Una metodología para sectorizar pacientes en el consumo de medicamentos aplicando datamart y datamining en un hospital

Recepción: Febrero de 2007 / Aceptación: Mayo de 2007

[1] Ivan Tapia Rivas
[2] María Ruiz Rivera
[3] Edgar Ruiz Lizama

RESUMEN

El presente trabajo, propone un método para el análisis de datos en la forma con que se consumen los medicamentos en un hospital peruano a fin de poder identificar algunas realidades o características no observables que producirían desabastecimiento o insatisfacción del paciente, el cual servirá como una herramienta para la toma de decisión sobre el abastecimiento de medicamentos en el hospital. Aquí se complementan técnicas de datamart, de extracción y carga de datos, así como algoritmos de minería de datos como K-means para sectorizar los consumos de medicamentos mencionados. Finalmente se muestran gráficas con algunos resultados extraídos de las pruebas.

Palabras Clave: Minería de datos, inteligencia de negocios, algoritmo K-Means, algoritmo de clasificación.

A METHODOLOGY TO CLASSIFY PATIENTS IN THE MEDICINE'S CONSUMPTIONS APPLYNG DATAMART AND DATAMINING A HOSPITAL

ABSTRACT

The present work, provides a method for the analysis of data in the way with which the medicines in a Peruvian hospital are consumed, in order to identify some nonobservable realities or characteristics which would produce shortage of supplies or insatisfaction of patient, serving as a tool for the decision making on the medicine supplying in the hospital. Here techniques of datamart are complemented, of extraction and load of data, as well as algorithms of mining suchs K-means to classify the mentioned medicine consumptions. Finally there are graphs with some extracted results from the tests.

Key Words: Datamining, business intelligence, K-Means algorithm, classification algorithm.

INTRODUCCIÓN

Las organizaciones dedicadas a la atención de la salud, 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 con el apoyo de reportes que les permite 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 indicadores que muestran si se llegó a un nivel mínimo para reabastecer.

Es aquí donde los algoritmos de minería de datos para la clasificación, nos permite saber cual es la composición del público con el que cuenta el hospital, y podrá saber realmente lo que sucede con el consumo de medicamentos, pudiendo así, abastecerse solo con los medicamentos necesarios.

PLANTEAMIENTO DEL PROBLEMA

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

No se tiene un modo de conocer, de antemano, porqué es que ciertos medicamentos son mas despachados, ni tampoco se conoce el tipo de pacientes que solicitan medicinas de un tipo u otro, o si los números que aparecen en los reportes, contienen algo más de información que solo valores estadísticos.

En resumen, no se tiene una clasificación de los pacientes que acuden al hospital, de acuerdo a las características de sus consumos de

(1) Ingeniero de Sistemas e Informática, UNMSM.
E-mail: diurvan@hotmail.com.
(2) Licenciada en Computación. Profesora de la Facultad de Ingeniería de Sistemas e Informática, UNMSM.
E-mail: mruizr@unmsm.edu.pe
(3) Magíster en Informática. Profesor de la Facultad de Ingeniería Industrial, UNMSM.
E-mail: eruizl@unmsm.edu.pe

medicamentos, que ayudaría a la toma de decisiones en el abastecimiento de medicamentos al hospital.

Para dar solución al problema se plantea se toma una muestra de un periodo determinado, desde la fuente de datos transaccional de un sistema desarrollado para el control de medicamentos e historias de pacientes.

La muestra correspondiente a un periodo de tiempo, pasa por una etapa de limpieza de datos, luego se modela y desarrolla una base de datos dimensional, convirtiendo la información extraída, en forma de cubos y dimensiones para poder ser analizada a manera de cubos.

La base de datos servirá para aplicar algoritmos de clasificación. En este caso, se utilizó el algoritmo Simple K-Means, el cual tiene por objetivo, formar grupos o "clusters"; donde cada "cluster" equivale a grupos con características similares. Obviamente la información que consiste en pacientes consumiendo medicamentos, utilizará solo algunas métricas para realizar el análisis.

MARCO CONCEPTUAL

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. Existen arquitecturas para poder crear datawarehouses:

1. **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. El diagrama de tablas relacionales en la figura 1, 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. [2].

2. **Esquema Copo de Nieve**
   La figura 2 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

Figura 1. Esquema Estrella

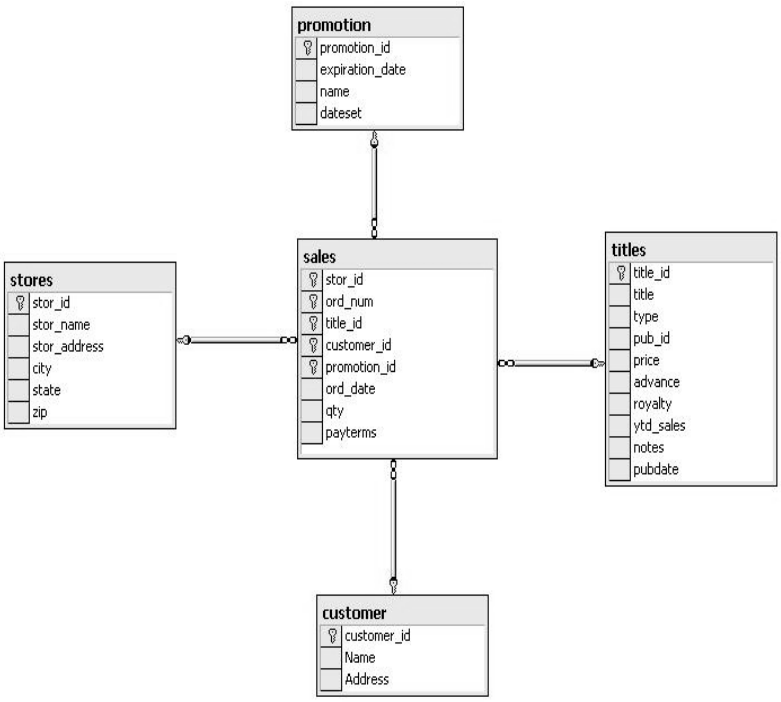
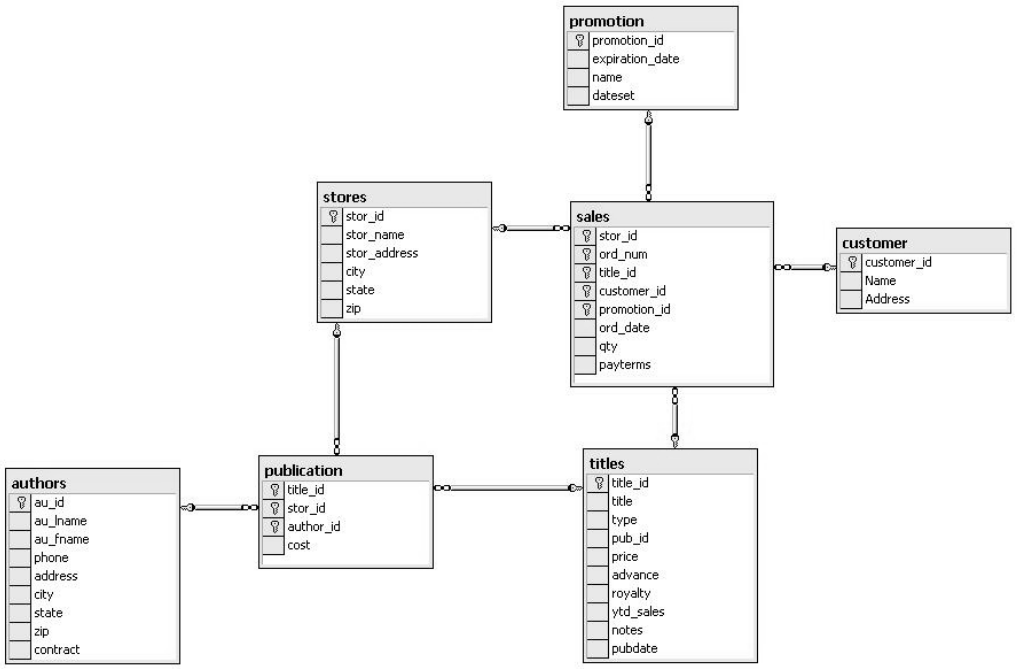


Figura 2. Esquema Copo de Nieve



unión para responder a las consultas, lo que empeora la performance, además del mantenimiento que requieren las tablas adicionales. [2]

Datamart
Son versiones más pequeñas de Datawarehouse. Estas versiones se crean usando algún criterio particular, como por ejemplo el lugar geográfico. Algunas corporaciones reemplazan completamente el concepto de tener un Datawarehouse central, por varios datamarts más pequeños que se alimenten directamente de los sistemas operacionales. [2]

Almacenamiento OLAP (procesamiento analítico on-line)
OLAP se define como el análisis multidimensional e interactivo de la información de negocios a escala empresarial. Consiste en combinar distintas áreas de la organización, y así ubicar ciertos tipos de información que revelen el comportamiento del negocio. [2]

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. [2]
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.

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.

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. [2]

Medidas o métricas
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, etc.

Jerarquías 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.

Datamining
Es 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 Datamining clasifican y predicen futuras tendencias y comportamientos.
Los algoritmos de clustering (o clasificación) identifican 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. [4]

Algoritmo K-Means
Uno de los algoritmos más utilizados para hacer clustering es el k-medias (kmeans), que se caracteriza por su sencillez. [3]

a. 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.
b. A continuación cada una de las instancias, ejemplos, es asignada al centro del cluster más cercano de acuerdo con la distancia Euclideana que le separa de él.
c. 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.
d. Finalmente se repite el proceso completo con los nuevos centros de los clusters.
e. 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. [3]

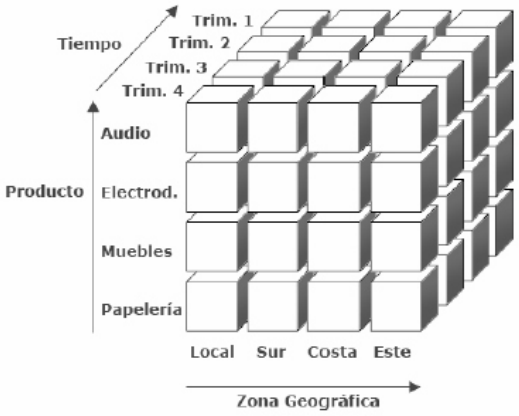
Para obtener los centroides, se calcula la media o la moda según se trate de atributos numéricos o simbólicos.

Ejemplo 1.
En la primera columna del cuadro 1, 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.
Se recalculan los centros, como el promedio de las distancias dentro de cada conglomerado, de la siguiente manera: (mínima d.) de cluster1 / nro. Ítems de cluster1.

Entonces: Cluster1: 17/4= 4.25. Cluster2: 17/6=2.83.
Los nuevos centroides entonces son: 4.25, 2.83.

Ahora, en este cuadro 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.

Figura 3. Cubo multidimensional.



Cuadro 1. Ejemplo de algoritmo K-Means

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

APLICACIÓN PRÁCTICA

A continuación se describen los pasos que se siguen en el trabajo:

Entender y analizar el problema
En esta etapa se empieza delimitando el interés del estudio a las tablas o consultas que se tomarán de toda la base de datos existente. Esta elección, implica conocimiento del negocio, debido a que se debe definir solo las tablas o consultas que serán de interés para el análisis.

En este estudio, se tomaron las tablas ADSERVIC (Servicios), ADTASEG (Tipo de Seguro), DIAGNOS (Diagnósticos), FM_PRECE (Presentación del medicamento), MEDICAME (Medicamentos), MEDICO, ADHISCLI (Pacientes), POLICLIN (Centros de Atención), TIEMPO, RECETA. Existen otras tablas como CENTRO_COSTO, PARAMETRO, UBIGEO, etc. que no son importantes para este análisis.

Limpiar los datos
Este paso es la etapa que toma más tiempo, debido a que generalmente las transacciones realizadas en el hospital en estudio, son hechas sin ningún tipo de control en el ingreso de datos. Es por esto, que hay que verificar en la mayoría de los datos, y ejecutar algunas consultas simples de transacciones (Transact-SQL), si la información existente es válida o tiene, en algunos casos, datos inconsistentes con el tipo de dato almacenado.

Ejemplos de esto es la validación de fechas, validación de saldos, campos en blanco, campos de error (caracteres inválidos), etc.

El trabajo se realiza de forma manual, o utilizando algunas consultas simples, pero básicamente se tiene que explorar dentro de la data, para encontrar alguna inconsistencia de dato, o integridad o de algún otro tipo.

Cuadro 2. Continuación ejemplo de algoritmo K-Means

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

Diseñar el esquema dimensional
Se definen las Dimensiones y Jerarquías:

Dimensión: DIAGNOSTICO.
Jerarquías: Descripción.
Dimensión: MEDICAMENTO.
Jerarquías: Control (Medicamento controlado o no), CodLog (Código Logístico)
Dimensión: MEDICO.
Jerarquías: Descripción.
Dimensión: PERSONA.
Jerarquías: Sexo, EstadoCivil.
Dimensión: POLICLINICO.
Jerarquías: Descripción.
Dimensión: PRESENTACION.
Jerarquías: Descripción.
Dimensión: SERVICIO.
Jerarquías: AbreviaciónServicio.
Dimensión: TIEMPO.
Jerarquías: Año, Trimestre, Mes, Dia.
Dimensión: TIPOSEGURO.
Jerarquías: Descripción.

Se define la Tabla de Hechos: FACT_HOSPITAL

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, primero se ha hecho la 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. Después de ejecutar la sentencia, se tienen tablas organizadas de la siguiente manera:

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, etc.

Selección de atributos para el análisis
Se definieron atributos a ser analizados:
SEXO, ESTADO CIVIL, CODIGO LOGISTICO DE MEDICAMENTO, CONTROLADO (Indica si un medicamento necesita evaluación de una Junta Médica para ser adscrita a un paciente), DIAGNOSTICO (Utiliza el Código C-DIAG de la OMS), PRESENTA (Tipos de Presentación del medicamento), SERVICIO (Especialidad del hospital), SEGURO (Clase de seguro con la que cuenta el paciente).

Figura 4. Esquema dimensional del Datamart.

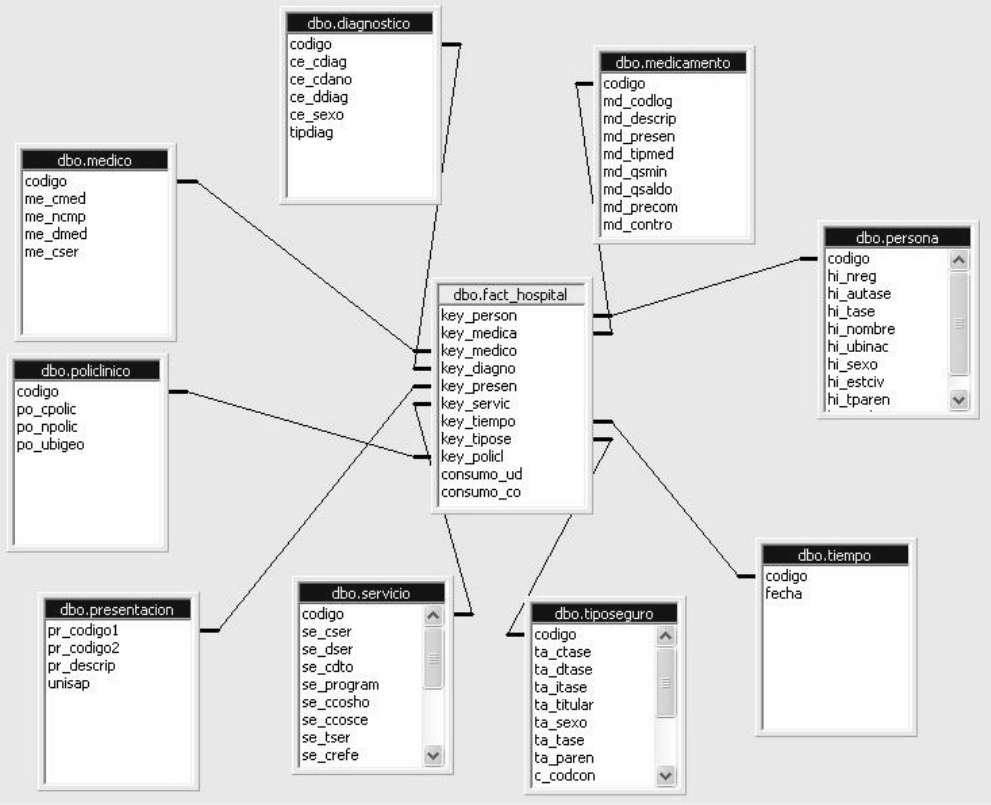
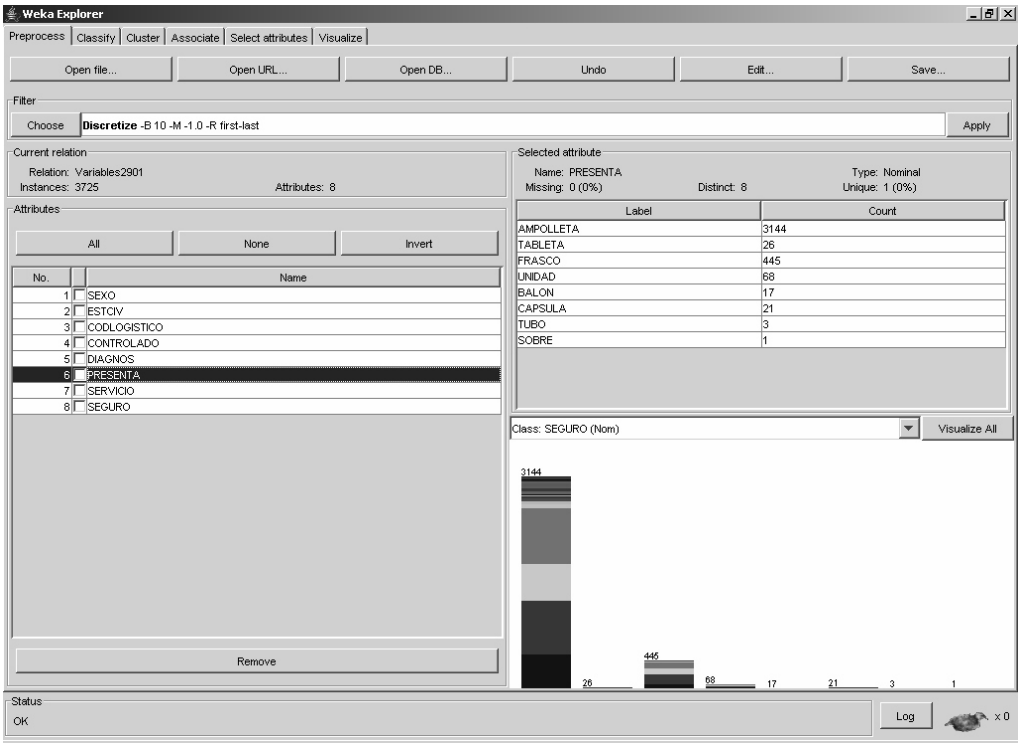


Figura 5. Visualización de data cargada a Weka.



Aplicación del algoritmo k-means

En este punto se utilizó un software de código abierto llamado Weka[5].

Se necesita transformar la consulta específica hacia un formato reconocible por este software, el cual, también es un estándar (Delimitados por comas o sino, el formato ARFF).

En la figura 5 se observa la información cargada en el software: En la parte izquierda, se ve los atributos seleccionados. En la parte derecha, se tiene los valores del atributo seleccionado. Y en la parte inferior se muestra una gráfica con la relación entre el atributo seleccionado, y otros atributos (Distribución de Frecuencia).

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 se emplean 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

El resultado después de ejecutar el algoritmo se muestra a continuación:

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

Los resultados de explican de la siguiente manera:

i. Mujeres con seguro de Obligatorio Dependiente, casadas 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 Hijas de asegurado, con diagnóstico de 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 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 diagnóstico 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 diagnóstico es Diabetes Mellitus no insulinodependiente, con complicaciones

Figura 6. Parámetros del algoritmo.

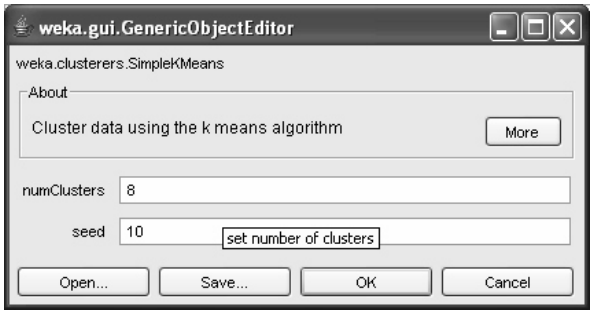


Figura 7. Clusters Vs. Servicios.

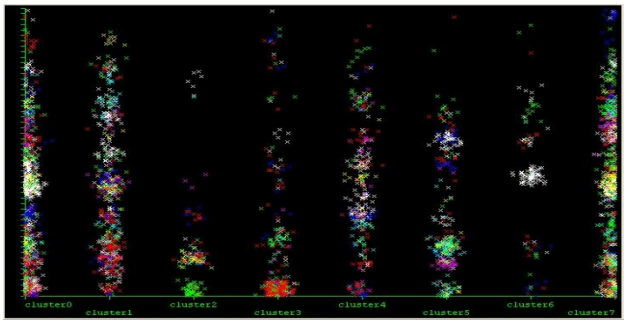
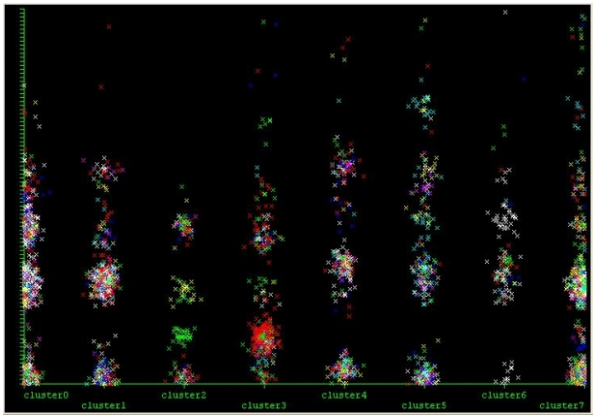


Figura 8. Clusters vs. Medicamento.



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 diagnóstico 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 diagnóstico 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 Cónyuge, casadas cuyo diagnóstico es Insuficiencia respiratoria aguda, procedentes de Urología. Los médicos que tratan, les recetan medicamentos no controlados, mayoritariamente Imipenem + Colastatin 500 Mg. + 500 Mg. en presentación de ampolleta.

CONCLUSIONES Y RECOMENDACIONES

En el trabajo se toman las técnicas mencionadas, y se propone una metodología tal que cumpla con el objetivo propuesto. En cada uno de los pasos de la metodología, se trató de aplicar la mejor técnica, ya que el trabajo no contempla la creación de ningún software para este fin, sino, explicar la metodología planteada.

Se considera que la técnica de minería de datos utilizada es oportuna y demuestra que se puede modelar sistemas de minería de datos, con algoritmos simples pero de mucha robustez para cualquier proyecto de clusterización.

El modelo sirve para comparar los resultados contra el de otras instituciones del mismo rubro pero en otras áreas de Latinoamérica, a modo de conocer las realidades de las poblaciones.

Se recomienda complementar el trabajo, aplicando el estudio hacia el análisis de diagnósticos, o aplicando algoritmos de predicción como redes neuronales.

REFERENCIAS BIBLIOGRÁFICAS

1. Fernandes, A. y Curvello, F. (2001). Tesis de Bachillerato: “Aspectos de criação e carga de um ambiente de data warehouse”. Universidad Federal do Rió De Janeiro, Brasil.

2. García Molina. (2004). Apunte o Articulo: “Técnicas de Análisis de Datos. Aplicaciones Prácticas utilizando Microsoft Excel y Weka”. José Manuel Molina López y Jesús García Herrera. Universidad Carlos III Madrid, España.

3. Servente M**.** (2002)**.** Tesis de Grado en Ing. Informática: “Algoritmos TDIDT Aplicados a la Minería de Datos Inteligente”. Facultad de Ingeniería. Universidad de Buenos Aires**,** Argentina.

4. Weka.
En: http://www.cs.waikato.ac.nz/ml/weka/
(Visitado: 30-03-07).

2. Zvenger, P. (2005). Tesis de Licenciatura: Introducción al soporte de Decisiones. Incorporación de Soluciones OLAP en entornos empresariales”. Dpto. Ciencias e Ing. de la Computación. Universidad Nacional del Sur, Argentina.