# Minería de Datos en Base de Datos de Servicios de Salud

**Monserrat, Sergio**
***Estudiante Maestría en Ingeniería en Sistemas de Información UTN – FRSF***
**Chiotti, Omar**
***Cidisi – UTN – FRSF, Ingar UTN- CONICET***

**Resumen**
*La minería de datos tiene por propósito procesar datos de un dominio para encontrar patrones de comportamiento útiles para la toma de decisiones. En el campo de la investigación aplicada se está trabajando en el descubrimiento de patrones de comportamiento de datos en diferentes áreas de la medicina, actividad conocida como informática biomédica.*
*En este trabajo se presenta el proceso de minería de datos que se lleva a cabo mediante una investigación aplicada al dominio del servicio de salud. Se utiliza como caso de estudio los datos de una empresa prestataria de servicios de salud humana. El propósito es descubrir patrones de comportamiento que generen conocimiento que puede ser útil para la empresa en particular y para el área de salud en general. En la primera etapa del trabajo se propuso como objetivo específico buscar patrones de consumo de medicamentos. Se presentan algunos detalles del proceso realizado y resultados obtenidos.*

**Palabras clave**: Minería, datos, informática, biomédica, consumo, medicamentos, salud.

## 1. Introducción

La globalización y el surgimiento de herramientas y conceptos informáticos como “marketplace”, “streamming” o red social, dio lugar al crecimiento de bases de datos con todo tipo de información. Grandes cantidades de datos de negocios, sociedades, ciencia, ingeniería, medicina y casi todo aspecto de la vida diaria residen y se procesan en las redes de computadoras y dispositivos de almacenamiento. Este explosivo crecimiento es el resultado de la informatización de la sociedad [1] y el fenómeno reconocido por muchos autores, como el impulso que dio lugar a la búsqueda de una nueva forma de analizar grandes cantidades de datos, con el fin de obtener información de ellos para la toma de decisiones.

La informatización de las organizaciones tiene como finalidad primera dar soporte a los procesos de negocio básicos. Una vez satisfecho este aspecto, aparece la necesidad de información para buscar nuevas metas de negocio [2].
Esta necesidad es también consecuencia de dos factores importantes, la incertidumbre y el costo por tomar decisiones incorrectas. Esto dio origen al concepto de economía basada en información y conocimiento [3], cuyo objetivo es estudiar las diferentes situaciones que presentan los datos, para adelantarse a ellas y tomar provecho. Uno de los mayores desafíos de las empresas es mantener una cartera de clientes lucrativa. En este escenario, es necesario adquirir conocimiento que ayude a interpretar las metas, expectativas y deseos de los clientes para poder satisfacerlos de la mejor manera [4].
La investigación y el desarrollo de herramientas para analizar grandes volúmenes de datos se hicieron cada vez más necesarios. Como resultado, surgieron dos metodologías o técnicas relacionadas entre sí, aunque con diferentes objetivos: “data warehousing” y minería de datos.
El “data warehousing” es un proceso de almacenamiento de datos históricos para su posterior análisis mediante reportes y tableros de comando que permite ver los movimientos a través del tiempo de las principales variables de un dominio. El “data warehouse”, si bien no de manera indispensable, se convierte en una buena fuente de datos para realizar minería.
La minería de datos puede realizarse a partir de un simple fichero. No obstante, las ventajas aumentan cuando se cuenta con grandes volúmenes de datos.

La minería de datos puede definirse como el procesamiento de los datos para encontrar patrones de comportamiento que sean de utilidad para la toma de decisiones. Se relaciona de manera estrecha con la estadística, usando técnicas de muestreo y visualización de datos, depuración y cálculo de indicadores. Es importante aclarar que, si bien en ocasiones también se la referencia como descubrimiento de conocimiento desde bases de datos (Knowledge Discovery from Databases) [5] [6], el conocimiento será el que el experto del dominio pueda descubrir a partir de los patrones encontrados. De modo que la parte humana juega un rol clave, por eso se puede definir como un proceso semi-automático [3].

Recibe también un aporte significativo de la inteligencia artificial que contribuye con algoritmos de procesamiento de los datos de tipo heurístico, basados en el modelo de razonamiento humano [7].

Si bien la preparación de los datos y la aplicación de los algoritmos de minería requieren conocimientos de estadística y de informática, la fase de descubrimiento e interpretación de la información recaerá en el experto del dominio.

Si bien los objetivos de aplicar minería sobre las bases de datos comerciales varían con el área de negocio, en general puede decirse que estarán orientados a determinar qué o quién realiza o posee qué cosa bajo qué circunstancia, para luego poder llevar a cabo una acción correctiva o beneficiosa al respecto.

A nivel de investigación básica se está trabajando en el mejoramiento de los algoritmos de minería de datos existentes y en el desarrollo de nuevos algoritmos, en tanto que a nivel de investigación aplicada se está trabajando en el descubrimiento de patrones de comportamiento de datos en diferentes áreas. Por ejemplo, en el área de medicina, actividad conocida como informática biomédica [8] [9] [10] [11] [12], o en el área de genética, actividad conocida como bioinformática.

Este trabajo tiene como objetivo describir el proceso de minería de datos que se lleva a cabo en una base de datos de una empresa prestataria de servicios de salud humana y presentar resultados obtenidos hasta el momento. El propósito de esta investigación aplicada es poder descubrir patrones de comportamiento que generen conocimiento útil para la empresa en particular y para el área de salud en general. La empresa podrá utilizar este conocimiento para tomar decisiones que le permitan mejorar los servicios provistos a sus clientes. En la primera etapa del trabajo se propuso como objetivo específico la búsqueda de patrones de consumo de medicamentos.

En lo que sigue, el trabajo se organiza del siguiente modo: en la Sección 2 se describe brevemente el caso de estudio; en la Sección 3 se describe el proceso de minería de datos llevado a cabo; en la Sección 4 se presenta un resumen de los principales resultados obtenidos; y finalmente, en la Sección 5 se presentan las conclusiones y trabajos futuros.

## 2. Caso de estudio

Se ha tomado como caso de estudio una empresa mutual que brinda servicios de salud. La empresa canaliza la atención de los afiliados a través de convenios con entidades de profesionales de la medicina e instituciones sanitarias como hospitales, sanatorios y clínicas. Tiene cobertura nacional con administración centralizada en Santa Fe. Las prestaciones incluyen prácticas médicas, internaciones, óptica, ortopedia, rehabilitación y farmacia.

El “data warehouse” de la empresa implementado con tecnología SQL Server cuenta con varios millones de registros con datos del plan médico del afiliado, ubicación geográfica, edad, ingresos, situación social y la historia de consumos de prácticas y de medicamentos.

### 2.1 Descripción de los datos

*Planes*: la cobertura del afiliado está reglada por marcos que establecen el porcentaje de descuento y/o monto que se

reconocen al momento de una práctica médica o adquisición de un medicamento. Estos beneficios se agrupan en categorías llamadas planes, los cuales se clasifican en planes para asalariados y planes para autónomos.

*Programas*: bajo ciertas condiciones médicas las personas pueden recibir la cobertura de un programa, que asigna descuentos y excepciones que el plan médico no reconoce. Por ejemplo, el plan materno otorga 100% de descuento en ciertos artículos y/o medicamentos, durante el período de embarazo.

*Códigos*: todo consumo está identificado con un código nomenclador. Para las prácticas médicas se utiliza una adaptación del Nomenclador Nacional. Este es un estándar para la facturación de las prácticas llevadas a cabo por los profesionales de la salud. Los diagnósticos que justifican los consumos están codificados según CIE-10, acrónimo de Clasificación Internacional de Enfermedades. Para los medicamentos se usa un código no estándar.

*Tipos de código*: existen códigos de prestaciones ambulatorias, internaciones, bioquímica, ortopedia, odontología, óptica y medicamentos.

Los códigos de diagnósticos son muy diversos, lo que dificulta un análisis de los mismos. El codificador bajo el cual se encuentran clasificados está dividido en varios capítulos.

Se tiene información sobre la monodroga y acción terapéutica de los medicamentos. Estos datos son muy variados y dan lugar a un gran número de clasificaciones si se los quisiera agrupar por ellos, sobre todo, porque no están codificados siguiendo un estándar y sus descripciones son muy variadas, incluso con caracteres fuera del alfabeto.

## 3. Proceso de minería de datos

Para llevar a cabo la tarea de minería de datos propuesta se tomó como basee el proceso de desarrollo de minería de datos CRISP (Cross-Industry Standard Process for Data Mining) [13]. La decisión se debió fundamentalmente a que es reconocido a nivel internacional como un estándar, no es propietario y es libremente disponible.

Tomando como base CRISP se llevaron a cabo las actividades que se describen en las siguientes secciones.

### 3.1 Relevamiento y muestreo de datos

A efectos de conocer el dominio y poder determinar las variables a incluir en el proceso de minería, plantear interrogantes y seleccionar los datos, se llevaron a cabo las tareas que se describen a continuación.

#### 3.1.1 *Definición de los datos a utilizar*

Para el objetivo específico de la primera etapa del trabajo, buscar patrones de consumo de medicamentos, se utilizaron los datos de consumo de medicamentos y se eligieron como potenciales variables explicativas (discriminantes) el sexo del paciente, la edad, el tipo de medicamento y la estación del año.

#### *3.1.2 Cálculo del tamaño de la muestra*

El tamaño de la muestra se determinó mediante un modelo de selección de muestra para poblaciones de tamaño identificado, $n = z^2\ \sigma^2\ N\ /\ (N\text{-}1)\ e^2 + z^2\ \sigma^2$, donde: $n$ es el tamaño de la muestra, $N$ es el tamaño de la población, $z$ es el nivel de confianza, $\sigma^2$ es la varianza de la población y $e$ es el error de la muestra.

#### *3.1.3 Proceso para seleccionar los datos*

La muestra se eligió tomando como base dos criterios: la probabilidad y la mecánica de selección. En relación al primer criterio se consideró a todos los registros con igual probabilidad de ser seleccionado. Respecto al segundo criterio, se optó por un mecanismo de selección sin reposición, de modo que sólo se tuvo una instancia de cada registro en cada muestra.

#### *3.1.4 Toma de la muestra*

Para tomar la muestra se decidió utilizar un proceso de selección aleatoria de registros desde el “data warehouse” [14]. La palabra aleatoria sugiere seleccionar los registros al

azar, con independencia de cualquier criterio [15]. Para ello se desarrolló un algoritmo basado en la generación de números aleatorios que permite definir los registros a incluir en la muestra. El algoritmo se implementó con el lenguaje Transact-SQL (T-SQL) del motor de bases de datos SQL Server [16].

### 3.2 Preparación de los datos

Los datos fueron acondicionados para garantizar resultados válidos. Esto implicó eliminar los valores incorrectos de los atributos, producidos por error humano, computacional, dato erróneo ingresado debido a campos de entrada obligatorios, entre otros. Las tareas realizadas se describen a continuación.

#### *3.2.1 Depuración de los datos*

Se eliminaron los registros con datos que podían sesgar los resultados de la minería. Para la búsqueda de datos irregulares se utilizaron principalmente métodos gráficos y métodos numéricos basados en la mediana y "cuartiles". Éstos permitieron evidenciar la presencia de agrupamientos y valores atípicos. Para esta tarea se utilizó el lenguaje de base de datos T-SQL.

#### *3.2.2 Transformación de los datos*

La principal transformación se realizó con los valores de la variable Tipo de Medicamento. Esto se debió a que no están codificados de manera estándar, y la clasificación utilizada es excesivamente extensa. Se usó el Sistema de Clasificación Anatómica, Terapéutica, Química ATC (Anatomical, Therapeutic and Chemical classification system) instituido por la Organización Mundial de la Salud. Es un índice de sustancias farmacológicas y medicamentos organizado según grupos terapéuticos (Tabla 1).

Los valores de la variable Edad fueron transformados a una escala de rangos etarios definida por los valores "niño" (hasta 13 años), "joven" (de 14 a 30 años), "adulto" (de 31 a 50 años) y "mayor" (más de 50 años).

Tabla 1: Sistema de Clasificación ATC

| **Clase** | **Descripción** |
|---|---|
| A | Tracto alimentario y metabolismo |
| B | Sangre y órganos formadores de sangre |
| C | Sistema cardiovascular |
| D | Dermatológicos |
| G | Sistema genitourinario y hormonas sexuales |
| H | Preparados hormonales sistémicos, excl. Hormonas sexuales e insulinas |
| J | Antiinfecciosos para uso sistémico |
| L | Agentes antineoplásicos e inmunomoduladores |
| M | Sistema músculo-esquelético |
| N | Sistema nervioso |
| P | Productos antiparasitarios, insecticidas y repelentes |
| R | Sistema respiratorio |
| S | Órganos de los sentidos |
| V | Varios |

Para utilizar algoritmos de segmentación, los datos fueron convertidos a escalas numéricas y normalizados usando el método Z-Score.

### 3.3 Minería propiamente dicha

Para el proceso de minería de datos propiamente dicho se utilizó una muestra de 23919 registros y se realizaron pruebas con varios algoritmos de segmentación ("clustering") con diversas métricas [1]. Los resultados más concluyentes se obtuvieron con el algoritmo K-means basado en el "centroide" usando como métrica la distancia Euclídea [1].

Se utilizó un algoritmo de árbol de clasificación a efectos de generar reglas de clasificación [1]. Estas reglas fueron utilizadas para generar un modelo de clasificación que fue validado con una segunda muestra de datos.

Se utilizó como herramienta RapidMiner, el cual es un sistema de código abierto desarrollado por Rapid-I [17]. Ofrece una programación visual a través de nodos que permiten llevar a cabo diferentes acciones sobre los datos: importación a repositorios propios, limpieza y transformación, y aplicación de múltiples algoritmos de minería.

### 3.4. Descubrimiento de información

Con la participación de expertos del dominio, a partir de los agrupamientos encontrados, se identificaron los patrones. Los principales resultados obtenidos de la primera etapa del trabajo se presentan en la siguiente sección.

## 4. Resultados obtenidos

Los mayores consumos en valores porcentuales corresponden a medicamentos de las clases C, A, N, J y R que explican el 80 % del consumo total de medicamentos (Tabla 2).

Tabla 2: % consumo por clase

| Clase | % |
|---|---|
| C | 21 |
| A | 21 |
| N | 15 |
| J | 15 |
| R | 8 |
| **TOTAL** | **80** |

En la Clase C, medicamentos para el *Sistema cardiovascular*, las variables discriminantes son Sexo y Edad. En la Tabla 3 se pueden observar tres grupos: *Alto Consumo*, que corresponde a Sexo== ”masculino” ∧ Edad==”mayor” con el 48% del consumo; *Mediano Consumo*, que corresponde a Sexo==”femenino” ∧ Edad==”mayor” con el 18% del consumo y a Sexo==”femenino” ∧ Edad==”adulto” con el 13% del consumo; y *Bajo Consumo*, que corresponde a Sexo==”masculino” ∧ Edad==(”adulto” ∨ ”joven” ∨ “niño”) cuyo consumo oscila entre el 8% y el 1%, y a Sexo==”femenino” ∧ Edad==(”joven” ∨ ”niño” ) cuyo consumo oscila entre el 7% y el 1%.

En la Clase A, medicamentos para el *Tracto Alimentario y Metabolismo*, las variables discriminantes son Sexo y Edad.

En la Tabla 4 se pueden observar tres grupos: Alto Consumo, que corresponde a Sexo==”masculino” ∧ Edad=”mayor” con el 38% del consumo; Mediano Consumo, que corresponde a Sexo==”femenino” ∧ Edad=”adulto” con el 21% del consumo y a Sexo==”femenino” ∧ Edad=”mayor” con el 19% del consumo; y Bajo Consumo, que corresponde a Sexo==”masculino” ∧ Edad==(”adulto” ∨ “niño” ∨ ”joven”) cuyo consumo oscila entre el 6% y el 1%, y a Sexo==”femenino” ∧ Edad==(”niño” ∨ ”joven”) cuyo consumo oscila entre el 5% y el 4%.

Tabla 3: Consumo de medicamento Clase C

| GRUPO | Sexo | Edad | % |
|---|---|---|---|
| Alto Consumo | masculino | mayor | 48 |
| Mediano Consumo | femenino | mayor | 18 |
| | femenino | adulto | 13 |
| Bajo Consumo | masculino | niño | 8 |
| | femenino | niño | 7 |
| | masculino | adulto | 4 |
| | femenino | joven | 1 |
| | masculino | joven | 1 |

Tabla 4: Consumo de medicamento Clase A

| GRUPO | Sexo | Edad | % |
|---|---|---|---|
| Alto Consumo | masculino | mayor | 38 |
| Mediano Consumo | femenino | adulto | 21 |
| | femenino | mayor | 19 |
| Bajo Consumo | masculino | adulto | 6 |
| | masculino | niño | 6 |
| | femenino | niño | 5 |
| | femenino | joven | 4 |
| | masculino | joven | 1 |

En la Clase N, medicamentos para el *Sistema Nervioso*, las variables discriminantes son Sexo y Edad. En la Tabla 5 se puede observar que Sexo==”femenino” ∧ Edad==(”adulto” ∨ ”mayor”) explica el 47% del consumo, mientras que Sexo== ”masculino” ∧ Edad=”mayor” explica el 37% del consumo.

Tabla 5: Consumo de medicamento Clase N

| GRUPO | Sexo | Edad | % |
|---|---|---|---|
| Alto Consumo | masculino | mayor | 37 |
| | femenino | adulto | 23 |
| | femenino | mayor | 24 |
| Bajo Consumo | masculino | adulto | 5 |
| | masculino | niño | 3 |
| | femenino | joven | 3 |
| | femenino | niño | 3 |
| | masculino | joven | 2 |

En la Clase J, medicamentos *Anti-infecciosos para uso sistémico*, las variables discriminantes son Sexo, Edad y Estación. En la Tabla 6 se pueden observar tres grupos: *Alto Consumo*, que corresponde a Sexo==”femenino” ∧ Edad==”adulto” con el 25% del consumo y a Sexo==”masculino” ∧ Edad==”mayor” con el 21% del consumo; *Mediano Consumo*, que corresponde a Sexo== (”masculino” ∨ ”femenino” ) ∧ Edad== ”niño” cuyo consumo oscila entre el 16% y el 14% y a Sexo==”femenino” ∧ Edad== ”mayor” con el 11% del consumo; y *Bajo Consumo*, que corresponde a Sexo== ”masculino” ∧ Edad==(”adulto” ∨ ”joven”) cuyo consumo oscila entre el 5% y el 3%, y a Sexo==”femenino” ∧ Edad==”joven” con el 5% del consumo.

Tabla 6: Consumo de medicamento Clase J

| GRUPO | Sexo | Edad | % |
|---|---|---|---|
| Alto Consumo | femenino | adulto | 25 |
| | masculino | mayor | 21 |
| Mediano Consumo | masculino | niño | 16 |
| | femenino | niño | 14 |
| | femenino | mayor | 11 |
| Bajo Consumo | masculino | adulto | 5 |
| | femenino | joven | 5 |
| | masculino | joven | 3 |

La variable Estación también actúa como discriminante en la Clase J. En la Tabla 7 se puede observar la distribución porcentual de consumos en las diferentes estaciones. El mayor valor se registra en Estación==”invierno” con un consumo superior al 50% en relación a la Estación==”verano” donde se registra el menor consumo.

Tabla 7: Distribución del consumo de medicamento Clase J por Estación

| verano | otoño | invierno | primavera |
|---|---|---|---|
| 20% | 25% | 30% | 25% |

En la clase R, medicamentos para el *Sistema Respiratorio*, las variables discriminantes son Sexo, Edad y Estación. En la Tabla 8 se pueden observar tres grupos: *Alto Consumo*, que corresponde a Sexo== ”masculino” ∧ Edad==”mayor” con el 23% del consumo y a Sexo==”femenino” ∧ Edad==”adulto” con el 21% del consumo; *Mediano Consumo*, que corresponde a Sexo==”masculino” ∧ Edad==”niño” y a Sexo==”femenino” ∧ Edad==(”niño” ∨ ”mayor”) cuyo consumo oscila entre el 17% y el 13%; y *Bajo Consumo*, que corresponde a Sexo== ”masculino” ∧ Edad==(”joven” ∨ ”adulto”) y a Sexo==”femenino” ∧ Edad==”joven” cuyo consumo oscila entre el 5% y el 2%.

Tabla 8: Consumo de medicamento Clase R

| GRUPO | Sexo | Edad | Estación | % | % |
|---|---|---|---|---|---|
| Alto Consumo | masculino | mayor | Invierno | 7,2 | 23 |
| | masculino | mayor | Otoño | 5,8 | |
| | masculino | mayor | Primavera | 5,7 | |
| | masculino | mayor | Verano | 4,1 | |
| | femenino | adulto | Invierno | 6,2 | 21 |
| | femenino | adulto | Otoño | 4,9 | |
| | femenino | adulto | Primavera | 5,2 | |
| | femenino | adulto | Verano | 4,1 | |
| Mediano Consumo | masculino | niño | | | 17 |
| | femenino | niño | | | 15 |
| | femenino | mayor | | | 13 |
| Bajo Consumo | masculino | adulto | | | 5 |
| | femenino | joven | | | 4 |
| | masculino | joven | | | 2 |

La variable Estación también actúa como discriminante, con una diferencia de consumo del 122% entre Estación== ”verano”, donde se registra el menor consumo, y Estación==”invierno”, donde se registra el mayor consumo. A modo de ejemplo, en la Tabla 8 se muestra un detalle de valores de esta variable para el grupo *Alto Consumo*. El patrón se repite en los otros dos grupos donde las diferencias son más marcadas.

## 5. Conclusiones

Los resultados obtenidos del proceso de minería de datos, realizado sobre la base de datos de una empresa de servicios de salud, permitieron alcanzar el objetivo propuesto de buscar patrones de consumo de

medicamentos por franja etaria, por sexo y por estaciones del año.
En términos cualitativos, el resultado del proceso de minería no ha producido cambios radicales en el conocimiento del dominio. El principal aporte está dado por la estructuración de dicho conocimiento. Esto proporciona la base para el desarrollo de un modelo de reglas de inferencia que la empresa podrá implementar para dar soporte a procesos de decisión gerenciales.
A futuro se trabajará en la búsqueda de patrones de comportamiento referidos al tratamiento de afecciones tratando de identificar, a partir de los medicamentos o monodrogas suministrados, el tipo de alteraciones de la salud más combatidas según las características del paciente; y afecciones comunes a partir de los diagnósticos informados en los consumos de prácticas médicas. En estos trabajos se analizará también la ubicación geográfica de las personas como potencial variable discriminante.

**Agradecimiento:** a Jerárquicos Salud, Av. Facundo Zuviria 4584, Santa Fe, Argentina.

## Referencias

1. Han, J., M. Kamber, J. Pei. Data Mining - Concepts and Techniques. Morgan Kaufmann, Third Edition (2011).
2. Pérez López, C., D. S. Gonzáles. Minería de datos. Técnicas y Herramientas. Ediciones Paraninfo (2007).
3. Barreiro Fernández, J., J. Diez de Castro, B. Barreiro Fernández. E. Ruzo Sanmartín, F. Losada Pérez (Coords.). Gestión Científica Empresarial: Temas de Investigación Actuales. Netbiblo (2003).
4. Vieira Braga, L., L. Ortiz Valencia, S. Ramirez Carvajal. Introducción a la Minería de Datos. E-papers (2009).
5. Lorose, D. T. Discoveryng knowledge in data: An introduction to data mining. Wiley-Interscience (2005).
6. Oracle Data Mining. Sitio de Oracle: Documentación sobre el producto (2012). http://docs.oracle.com/cd/E16338_01/datamine.112/e16808.pdf.
7. Gorunescu, F. Data Mining -Concepts, Models and Techniques- Intelligent Systems Reference Library, Volume 12. Springer (2011).
8. Bellazzi, R., L. Sacchi, S. Concaro. Methods and tools for mining multivariate temporal data in clinical and biomedical applications. Conference Proceeding IEEE Eng and Medical Biology Society. 2009:5629-5632. doi: 10.1109/IEMBS.2009.5333788 (2009).
9. Rezapour, M., M. Khavanin Zadeh, M. Sepehri. Implementation of Predictive Data Mining Techniques for Identifying Risk Factors of Early AVF Failure in Hemodialysis Patients. Computatinal and Mathematical Methods in Medicine. Jun 4 (2013), doi:10.1155/2013/830745.
10. M. K. Obenshain, Application of Data Mining Techniques to Healthcare Data. Infection Control and Hospital Epidemiology Journal, 25(8):690-695. Aug (2004)
11. Bellazzi R, Diomidous M, Sarkar IN, Takabayashi K, Ziegler A, McCray AT. Data analysis and data mining: current issues in biomedical informatics. Journal Methods of Information in Medicine. 50 (6):536-44. (2011)
12. Venkatadri. M and Lokanatha C. Reddy. A Review on Data mining from Past to the Future International Journal of Computer Applications, 15(7):19-22, Feb (2011)
13. CRISP-DM 1.0. Chapman. P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth. Copyright (1999, 2000). ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf.
14. Martínez Bencardino, C. Estadística básica aplicada. Ecoe Ediciones (2006).
15. Vivanco M. Muestreo Estadístico. Diseño y Aplicaciones. Editorial Universitaria (2005).
16. Referencia de Transact-SQL. Microsoft (2013). http://msdn.microsoft.com/eses/library/ms189826(v=sql.90).aspx.
17. http://rapid-i.com/content/view/181/ (2013).