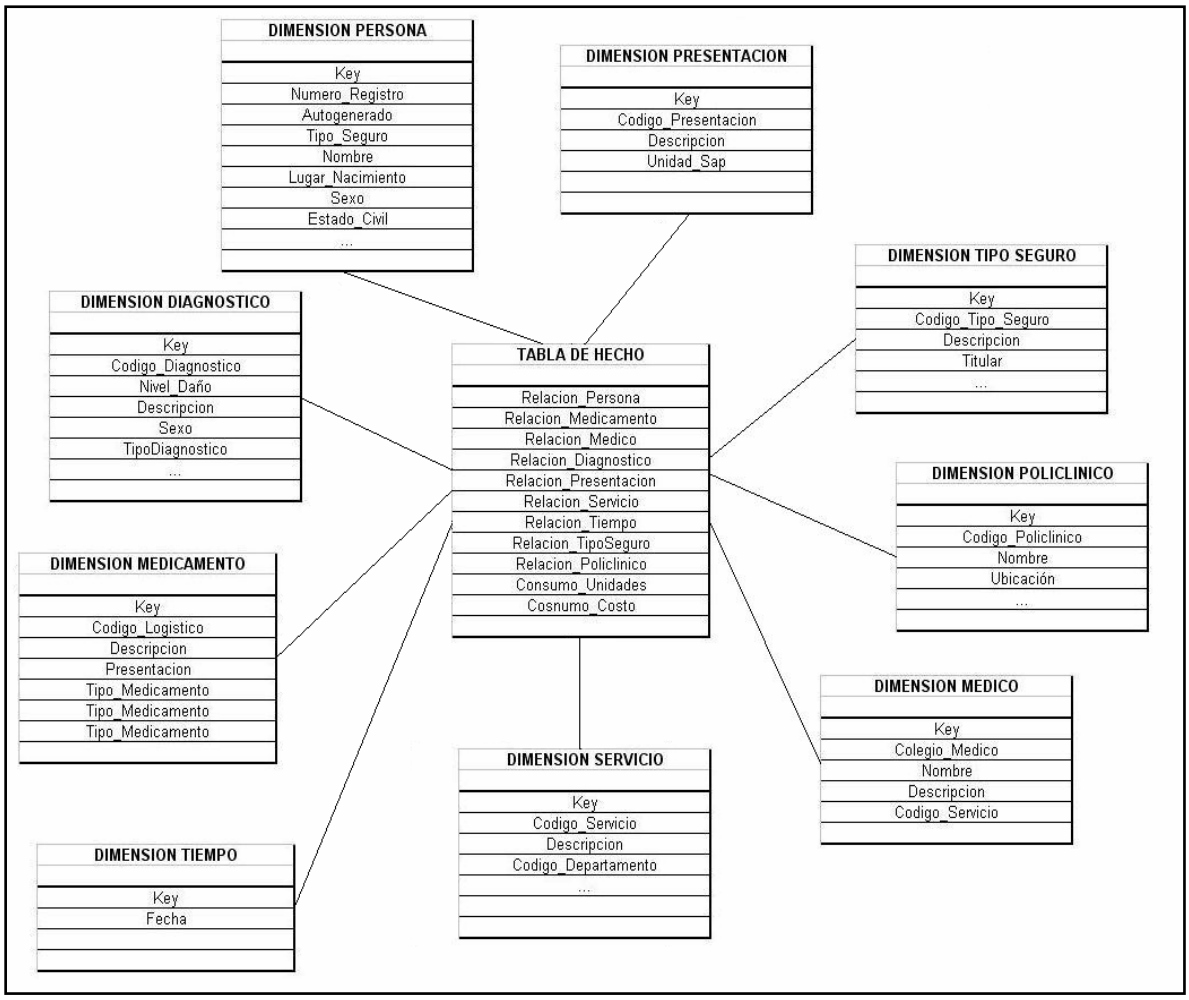


Figura 3.2. Esquema Dimensional del Datamart.

Ahora, se cuenta con las tablas necesarias para migrar la información a un modelo dimensional, donde cada tabla generada con el procedimiento anterior corresponderá a las dimensiones y a la tabla de hechos.

De esta manera se tendrá que:

La tabla MEDICO, se transformará en la **dimensión MEDICO**.

La tabla TIPOSEGURO, se transformará en la **dimensión TIPOSEGURO**.

La tabla POLICLINICO, se transformará en la **dimensión POLICLINICO**.

La tabla PRESENTACION, se transformará en la **dimensión PRESENTACION**.

La tabla DIAGNOSTICO, se transformará en la **dimensión DIAGNOSTICO**.

La tabla MECICAMENTO, se transformará en la **dimensión MEDICAMENTO**.

La tabla SERVICIO, se transformará en la **dimensión SERVICIO**.

La tabla PERSONA, se transformará en la **dimensión PERSONA**.

La tabla TIEMPO, se transformará en la dimensión **TIEMPO**.

Sobre este punto, cabe resaltar que se pueden utilizar funciones propias de los manejadores de base de datos para poder tener todas las jerarquías necesarias en la dimensión Tiempo. A continuación un ejemplo:

```
USE FACT_ENTIDAD_RELACION
GO
SELECT DISTINCT IDENTITY(integer,0, 1) AS Tiempo_key,
 Dia = S.fecha,
 DiadeSemana = DateName(dw,S.fecha),
 Mes = DatePart(mm,S.fecha),
 Ano = DatePart(yy,S.fecha),
 Trimestre =DatePart(qq,S.fecha),
 DiadeAno = DatePart(dy,S.fecha),
 Feriado = 'N',
 FindeSemana = case DatePart(dw,S.fecha)
                  when (1 ) then 'Y'
                  when (7) then 'Y'
                  else 'N'  end,
 MesdeAno = DateName(month, S.fecha) + '_' + DateName(year,S.fecha),
 SemanadeAno =DatePart(wk,S.fecha)
INTO TIEMPO
FROM Tiempo1 S
WHERE S.fecha IS NOT NULL
```

Por último, la tabla **FACT_CONSUMO**, también se transformará en la Tabla de Hechos **FACT_HOSPITAL**.

Las tareas descritas hasta el momento, pueden ser automatizados, contando con la posibilidad de poder actualizar nuestra Base de Datos Relacional con información de las tablas transaccionales, programando las tareas de carga de datos.

El siguiente paso, ahora, es definir la plataforma sobre la que será soportada todas las consultas desde las tablas del modelo entidad-relación hacia el datamart.

Lo que se ha realizado hasta el momento es definir un esquema relacional, sobre el que ejecutaremos Análisis Multidimensional (MOLAP) recordando que, según [PATRICIA ZVENGER], este modelamiento, nos brindará un análisis multidimensional de los datos.

Con respecto a los requerimientos sobre modelamiento OLAP, determinados en los análisis de [PATRICIA ZVENGER] y [ANDRE-FELIPE], debemos recalcar que se están tomando algunos criterios, pero no en su totalidad, debido a que no es punto de estudio del presente trabajo.

Es por eso que en temas como el análisis acerca de la necesidad de una reserva inicial de espacio, para un crecimiento futuro, así como de la necesidad de particionamiento de la Base de Datos y demás factores definidos, no son considerados al 100% en este trabajo.

3.2.5 SELECCIÓN DE ATRIBUTOS PARA EL ANÁLISIS DEL ALGORITMO.

A continuación definiremos los atributos a ser evaluados por el algoritmo:

SEXO:

Género de las personas que acuden al hospital. Sus posibles valores son:

M : Masculino

F: Femenino

ESTCIV:

Estado Civil de las personas que acuden al hospital. Sus posibles valores son:

S : Soltero

C: Casado

V: Viudo

D: Divorciado

CODLOGISTICO:

Código Logístico con el que se adquieren los medicamentos. En la data considerada, se encontró que se registraron 94 medicamentos. Sus posibles valores son:

A010250008 AMIKACINA 500 MG/2 ML
A010250007 AMIKACINA 100 MG/2 ML
A010250041 CEFTAZIDIMA 1 G
A010250139 VANCOMICINA 500 MG P/INF IV

A011100037	PIRIDOXINA 50 MG
A011100050	TIAMINA 100 MG
A010500020	SALBUTAMOL 100 MG P/INHAL AEROSOL
A010250086	GENTAMICINA 80 MG
A011050016	CLORURO DE SODIO 0.9 % X 1,000 ML
A011050042	MANITOL 20 % X 500 ML P/INF.IV
A011050074	SOLUCION PARA DIALISIS PERITONEAL (SISTEMA DESCONEXION)
A011050072	SOLUCION PARA DIALISIS PERITONEAL (SD) 1.5 % X 2 L
A011050076	SOLUCION PARA DIALISIS PERITONEAL (SD) 4.25 % X 2 L
A020101394	OBTURADOR DE PLASTICO PARA PROLONGADOR
A011050058	SOLUCION PARA DIALISIS PERITONEAL 1.5 % X 5 LITROS
A020100998	EQUIPO DE VENOCLISIS EMPAQUE INDIVIDUAL ESTERIL DESCART
A010850037	SELEGILINA 5 MG
A011000001	ALPRAZOLAM 0.5 MG
A010400046	ORCIPRENALINA 0.5 MG/ML
A010250101	METRONIDAZOL 500 MG
A010450021	LACTULOSA 3,33 G/5 ML JARABE X 100 ML O MAS
A010850022	FENOBARBITAL 100 MG/ML X 2 ML
A010250089	IMIPENEM + CILASTATIN 500 MG + 500 MG
A010900011	DERIVADOS DE METILCELULOSA GOTAS OFTALMICAS
A010050009	CODEINA FOSFATO 60 MG
A010750018	INSULINA NPH HUMANA 100 U.I./ML
A010250042	CEFTRIAXONA 1 G
A010150008	METILPREDNISOLONA (SODIO SUCCINATO, ACETATO) 500 MG
A010250080	FLUCONAZOL 100 MG P/INF.IV
A010750016	INSULINA CRISTALINA HUMANA 100 U.I./ML
A010250103	METRONIDAZOL 500 MG/100 ML
A010250003	ACICLOVIR 250 MG P/INF.IV
A010250095	KANAMICINA 1 G
A010750031	OCTREOTIDE (ANALOGO DE SOMATOSTATINA) 0.2 MG/ML
A011050014	CLORURO DE POTASIO 20 % X 10 ML
A010400093	GELATINA ENLAZADA 4 % X 500 ML
A011050024	DEXTROSA 10 % X 1,000 ML
A010050045	TRAMADOL (CLORHIDRATO) 50 MG/ML
A010350015	CICLOSPORINA 100 MG/ML X 50 ML SOLUCION O MICROEMULSION
A010700029	HEPARINA SODICA 5,000 U.I./ML
A010500005	BECLOMETASONA 50 MG P/INHAL AEROSOL

A011050010	ALBUMINA HUMANA 25 % X 50 ML
A010250036	CEFEPIME 1 G
A010400021	ENALAPRIL 10 MG
A010800026	VACUNA CONTRA LA HEPATITIS B MONODOSIS
A010400020	DOPAMINA 200 MG/5 ML P/INF.IV
A010700044	CONCENTRADO DE FACTOR VIII 250 U.I.
A010350017	CICLOSPORIN 50 MG (MICROEMULSION)
A010700014	FACTOR DE CRECIMIENTO DE COLONIAS GRANULOCITICAS Y MACR
A011050061	SOLUCION PARA DIALISIS PERITONEAL 4.25 % X 5 LITROS
A010200006	FLUMAZENIL 0.1 MG/ML X 5 ML
A011050027	DEXTROSA 5 % X 1,000 ML
A990000006	AGUA DE BICARBONATADA
A010250045	CIPROFLOXACINO 200 MG
A010250024	BENCILPENICILINA PROCAINICA 1,000,000 U.I. (CON DILUYEN
A010350042	FOLINATO CALCICO 15 MG
A010250009	AMINOPENICILINA/INHIBIDOR DE BETALACTAMASA 1,000/200-50
A010250043	CEFUROXIMA 750 MG
A010850006	BIPERIDENO 5 MG/ML
A010700001	ACIDO FOLICO 0.5 MG
A010650028	NISTATINA 25,000 U.I./G X 60 G CREMA VAGINAL
A010550012	CLOTRIMAZOL 1 % CREMA
A010650016	ESTROGENOS CREMA
A010750041	CARBONATO DE CALCIO 500 MG O MAS DE ION CA
A010650030	OXITOCINA 10 U.I./ML
A010700008	ERITROPOYETINA HUMANA 2,000 U.I.
A010350016	CICLOSPORINA 25 MG (MICROEMULSION)
A010700002	ACIDO TRANEXAMICO 1 G
A010250037	CEFOTAXIMA 0.5 G
A010250084	GANCICLOVIR 500 MG
A010400019	DOBUTAMINA 250 MG/20 ML P/INF.IV
A010250021	ANFOTERICINA B 50 MG P/INF.IV
A011050002	AGUA DESTILADA X 1,000 ML
A010250047	CIPROFLOXACINO 500 MG (TABLETA RANURADA)
A010400037	ISOSORBIDE DINITRATO 5 MG SUBLINGUAL
A010400060	NITROGLICERINA 5 MG/ML
A011050031	DEXTROSA 50 % X 1000 ML
A010450034	SALES DE REHIDRATACION ORAL (FORMULA OMS) 27.9 G PARA D
A010050035	PARACETAMOL 500 MG
A010250159	ACICLOVIR 400 MG

A010050018 IBUPROFENO 400 MG
A010250061 DICLOXACILINA 500 MG
A010050002 ALOPURINOL 100 MG
A010400039 LOVASTATINA 20 MG
A010850017 FENITOINA 100 MG
A011000009 DIAZEPAM 10 MG
A010750021 LEVOTIROXINA SODICA 0.1 MG
A010350057 MERCAPTOPURINA 50 MG
A010250133 SULFAMETOXAZOL + TRIMETROPRIMA 400 + 80 MG
A010500017 IPRATROPIO BROMURO 20 MG P/INHAL AEROSOL
A010250051 CLINDAMICINA 600 MG
A010100014 LIDOCAINA 2 % X 20 ML
A010250035 CEFAZOLINA 1 G
A010500013 FENOTEROL 0.5 % X 20 ML P/INHAL

CONTROLADO:

Indica si el medicamento despachado, es controlado o no. Esto quiere decir, que para su adscripción, necesita pasar por una Comisión especializada o no. Generalmente corresponde a medicamentos para cuyo diagnostico correspondiente son de gravedad. Sus posibles valores son:

S : SI

N: NO

DIAGNOS:

Corresponden a todos los diagnósticos catalogados en la Organización Mundial de la Salud. Los valores se detallan en [OMS].

PRESENTA:

Nos muestra las presentaciones con las que pueden presentarse los medicamentos. Sus posibles valores son:

AMPOLLETA
BALON
CAPSULA
CARTUCHO
CENTIMETRO CUBICO
EQUIPOS
FRASCO
GRAMO
OVULOS
POTE
SOBRE
SUPOSITORIO
TABLETA
TONELADA
TUBO
UNIDAD

SERVICIO:

Son los Servicios con los que cuenta el hospital. Sus posibles valores son:

CPQ: CIRUGIA PLASTICA

MI1: MED.1

NEF: NEFROLOGIA

URO: UROLOGIA

NER: NEUROLOGIA DES. VAS. CEREBRO

NEC: NEUROCIRUGIA

OTO: OTORRINO LARINGOLOGIA

NEU: NEUMOLOGIA

CTC: CIR. DE TORAX

HEM: HEMATOLOGIA CLINICA

ONC: ONCOLOGIA

CG1: CIR.1

GAS: GASTROENTEROLOGIA

END: ENDOCRINOLOGIA

DER: DERMATOLOGIA

GIN: GINECOLOGIA GENERAL

PQG: PSIQ. GENERAL

REU: REUMATOLOGIA

CIM: CIR. DE MANOS Y MICROCIR. EXTREMIDADES

OBA: OBSTETRICIA ALTO RIESGO

TRA: TRAUMATOLOGIA ORTOPEDIA

CLP: CLIN. PEDIATRICA

MI2: MED.2

MI3: MED.3

MI5: MED.5

CG2: CIR.2

CG3: CIR.3

CG5: CIR.5

CIP: CIR. PEDIATRICA

CCC: CIR. CAB. CUELLO. MAXIMOFACIAL

EME: EMERGENCIA

GER: GERIATRIA

UCI: SERV. CUIDADOS INTERMEDIOS

UTI: SERV. CUIDADO INTENSIVOS

ANE: ANESTESIOLOGIA

CGV: CIR.5 (PARES)

UCP: UTI PEDIATRICA

NUI: INTERMEDIOS-NEUROCIRUGIA

NUC: UCI NEUROCIRUGIA

UQT: U. QUEMADOS INTERMEDIO

ODO: ODONTOLOGIA

OFT: OFTAMOLOGIA

CG4: CIR.4

PTH: TRANSPLANTE/CIR. DE HIGADO.

UCN: UNID. CUID. INTERM. DE NEUMOLOGIA

UOB: SALA DE CUIDADOS ESPEC. PERINATALES

UIM: UNID. CUID. INTERM. DE MI2

UM1: UNID. CUID. INTERM. DE MI1

UNC: UNID. CUID. INTERM. DE NER

UM5: UNID. CUID. INTERM. DE MI5

URG: UROLOGIA GENERAL DAMAS

CPO: CLINICA PEDIATRICA ONCOLOGICA

UC3: SERV. CUIDADOS INTERMEDIOS

CLC: SERV. ATENC. DOMICILIARIA ESPECIALIZADA.

SEGURO:

Son los Servicios con los que cuenta el hospital. Sus posibles valores son:

HIJO
OBLIGATORIO__DEPEND.
CONYUGE
PENSIONISTA
VIUDEZ
SEGURO_PERSONAL
TERCERO
SERVIDORA_DEL_HOGAR
INVALIDEZ
SEGURO_FAMILIAR
FOPASEF
SEG.UNIVERSITARIO
SEG.INDEPENDIENTE
TRABAJADOR_IPSS
FACULT._CONTINUADOR
CONCUBINO
FACULT._INDEPENDTE
CONSTRUCCION_CIVIL
MAGISTERIO
HIJO_INCAPACITADO
AMA_DE_CASA
POR_REGULARIZAR
CHOFER_PROFESIONAL

3.2.6 APLICAR EL ALGORITMO K-MEANS PARA EL PROCESO DE CLASIFICACION E INTERPRETACION.

Tomando en cuenta que el algoritmo a utilizar es el K-means, vamos a explicar como trabaja este algoritmo sobre nuestra data, para finalmente poder obtener nuestros grupos (o sectores) de pacientes que es lo que finalmente se desea.

En el Capitulo II, se ve más detalladamente la forma como trabaja el algoritmo sobre los datos. Según lo visto en [GUSTAVO ADOLFO], y tomando como referencia una de las dimensiones de nuestra data, podemos correr el algoritmo de manera equivalente.

Utilizando una herramienta que contempla este algoritmo [WEKA], podemos realizar 3 visualizaciones:

1. La fuente de datos, producto de una consulta desde nuestro datamart:

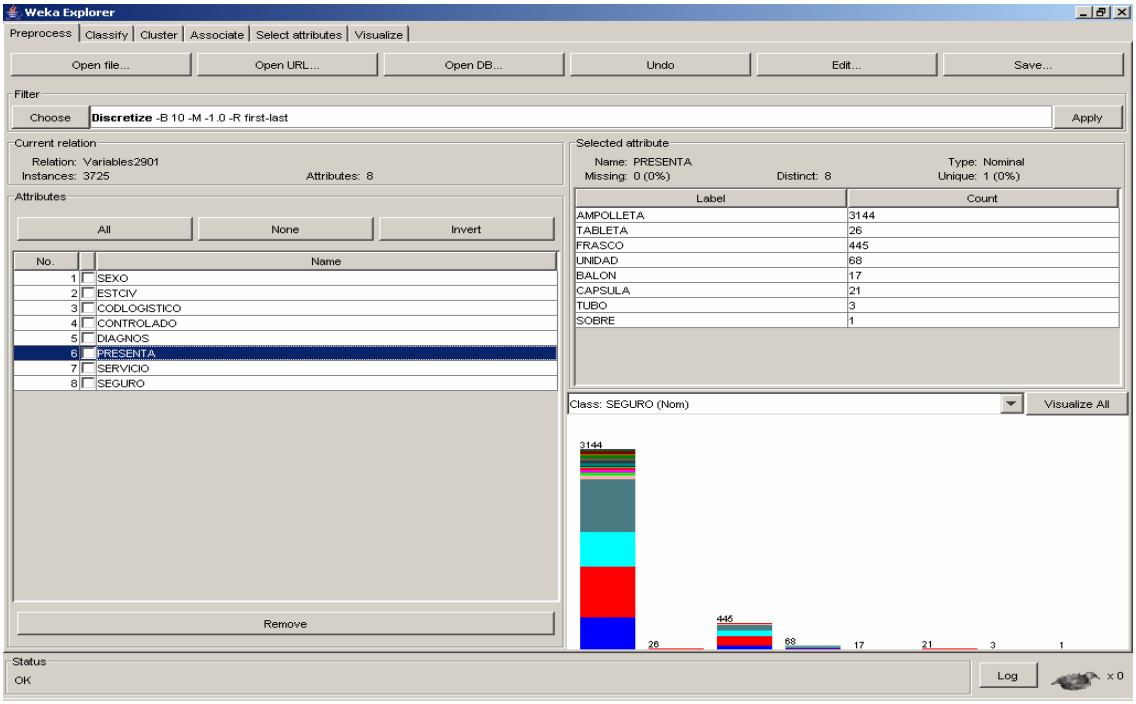


Figura 3.3. Visualización de Data Cargada.

En la parte izquierda, vemos los atributos seleccionados. En la parte derecha, vemos los valores del atributo seleccionado. Y en la parte inferior, vemos una gráfica con la relación entre el atributo seleccionado, y otros atributos (Distribución de Frecuencia).

2. A continuación, se trata debe hacer muchas pruebas de aplicación del algoritmo a la fuente de datos, con el objetivo de encontrar el mejor número de clusters para el proyecto.

El punto en el que el algoritmo encuentra los clusters adecuados, es cuando las características de cada cluster, no varían de iteración en iteración.

En este punto, usamos los siguientes parámetros:

NumClusters = 2 Speed=10

NumClusters = 3 Speed=10

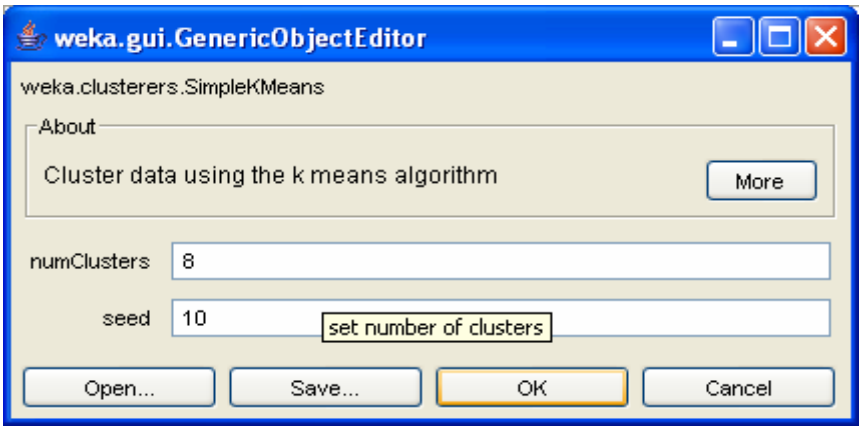
NumClusters = 4 Speed=10

NumClusters = 5 Speed=10

NumClusters = 6 Speed=10

NumClusters = 7 Speed=10

NumClusters = 8 Speed=10



```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 8 -S 10
Relation:     Variables
Instances:    3725
Attributes:   8
              SEXO
              ESTCIV
              CODLOGISTICO
              CONTROLADO
              DIAGNOS
              PRESENTA
              SERVICIO
              SEGURO
Test mode:    split 50% train, remainder test


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 12409.0

Cluster centroids:

Cluster 0
              Mean/Mode:  F C A010700029 N E11.5 AMPOLLETA MI2 PENSIONISTA
              Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 1
              Mean/Mode:  F C A010250042 N R50.9 AMPOLLETA MI3 CONYUGE
              Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 2
              Mean/Mode:  F C A010700029 N I77.0 AMPOLLETA URO CONYUGE
              Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 3
              Mean/Mode:  F C A011050072 N N18.0 FRASCO NEF OBLIGATORIO__DEPEND.
              Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 4
              Mean/Mode:  F C A010250139 S N39.0 AMPOLLETA UCI PENSIONISTA
              Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 5
              Mean/Mode:  M C A010250008 S G93.4 AMPOLLETA UTI OBLIGATORIO__DEPEND.
              Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 6
              Mean/Mode:  F C A010250080 S J15.9 AMPOLLETA UCN OBLIGATORIO__DEPEND.
              Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 7
              Mean/Mode:  M S A010250042 N R50.9 AMPOLLETA MI3 HIJO
              Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
```

```
=== Model and evaluation on test split ===

kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 5981.0

Cluster centroids:

Cluster 0
        Mean/Mode:  F C A010250042 N J15.9 AMPOLLETA MI3 OBLIGATORIO__DEPEND.
        Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 1
        Mean/Mode:  F S A010250042 N C95.9 AMPOLLETA MI3 HIJO
        Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 2
        Mean/Mode:  F S A011050072 N N18.0 FRASCO NEF CONYUGE
        Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 3
        Mean/Mode:  M S A010250139 S K70.3 AMPOLLETA UTI OBLIGATORIO__DEPEND.
        Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 4
        Mean/Mode:  M S A010250042 N E11.5 AMPOLLETA URO HIJO
        Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 5
        Mean/Mode:  M C A010250041 N J96.9 AMPOLLETA MI2 PENSIONISTA
        Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 6
        Mean/Mode:  M C A010250080 S N39.0 AMPOLLETA UCI PENSIONISTA
        Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 7
        Mean/Mode:  F C A010250089 N J96.0 AMPOLLETA URO CONYUGE
        Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A

Clustered Instances

0       517 ( 28%)
1       254 ( 14%)
2       179 ( 10%)
3       157 (  8%)
4       241 ( 13%)
5       246 ( 13%)
6       121 (  6%)
7       148 (  8%)
```

4. Con estos resultados, vemos que los clusters quedan como sigue:

    - El número de instancias que el algoritmo ha utilizado. Es el número de registros sobre el cual

actúa el algoritmo. En este caso son 3725 ítems o registros.

- El número de atributos por los que se van a agrupar u obtener los clusters. En este caso son 8: Sexo, Estado Civil, Controlado, Código Logístico, Diagnostico, Presentación, Servicio y Seguro.
- Vemos que el grupo que contiene mayor cantidad de población lo tiene el primer cluster (28%). Esto quiere decir, que la mayor cantidad de pacientes atendidos presenta las características del cluster encontrado.
- Los clusters encontrados en la data seleccionada utilizando el algoritmo K-Means:

Cluster 0: F C A010250042 N J15.9 AMPOLLETA MI3 OBLIGATORIO_DEPENDIENTE.
Sexo: FEMENINO
Estado Civil: CASADO
CodLog: A010250042 (CEFTRIAXONA 1 G)
Control: NO CONTROLADO
Diagnostico: J15.9 (NEUMONIA BACTERIANA, NO ESPECIFICADA)
Tipo de Presentación: AMPOLLETA
Servicio: MI3
Tipo de Seguro: OBLIGATORIO_DEPENDIENTE

Cluster 1: F S A010250042 N C95.9 AMPOLLETA MI3 HIJO
Sexo: FEMENINO
Estado Civil: SOLTERO
CodLog: A010250042 (CEFTRIAXONA 1 G)
Control: NO CONTROLADO
Diagnostico: C95.9 (LEUCEMIA, NO ESPECIFICADA)
Tipo de Presentación: AMPOLLETA
Servicio: MI3
Tipo de Seguro: HIJO

Cluster 2: F S A011050072 N18.0 N FRASCO NEF CONYUGE
Sexo: FEMENINO
Estado Civil: SOLTERO

CodLog: A011050072 (SOLUCION PARA DIALISIS PERITONEAL (SD) 1.5 % X 2 L)
Control: NO CONTROLADO
Diagnostico: N18.0 (INSUFICIENCIA RENAL TERMINAL)
Tipo de Presentación: FRASCO
Servicio: NEFROLOGÍA
Tipo de Seguro: CONYUGE

Cluster 3: **M S A010250139 N K70.3 AMPOLLETA UTI OBLIGATORIO__DEPENDIENTE**
Sexo: MASCULINO
Estado Civil: SOLTERO
CodLog: A010250139 (VANCOMICINA 500 MG P/INF IV)
Control: NO CONTROLADO
Diagnostico: k70.3 (CIRROSIS HEPATICA ALCOHOLICA)
Tipo de Presentación: AMPOLLETA
Servicio: UTI
Tipo de Seguro: OBLIGATORIO DEPENDIENTE.

Cluster 4 **M S A010250042 E11.5 N AMPOLLETA URO HIJO**
Sexo: MASCULINO
Estado Civil: SOLTERO
CodLog: A010250042 (CEFTRIAXONA 1 G)
Control: NO CONTROLADO
Diagnostico: E11.5 (DIABETES MELLITUS NO INSULINODEPENDIENTE, CON COMPLICACIONES CIRCULATORIAS PERIFERICAS)
Tipo de Presentación: AMPOLLETA
Servicio: URO
Tipo de Seguro: HIJO

Cluster 5: **M C A010250041 J96.9 N AMPOLLETA MI2 PENSIONISTA.**
Sexo: MASCULINO
Estado Civil: CASADO
CodLog: A010250041 (CEFTAZIDIMA 1 G)
Control: NO CONTROLADO
Diagnostico: J96.9 (INSUFICIENCIA RESPIRATORIA, NO ESPECIFICADA)
Tipo de Presentación: AMPOLLETA
Servicio: MI2
Tipo de Seguro: PENSIONISTA

Cluster 6: **M C A010250080 N39.0 N AMPOLLETA UCI PENSIONISTA**
Sexo: MASCULINO
Estado Civil: CASADO
CodLog: A010250080 (FLUCONAZOL 100 MG P/INF.IV)
Control: NO CONTROLADO

Diagnostico: N39.0 (INFECCION DE VIAS URINARIAS, SITIO NO ESPECIFICADO)
Tipo de Presentación: AMPOLLETA
Servicio: UCI
Tipo de Seguro: PENSIONISTA

Cluster 7 F C A010250089 J96.0 N AMPOLLETA URO CONYUGUE.
Sexo: FEMENINO
Estado Civil: CASADO
CodLog: A010250089 (IMIPENEM + CILASTATIN 500 MG + 500 MG)
Control: NO CONTROLADO
Diagnostico: J96.0 (INSUFICIENCIA RESPIRATORIA AGUDA)
Tipo de Presentación: AMPOLLETA
Servicio: URO
Tipo de Seguro: CONYUGUE

Esto quiere decir que dentro de toda nuestra información almacenada, tenemos varios grupos o sectores cuyo centro (*centroide*) presenta las siguientes características:

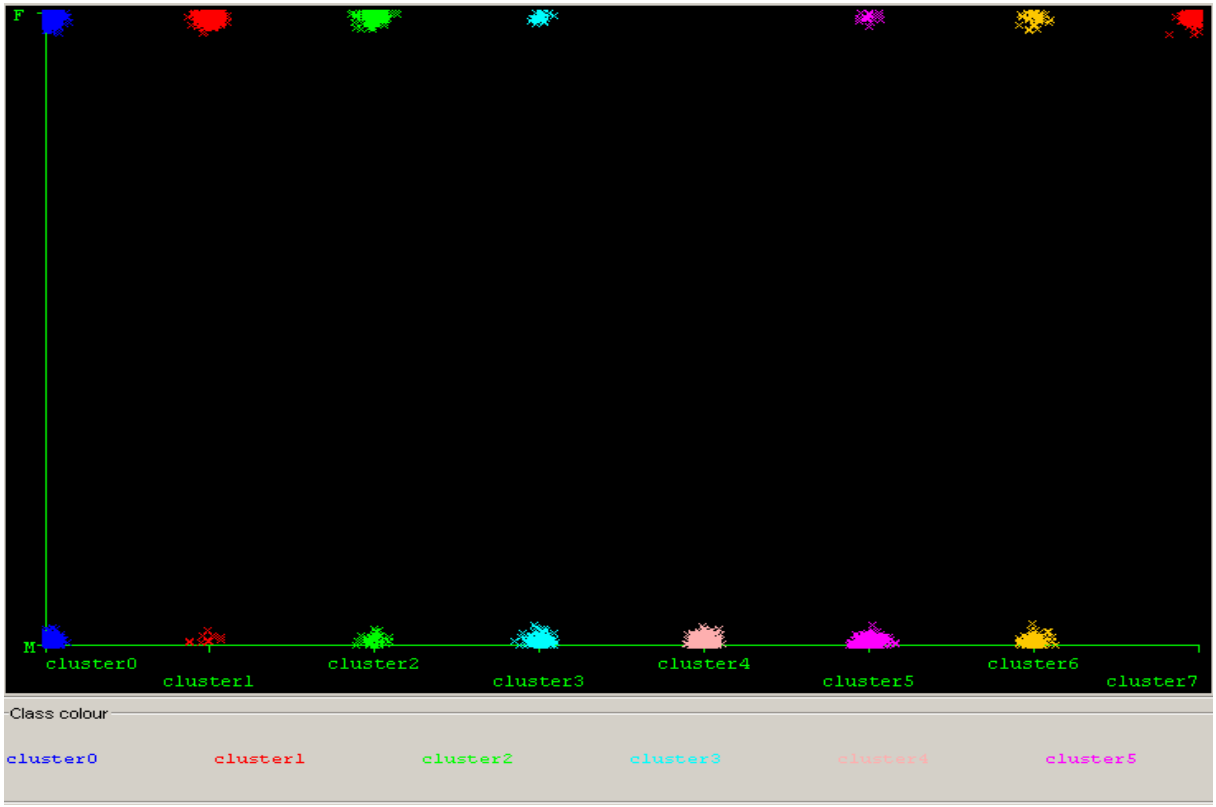
i. Mujeres con seguro de Obligatorio Dependiente, casados cuyo diagnóstico es Neumonía bacteriana, no especificada, procedentes de Medicina Interna 3. Los médicos que los tratan, les recetan medicamentos no controlados, mayoritariamente Ceftriaxona 1G. en presentación de Ampolleta.
ii. Mujeres con seguro de Hijo, solteros cuyo diagnostico es Leucemia, no especificada, procedentes de Medicina Interna 3. Los médicos que los tratan, les recetan medicamentos no controlados, mayoritariamente Ceftriaxona 1 G en presentación de ampolleta.
iii. Mujeres con seguro de Cónyuge, solteras cuyo diagnostico es Insuficiencia Renal Terminal, procedentes de Nefrología. Los médicos que los tratan, les recetan medicamentos no controlados, mayoritariamente Solución para diálisis peritoneal (SD) 1.5% x 2L en presentación de frasco.
iv. Varones con seguro de Obligatorio Dependiente, solteros cuyo diagnostico es Cirrosis hepática alcohólica, procedentes de UTI. Los médicos que los tratan, les recetan medicamentos No

Controlados, mayoritariamente Vancomicina 500 mg. p/inf IV en presentación de ampolleta.

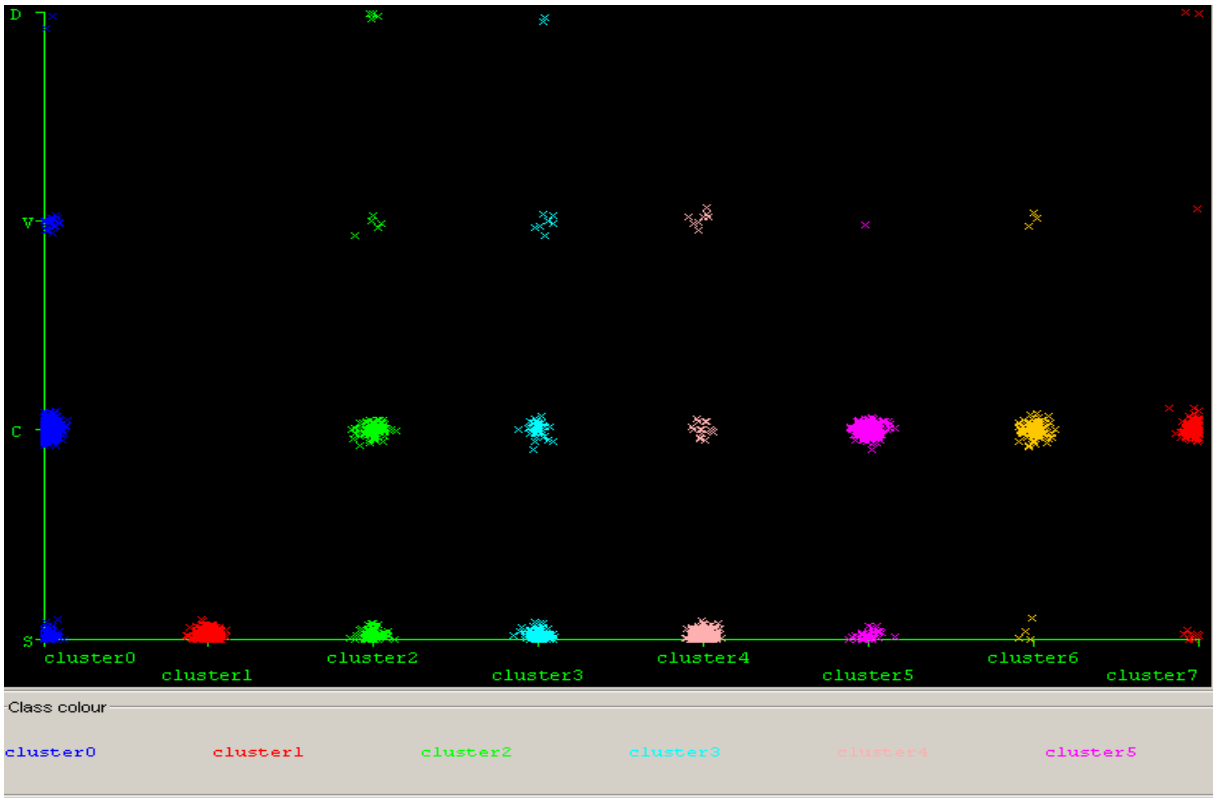
v. Varones con seguro de Hijo, solteros cuyo diagnostico es Diabetes mellitus no insulinodependiente, con complicaciones circulatorias periféricas, No Especificada, procedentes de Urología. Los médicos que los tratan, les recetan medicamentos No Controlados, mayoritariamente Ceftriaxona 1G en presentación de ampolleta.
vi. Varones con seguro de Pensionista, casados cuyo diagnostico es Insuficiencia Respiratoria, no especificada, procedentes de Medicina Interna 2. Los médicos que los tratan, les recetan medicamentos No Controlados, mayoritariamente Ceftazidima 1 G en presentación de ampolleta.
vii. Varones con seguro de Pensionista, casados cuyo diagnostico es Infección de vías urinarias, sitio no especificado, procedentes de Unidades de Cuidados Intermedios. Los médicos que los tratan, les recetan medicamentos No controlados, mayoritariamente Fluconazol 100 Mg. p/inf IV en presentación de ampolleta.
viii. Mujeres con seguro Cónyugue, casadas cuyo diagnostico es Insuficiencia respiratoria aguda, procedentes de Urología. Los médicos que los tratan, les recetan medicamentos no controlados, mayoritariamente Imipenem + Colastatin 500 Mg. + 500 Mg. en presentación de ampolleta.

5. Se inició utilizando como variables a evaluar o a agrupar, datos como: Sexo, Estado Civil, etc. pues lo que se desea, es confirmar los consumos de los mismos, y la existencia de **tipos** o **grupos** de pacientes que consumen medicamentos.
6. A continuación presentamos algunos gráficos correspondientes a los resultados del algoritmo:

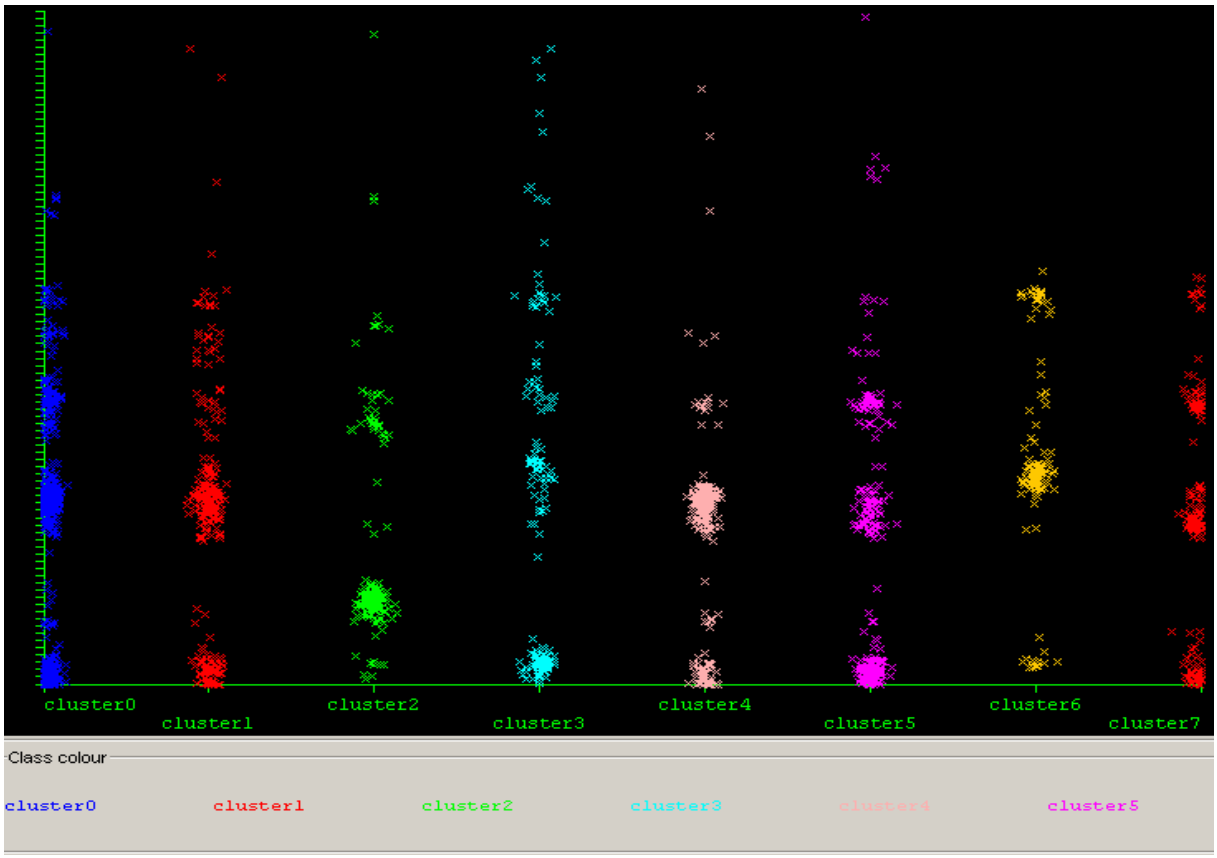
1. Resultado CLUSTERS vs. Atributo SEXO



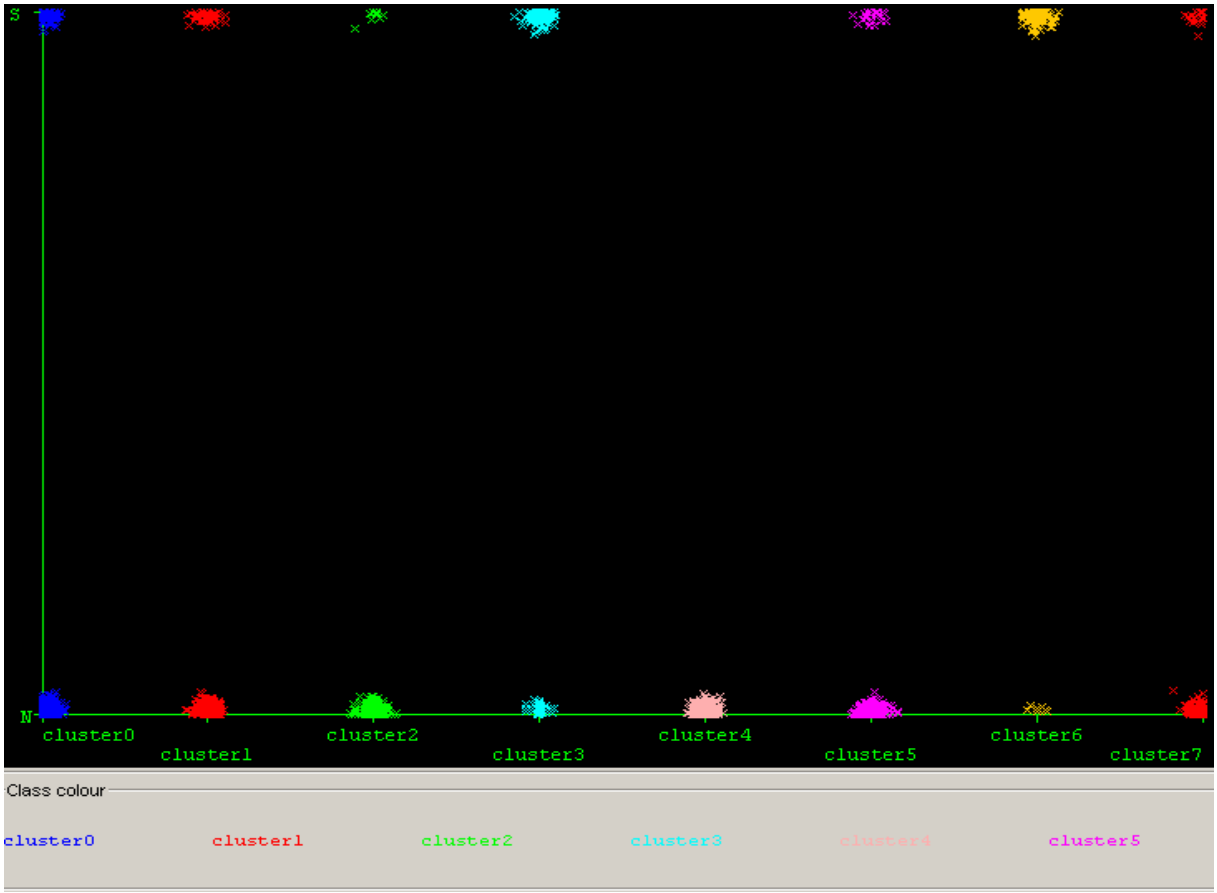
2. Resultado CLUSTERS vs. Atributo ESTADOCIVIL



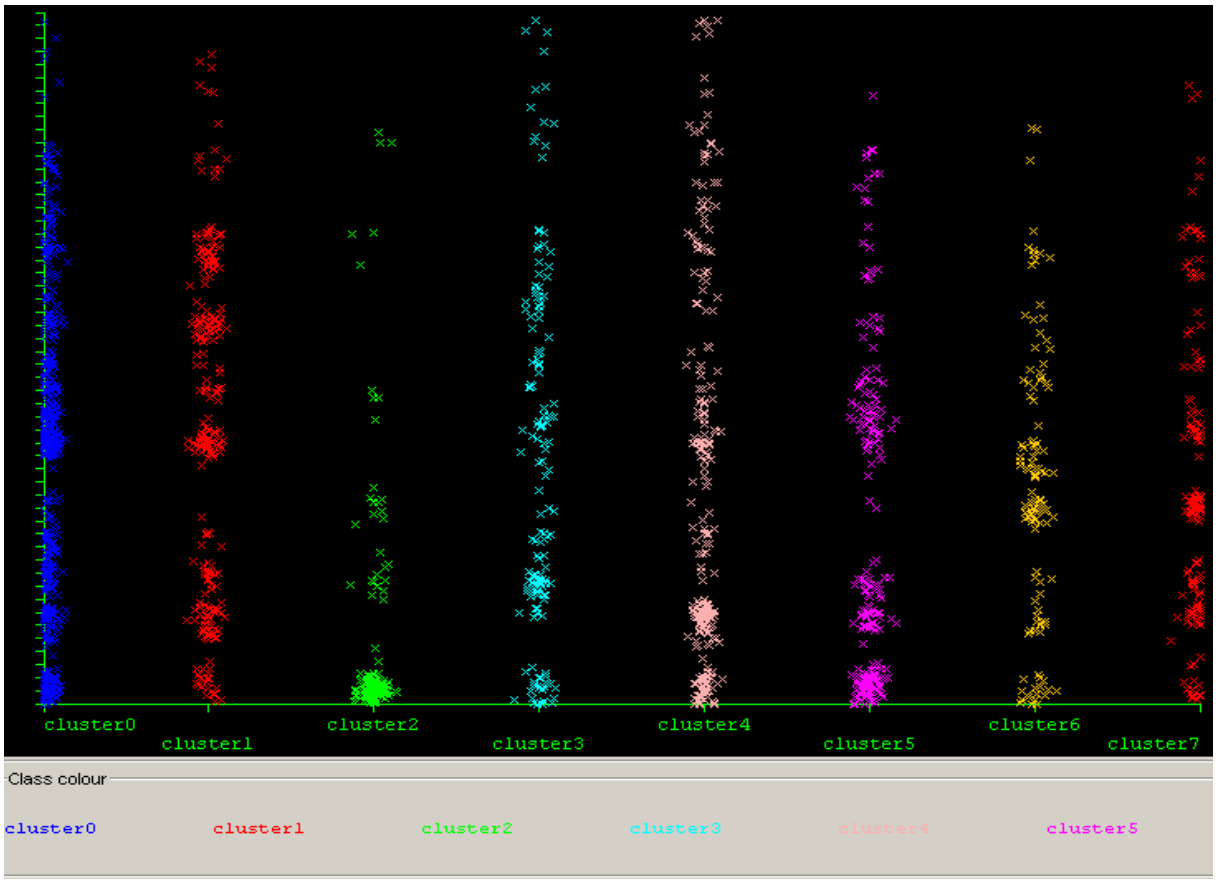
3. Resultado CLUSTERS vs. Atributo MEDICAMENTO



4. Resultado CLUSTERS vs. Atributo CONTROLADO



5. Resultado CLUSTERS vs. Atributo SERVICIO



Con estos gráficos, podemos analizar algunas características de los pacientes atendidos, en relación a sus consumos de medicamentos.

7. La idea, es poder reconocer qué pacientes, se encuentran en estos Clusters, ya que es lo que, a sugerencia de los médicos, les interesa analizar. Los médicos llevan un historial por grupos de pacientes. Este trabajo se focaliza en obtener grupos de clusters, e identificar qué pacientes se encuentran en estos clusters, para finalmente entregar los resultados a las áreas médicas para su respectivo análisis.

CAPÍTULO IV
RESULTADOS, DISCUSIÓN E INTERPRETACIÓN DE LA INVESTIGACIÓN

4.1 ANÁLISIS DEL ENTORNO DEL HOSPITAL NACIONAL GUILLERMO ALMENARA IRIGOYEN.

El Hospital Nacional Guillermo Almenara, es una institución con muchos años de existencia. La gran mayoría de la población peruana acude a sus instalaciones, para su atención y tratamiento.

El área que es de nuestro interés (el área de farmacia), consta de 12 módulos. Cada uno orientada a un grupo determinado de pacientes dentro del hospital. El o los módulos que son de nuestro interés, desarrollan sus labores concentrándose en el registro de los movimientos de cada paciente desde su ingreso hasta su alta. Proceso durante el cual, el paciente puede tener muchos movimientos y acudir una o varias veces al hospital por dolencias parecidas o no.

La información, actualmente, se registra por módulo. Es decir, la información de un paciente, como los movimientos de consumo de medicamentos puede ser registrada en distintos archivos para un mismo paciente. Esto es, debido a que los registros donde se almacenan la información, están divididos por archivo físico, más no lógico. Es decir, existen un archivo físico, para almacenar información de un paciente de un módulo, existe otro archivo físico, para almacenar información de un paciente de otro módulo, etc.

Los decidores del área, deben necesitar información del cruce de estos repositorios, debido a que se debe analizar la

información a un nivel más macro; viendo la información como un todo y no como islas de movimientos.

Por este motivo, al querer hacer un análisis de cómo se encuentran los consumos de medicamentos con una visión más global del área, se tiene que partir desde el análisis individual de cada módulo, y crear reportes cruzando las informaciones de módulos separados.
Con los sistemas tradicionales desarrollados en sistemas antiguos (FoxPro 2.6 para DOS) se preparan reportes ad-hoc para encontrar las respuestas a algunas las preguntas, pero se necesita dedicar aproximadamente un 60 % del tiempo asignado al análisis de localización y presentación de los datos, como también asignación de recursos humanos y de procesamiento del departamento de sistemas para poder responderlas, sin tener en cuenta la degradación de los sistemas transaccionales. Esta problemática se debe a que dichos sistemas transaccionales no fueron construidos con el fin de brindar síntesis, análisis, consolidación, búsquedas y proyecciones.

En todos los casos se observa la necesidad de considerar como punto de partida la información existente en las bases de datos de la institución. A continuación detallamos el sistema que hemos utilizado para la explotación de la información.

El Sistema de Gestión Hospitalaria, que está compuesto por:

- **Modulo de Admisión.-** Donde se manejan los datos de los pacientes, ya sean asegurados o no.

- **Modulo de Historia Clínica.-** Donde se maneja información propia de los pacientes que se encuentra almacenada en un documento llamado “Historia Clínica”.

- **Modulo de Consulta Externa.-** Aquí se registran los pacientes, previa cita, y que necesitan atenderse por alguna enfermedad, así como los que requieren de chequeos médicos, o tratamientos prolongados y programados por un médico tratante.

- **Módulo de Farmacia.-** Contiene información de los consumos de medicamento que son requeridos por los pacientes del hospital, prescritos por sus médicos tratantes. Este módulo se divide, de acuerdo a la procedencia de atención del paciente: Farmacia de Emergencia, Farmacia de Consulta Externa, Farmacia de Pensionistas, Farmacia de Dosis Unitaria, Farmacia de Hospitalización, Farmacia de Áreas Especiales, Farmacia de Transplante de Médula, Farmacia de Centro quirúrgico.

- **Módulo de Laboratorio.-** Se registran los exámenes clínicos de los pacientes de todo el hospital.

- **Módulo de Imagenología.-** Aquí se manejan los exámenes radiográficos de los pacientes.

- **Módulo de Anatomía Patológica.-** Aquí se registran las biopsias realizadas a los pacientes con sospecha de infecciones como cáncer, quistes, etc.

- **Módulo de Hospitalización.-** Aquí se registran a los pacientes que tienen estancia en el área de hospitalización. Generalmente, son de estadía prolongada.
- **Módulo de Emergencia.-** Aquí se registran aquellos pacientes que deben ser tratados con rapidez, aunque también pueden llegar a hospitalizarse.

- **Módulo de Centro Quirúrgico.-** Donde se registra los movimientos de los pacientes que se operan, es decir, los que ingresan a Sala Quirúrgica.

- **Módulo de Centro de Depósitos.-** Aquí se registran los materiales requeridos para un proceso de operación quirúrgica.

- **Módulo de Estadística.-** Opciones estadísticas de diferente índole.

- **Módulo de Facturación.-** A través de este módulo, se hace seguimiento a los pagos de los no asegurados.

Nuestra atención para este proyecto, será el Modulo de Farmacia, porque de aquí extraeremos la información sobre los consumos de medicamentos.

Módulo de Farmacia.

Este módulo, consta de 12 farmacias. A continuación, detallaremos las más importantes:

- **Farmacia de Emergencia:** Se registra el consumo de medicamentos de pacientes que están siendo tratados en el Departamento de Emergencia.

- **Farmacia de Consulta Externa:** Se registran los consumos de los pacientes con cita a los módulos de Consulta Externa.

- **Farmacia de Pensionistas:** Se registran los consumos de pacientes jubilados.

- **Farmacia de Hospitalización:** Se registran los medicamentos consumidos por los pacientes con permanencia en el hospital.

- **Farmacia de Centro Quirúrgico:** Se registra el consumo de medicamentos en las intervenciones quirúrgicas que se realizan a los pacientes.

Una característica en común con que cuentan todas las farmacias, es que existe una Comisión de Médicos (Comisión Evaluadora), que se encarga de controlar, qué tipo de medicamentos pueden ser prescritos por un médico, debido a que existen medicamentos que deben ser orientados a tratamientos específicos y críticos y no a tratamientos comunes o ambulatorios.

4.2 PRESENTACIÓN, ANÁLISIS E INTERPRETACIÓN DE LOS RESULTADOS.

4.2.1 DATAMART.

Interrogante:

¿Considera Ud. que se hace necesario el modelado y la construcción de un Datamart?

Interpretación:

De 50 personas encuestadas se obtuvo como resultados los siguientes:

- 12 respondieron “definitivamente si”, los cuales alcanzaron el 24.00% del total.
- 17 respondieron “probablemente si”, quienes fueron el 34.00% del total encuestado.
- 07 respondieron “probablemente no”, que fueron el 14.00% del total.
- 03 respondieron “definitivamente no”, que sólo conformaron el 6.00% del total.
- 11 fueron los que se mostraron nulos en su respuesta, ellos representan el 22.00% del total.

Se pudo apreciar que la mayoría de los participantes considera necesario el modelamiento y la construcción de un Datamart por ser una herramienta que permite hacer consultas complejas y de alto rendimiento a nivel del área de farmacia del Hospital Nacional Guillermo Almenara Irigoyen.

TABLA Nº 4.1

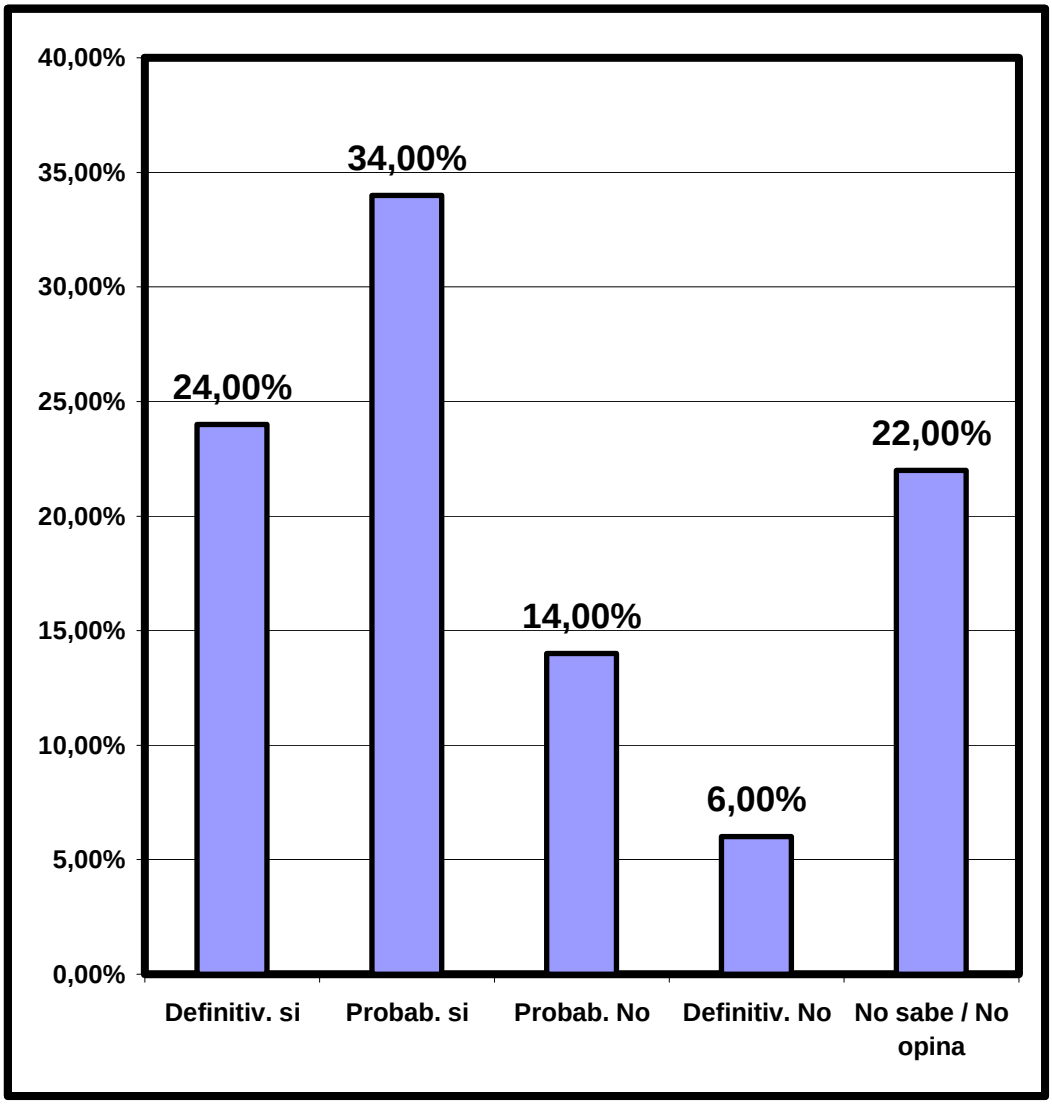
DATAMART

ALTERNATIVA	TOTAL PARCIAL	PORCENTAJE
Definitivamente si	*12*	*24,00%*
Probablemente si	*17*	*34,00%*
Probablemente no	*7*	*14,00%*
Definitivamente no	*3*	*6,00%*
No sabe / No opina	*11*	*22,00%*
TOTAL	**50**	**100,00%**

Fuente: Personas encargadas del Área de Informática y Jefaturas del Departamento de Farmacia del HNGAI.

FIGURA Nº 4.1

DATAMART



Fuente: Personas encargadas del Área de Informática y Jefaturas del Departamento de Farmacia del HNGAI.

4.2.2 IDENTIFICAR PACIENTES EN EL CONSUMO DE MEDICAMENTOS.

Interrogante:

¿Está Ud. de acuerdo con que se debe identificar pacientes sectorizándolos en el consumo de medicamentos?

Interpretación:

De todos los encuestados, 50 personas, se dieron como resultados los siguientes:

- 16 dijeron estar “totalmente de acuerdo”. Porcentualmente alcanzaron el 32.00% del total.
- 17 estuvieron “de acuerdo”. Porcentualmente fueron el 34.00% del total encuestado.
- 11 se mostraron nulos en su respuesta y fueron el 22.00% del total.
- 03 fueron los que estuvieron “en desacuerdo” y “totalmente en desacuerdo” cada uno de ellos; ellos sumaron el 12.00%

De los resultados obtenidos, la gran mayoría estuvo de acuerdo con la identificación de pacientes sectorizándolos y fundamentaron que a través de esta técnica se podrá llevar un control sobre los medicamentos que más se consumen por cada grupo de los pacientes, asimismo, tener un conocimiento de los medicamentos que deben ser abastecidos de manera inmediata por ser los de mayor aceptación por los pacientes.

TABLA Nº 4.2

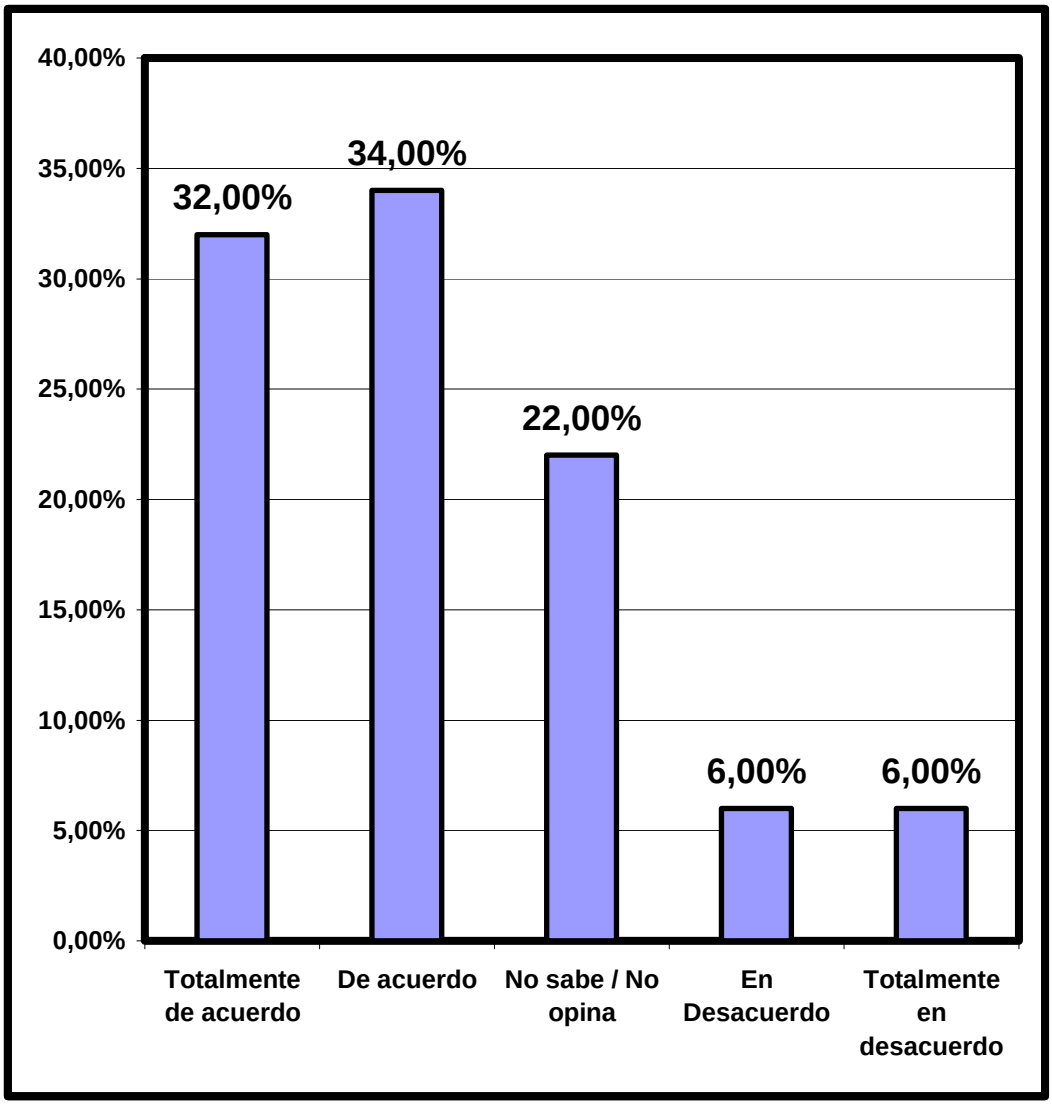
IDENTIFICAR PACIENTES EN EL CONSUMO DE MEDICAMENTOS

ALTERNATIVA	TOTAL PARCIAL	PORCENTAJE
Totalmente de acuerdo	*16*	*32,00%*
De acuerdo	*17*	*34,00%*
No sabe / No opina	*11*	*22,00%*
En Desacuerdo	*3*	*6,00%*
Totalmente en desacuerdo	*3*	*6,00%*
TOTAL	**50**	**100,00%**

Fuente: Personas encargadas del Área de Informática y Jefaturas del Departamento de Farmacia del HNGAI.

FIGURA Nº 4.2

IDENTIFICAR PACIENTES EN EL CONSUMO DE MEDICAMENTOS



Fuente: Personas encargadas del Área de Informática y Jefaturas del Departamento de Farmacia del HNGAI.

4.2.3 DATAMART Y ESTRATEGIAS EN LA TOMA DE DECISIONES

Interrogante:

¿Cree Ud. que con un Datamart para consultas OLAP se podría establecer mejores estrategias en la Toma de Decisiones?

Interpretación:

De los encuestados, que fueron un total de 50 personas, se obtuvo los siguientes resultados:

- 27 respondieron “definitivamente si”, y conformaron el 54.00% del total.
- 10 respondieron “probablemente si”, y fueron el 20.00% del total encuestado.
- 02 lo hicieron por “probablemente no”, y alcanzaron el 4.00%.
- 02 respondieron “definitivamente no”, y sólo alcanzaron el 4.00% del total.
- 09 optaron por “no sabe / no opina” y alcanzaron el 18.00% del total.

En definitiva, debido a la funcionalidad que tiene el Datamart, se pueden delinear nuevas y mejores estrategias que ayuden en la Toma de Decisiones a nivel del Área de Farmacia del HNGAI, tal como se muestra en el cuadro y gráfico siguientes.

TABLA Nº 4.3

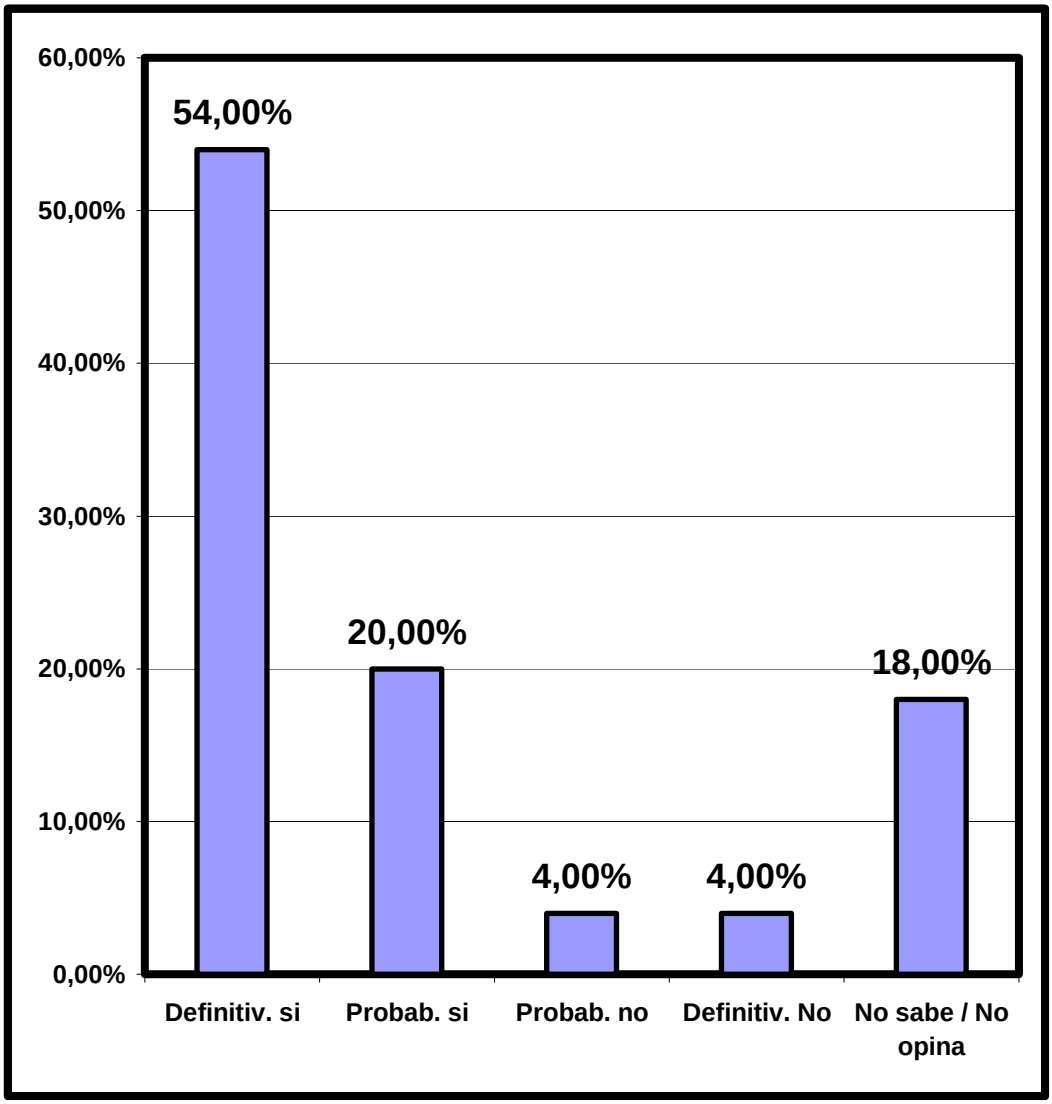
DATAMART Y ESTRATEGIAS EN LA TOMA DE DECISIONES

ALTERNATIVA	TOTAL PARCIAL	PORCENTAJE
Definitivamente si	*27*	*54,00%*
Probablemente si	*10*	*20,00%*
Probablemente no	*2*	*4,00%*
Definitivamente no	*2*	*4,00%*
No sabe / No opina	*9*	*18,00%*
TOTAL	**50**	**100,00%**

Fuente: Personas encargadas del Área de Informática y Jefaturas del Departamento de Farmacia del HNGAI

FIGURA Nº 4.3

DATAMART Y ESTRATEGIAS EN LA TOMA DE DECISIONES



Fuente: Personas encargadas del Área de Informática y Jefaturas del Departamento de Farmacia del HNGAI

4.2.4 MINERÍA DE DATOS.

Interrogante:

¿Considera Ud. necesaria la generación de pruebas de clasificación mediante la Minería de Datos para encontrar características similares en la información?

Interpretación:

De todos los encuestados, 50 personas, se muestra los resultados siguientes:

- 32 respondieron “probablemente si”, y sumaron el 64.00% del total.
- 08 respondieron “probablemente no”, y fueron el 16.00% del total.
- 10 respondieron “no sabe / no opina”, quienes fueron el resto de los encuestados, 20.00% del total.

Con respecto a los resultados de este numeral, observamos categóricamente que la Minería de Datos, mediante la generación de pruebas de clasificación, se podrá encontrar características similares de información, ya que a través de este proceso se logra extraer información y patrones de comportamiento que permanecen ocultos entre grandes cantidades de información.

TABLA Nº 4.4

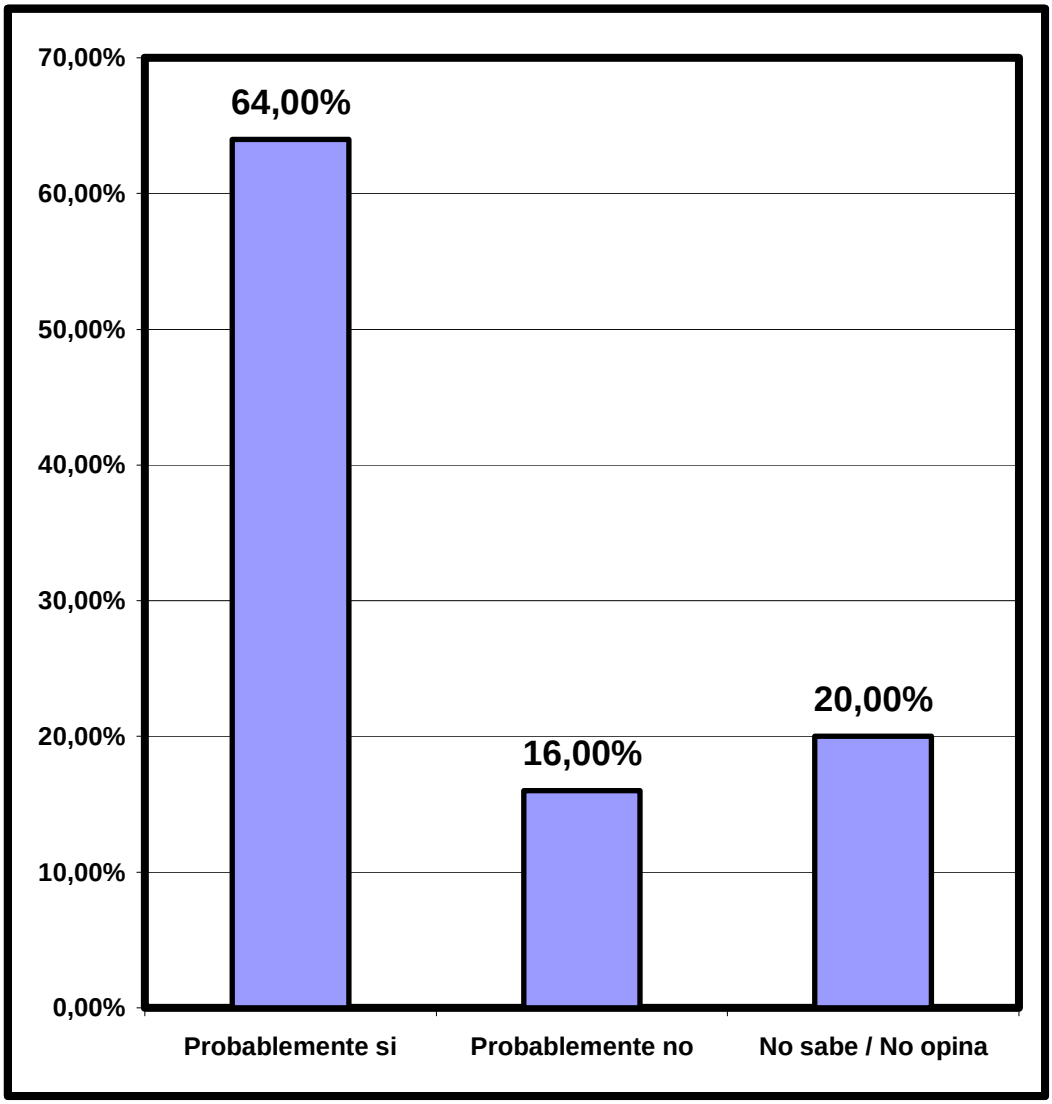
MINERÍA DE DATOS

ALTERNATIVA	TOTAL PARCIAL	PORCENTAJE
Probablemente si	*32*	*64,00%*
Probablemente no	*8*	*16,00%*
No sabe / No opina	*10*	*20,00%*
TOTAL	**50**	**100,00%**

Fuente: Personas encargadas del Área de Informática y Jefaturas del Departamento de Farmacia del HNGAI

FIGURA Nº 4.4

MINERÍA DE DATOS



Fuente: Personas encargadas del Área de Informática y Jefaturas del Departamento de Farmacia del HNGAI

4.2.5 MINERÍA DE DATOS Y LOS PROCEDIMIENTOS.

Interrogante:

¿Esta Ud. de acuerdo con que la Minería de Datos mejoraría los procedimientos durante la Toma de Decisiones?

Interpretación:

De todos los encuestados, 50 personas, se dieron como resultados los siguientes:

- 24 dijeron estar “totalmente de acuerdo”. Porcentualmente alcanzaron el 48.00% del total.
- 16 estuvieron “de acuerdo”. Porcentualmente fueron el 32.00% del total encuestado.
- 10 se mostraron nulos en su respuesta y fueron el 20.00% del total.
- 00 fueron los que estuvieron “en desacuerdo” y “totalmente en desacuerdo.

De acuerdo a los resultados obtenidos, y en concordancia con el ítem anterior, podemos deducir que la mayoría de encuestados está de acuerdo con que la Minería de Datos logrará mejorar los procedimientos en la Toma de Decisiones, ya que facilitará la extracción de datos que tiene que tomar en cuenta los Directores cuando van a tomar decisiones importantes que les compete a cada uno de ellos.

TABLA Nº 4.5

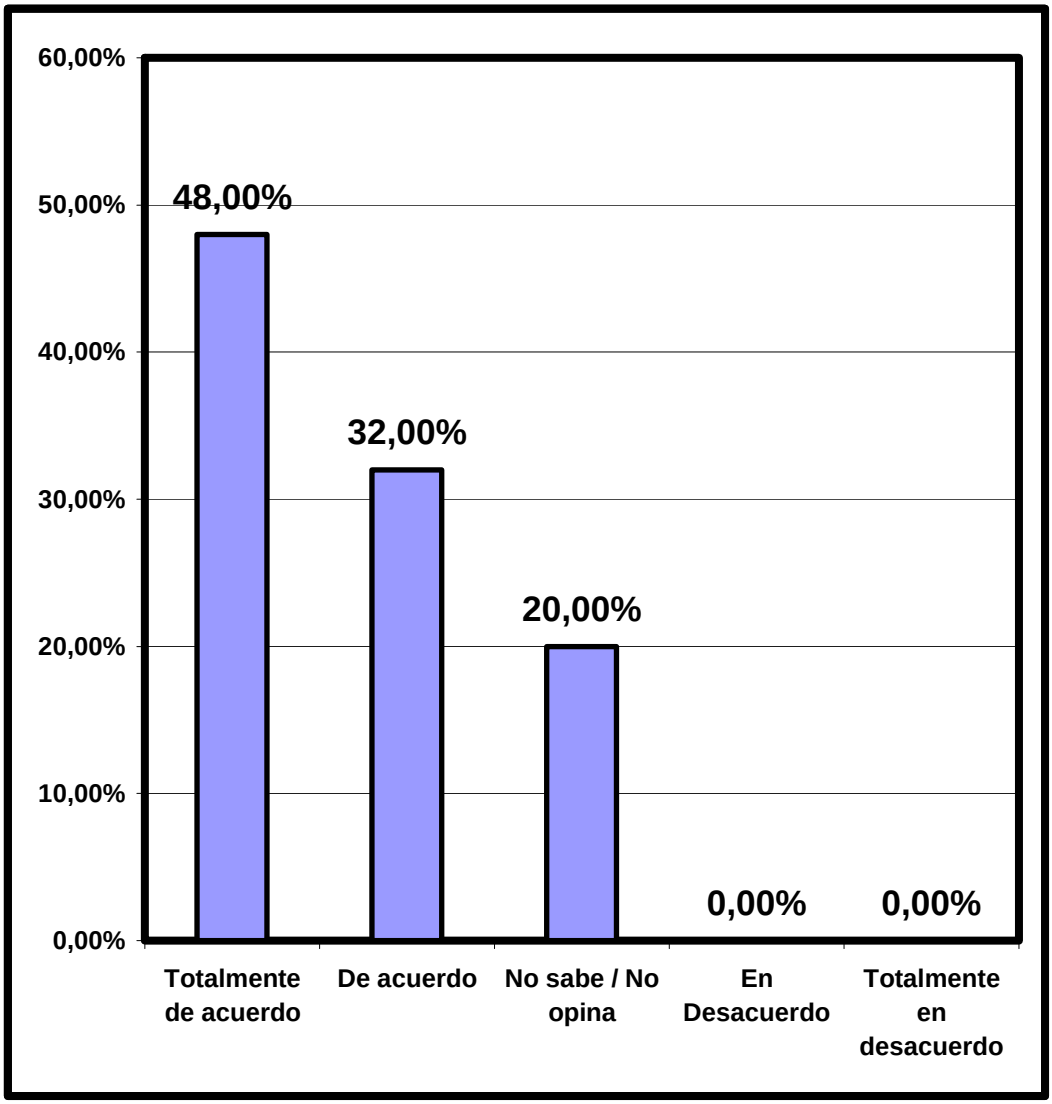
MINERÍA DE DATOS Y LOS PROCEDIMIENTOS

ALTERNATIVA	TOTAL PARCIAL	PORCENTAJE
Totalmente de acuerdo	*24*	*48,00%*
De acuerdo	*16*	*32,00%*
No sabe / No opina	*10*	*20,00%*
En Desacuerdo	*0*	*0,00%*
Totalmente en desacuerdo	*0*	*0,00%*
TOTAL	**50**	**100,00%**

Fuente: Personas encargadas del Área de Informática y Jefaturas del Departamento de Farmacia del HNGAI

FIGURA Nº 4.5

MINERÍA DE DATOS Y LOS PROCEDIMIENTOS



Fuente: Personas encargadas del Área de Informática y Jefaturas del Departamento de Farmacia del HNGAI

4.3 CONCLUSIONES.

Este proyecto, tomó las técnicas anteriormente mencionadas, y propuso una metodología que cumpla con el objetivo final definido.
En cada uno de los pasos de la metodología, se trató de aplicar la mejor técnica, ya que este proyecto no contempla la creación de ningún software para este fin, sino, explicar todos los métodos usados que cumplieron el objetivo planteado.

Los aportes principales de la tesis son:

- **Utilización de herramientas como Servicios de Minería.** Con lo cual se ha demostrado que se puede modelar sistemas de minería de datos, con algoritmos simples pero de mucha robustez para cualquier proyecto de clusterización.

- **El proyecto se convierte en el primer proyecto, enfocado en el análisis del consumo de medicamentos utilizando K-means.**
  Anterior a este proyecto, la institución no contaba con metodologías para analizar la información de manera distinta a la operacional o transaccional. Con este proyecto, se logra dos cosas:

  - Definir un modelo de trabajo para analizar cualquier área de interés de análisis de datos.

- Identificar al algoritmo K-means como el ideal para este tipo de proyectos donde se pretende **clasificar** la información contenida.

- **Existen otras áreas de estudio que también pueden resultar provechosas para la institución.**

  Del proyecto podemos deducir que, a pesar de constituir una metodología adecuada para encontrar algunas debilidades en el abastecimiento de medicinas, también se encontró una oportunidad, si se aplica la metodología a un área mas especifica como sería en el área de diagnósticos, debido a las siguientes razones:

  - Identificar los diagnósticos por los que los pacientes peruanos acuden a un hospital como la red EsSalud, nos permitiría conocer también, la realidad peruana en cuanto al tema de Salud personal y ambiental.

  - Identificar los diagnósticos mas comunes de la realidad peruana, también nos permitiría conocer la manera, como los medicamentos sugeridos por el personal medico, causan efecto a los pacientes. Si se llega a determinar que el consumo de un determinado medicamentos no causa cambios en el seguimiento de un determinado diagnostico, se podría pensar, que los criterios médicos, no es suficiente para solucionar problemas comunes, o, en su defecto, se podría evaluar la fórmula de

constitución de ese medicamento, como caso extremo.

Se pretende que este proyecto, sirva como modelo para futuros proyectos que tengan relación con la medicina, la psicología, y en todo campo donde se puede identificar tendencias de conductas o patrones de las mismas.

4.4 RECOMENDACIONES.

A continuación se describen las posibles ampliaciones del sistema que podrán implementarse para brindar más servicios de ayuda a la toma de decisiones, ya sea incorporando nuevas herramientas o nuevas funcionalidades.

1. **Complementar el trabajo, con Herramientas Especializadas de Inteligencia de Negocios.**
   El datawarehouse y datamart están diseñados de manera de facilitar la ampliación y crecimiento del proyecto. Se pueden adoptar herramientas como:

   - **Herramientas de Reporting:** Construcción de consultas avanzadas, distribución y visualización de información orientadas al usuario final. Adicionalmente, estas herramientas incorporan facilidades en la distribución de los reportes en la empresa.

- **Aplicación de Algoritmos de Predicción:** La mejora en este tema, estaría por el lado de la predicción. Existen algoritmos que facilitan la obtención de patrones de comportamiento, llegando al punto de predecir comportamientos de determinado segmento de pacientes. Para este trabajo en particular, se aplicaría a predecir justamente, los medicamentos que podrían sufrir desabastecimiento para cualquier área.

2. Ampliación de Áreas y Departamentos.

La oportunidad en este punto, es básicamente, extender el estudio hacia otras áreas como Laboratorio, las áreas de diagnóstico, etc. y poder complementar el análisis realizado en este trabajo.

A efectos de incorporar esta ampliación será necesario tener en cuenta la integración de todos los sistemas transaccionales, ya que determinada información que disponen algunas áreas y departamentos no está integrada con los principales sistemas del hospital.

BIBLIOGRAFÍA

- [ANDRE-FELIPE].
ASPECTOS DE CRIAÇÃO E CARGA DE UM AMBIENTE DE DATA WAREHOUSE.
André Fernandes Da Costa / Felipe Curvello A Anciães.
Universidad Federal do Rio De Janeiro. Diciembre 2001

- [FERNANDO-CARPANI].
CMDM: Un Modelo Conceptual para la Especificación de Base Multidimensionales
Fernando Carpani.
Inst. Computación.Universidad de la República. Agosto 2000

- [PATRICIA ZVENGER].
Introducción al soporte de Decisiones. Incorporación de Soluciones OLAP en entornos empresariales.
Patricia Andra Zvenger.
Dpto. Ciencias e Ing. De la Computación. Univ. Nacional del Sur. Diciembre 2005.

- [MOLINA GARCIA].
Técnicas de Análisis de Datos. Aplicaciones Prácticas utilizando Microsoft Excel y Weka.
José Manuel Molina López y Jesús García Herrera.
Universidad Carlos III – Madrid. 2004.

- [LUIS GABRIEL].
Uso de Técnicas de clasificación en conglomerados para describir perfiles en grandes bases de datos educativas.
Luis Gabriel Jaimes
Universidad de Puerto Rico. 2004

- [GUSTAVO ADOLFO].
Detección de Patrones en Imágenes Médicas.
Ing. Gustavo Adolfo Ferrero.
Instituto Tecnológico Buenos Aires - 2006

- [MAGDALENA SERVENTE].
Algoritmos TDIDT Aplicados a la Minería de Datos Inteligente.
Srta. Magdalena Servente.
Facultad de Ingenieria. Univ. Buenos Aires – 2002

- [SANDRA CARTAGENOVA].
Detección Automática de Reglas de Asociación.
Lic. Sandra G. Cartagenova.
Especialidad en Ingeniería en Sistemas Expertos. ITBA – 2005

MATERIAL DIGITAL

- [OMS] ORGANIZACIÓN MUNDIAL DE LA SALUD.
http://es.wikipedia.org/wiki/Lista_de_c%C3%B3digos_ICD-10
http://www.who.int/entity/es/

- [WEKA] SOFTWARE DE MINERIA DE DATOS DE LA UNIVERSIDAD DE WAIKATO.
http://www.cs.waikato.ac.nz/~ml/weka/

- [FREEDW] TUTORIAL DISEÑO DATAWAREHOUSE.
http://freedatawarehouse.com/tutorials/default.aspx

LISTA DE ABREVIATURAS

1. ALGORITMO “Archivo Invertido”:
   Un archivo invertido es aquel que permite hacer una búsqueda, en una colección de palabras, se compone de dos elementos: vocabulario y ocurrencias.

2. ALGORITMO “Binario”:
   Un algoritmo binario, tiende a generar árboles de muchos niveles. Por ello, el árbol resultante puede que no presente los resultados de manera eficiente, sobre todo si la misma variable ha sido utilizada para la división de varios niveles consecutivos.

3. ALGORITMO “B-tree”:
   Es un caso de clasificación "balanceado" (b = balanceado) de un directorio representado como un árbol. Trata de conservar las ramas del árbol de la misma longitud, para minimizar la lectura de nodos de directorio.

4. ALGORITMO “Hash”:
   Algoritmo que trasforma los datos en un resumen irreversible. Es decir, si pasamos los datos por un algoritmo Hash obtendremos un resumen único de esos datos, pero a partir del resumen no seremos capaces de volver a obtener los datos.

5. ALGORITMO “Sparse”:
   Son archivos en los que no se asigna bloques para almacenar secuencias de todos 0s, sino que las describe mediante meta-información almacenada.

6. BDT:
   Base de Datos Transaccionales. Es el almacén de datos, de actualización constante, donde se registran las operaciones diarias.

7. DATAMART:
   Es un subconjunto de un datawarehouse para un propósito específico (por ejemplo: Un datamart financiero, uno de marketing, etc.).

8. DATAWAREHOUSE:
Es una colección de datos orientadas a un dominio, integrado, no volátil y varía en el tiempo que ayuda a la toma de decisiones de la empresa.

9. DBMS:
DataBase Manager System. Sistema Administrador de Base de Datos.

10. DRILLDOWN:
Mostrar información de una fila en concreto, a un nivel mas detallado.

11. ERP:
Enterprise Resource Planning. Planificación de Recursos Empresariales. Consiste en una gran variedad de paquetes software, generalmente multi-modulares, que ofrecen soluciones integradas diseñadas para dar soporte a múltiples procesos de negocio.

12. ETL:
Extract, Transform and Load. Extracción, Transformación y Carga. Se refiere a manipular datos desde distintos origines, y trasformarlos hacia otro destinos de dato.

13. FOREIGN KEYS:
Llaves Foraneas. Concepto de base de datos que se refiere a todo campo que sirve para enlazar una tabla principal con otra secundaria.

14. HOLAP:
Hybrid OLAP: Es una combinación de los dos anteriores. Los datos agregados y precalculados se almacenan en estructuras multidimensionales y los de menor nivel de detalle en el relacional. Requiere un buen trabajo de análisis para identificar cada tipo de dato.

15. KDD:
Knowledge Discovery in Databases. Descubrimiento de Conocimiento en Base de Datos. Es el proceso de descubrir conocimiento útil dentro de los datos.

16. MDDB:
MultiDimensional DataBase. Base de Datos Multidimensional.

17. MOLAP:
Multidimensional OnLine Analytical Processing. Procesamiento Analítico Multidimensional en Línea. Es un proceso analítico en línea (OLAP) que indexa directamente en una base de datos multidimencional.

18. OLAP:
On-Line Analitical Process. Proceso analítico en línea, que se complementa con el concepto de datawarehouse, para que, juntos, brinden todas las funcionalidades de analisis multidimensional.

19. RDBMS:
Relational DataBase Manager System. Sistema Administrador de Base de Datos Relacional. Se refiere a cualquier fuente de datos con estructuras definidas para la administración de datos.

20. ROLAP:
Relational OnLine Analytical Processing. Proceso Analítico Relacional en Línea, es una forma de procesamiento analítico en línea (OLAP) que ejecuta análisis multidimencional sobre datos almacenados en una base de datos relacional, en vez de una base de datos multidimensional, como se considera el estándar de OLAP.

21. SGH:
Sistema de Gestión Hospitalaria. Es el sistema de propiedad de EsSalud del Perú, donde se almacena toda la información de los movimientos de los pacientes.

22. TimeStamp:
Grupo de fecha/hora. Proveniente del mundo Unix, como una forma genérica de designar el tiempo cronológico de forma completa: fecha, hora, minutos y segundos

23. TRANSACT-SQL:
Transact-SQL incluye instrucciones de control de flujo, y la capacidad de definir y usar procedimientos almacenados que incluyen ejecución condicio ucles.

24. WEKA:
Es una herramienta que permite realizar minería de datos con una interfaz gráfica lo que facilita su utilización. Además, permite una comparación con los distintos métodos que se utilizan para el pre-procesamiento, clasificación de información, clustering y meta-aprendizaje. WEKA proporciona una plataforma para evaluar un problema con distintas combinaciones de algoritmos y poder extraer conocimiento interesante.