Descubrimiento de conocimiento en los negocios

JOHANY ARMANDO CARREÑO GAMBOA

LOS EMPRESARIOS FRENTE A LA INFORMACIÓN

El procesamiento de los datos ha sido siempre un tema importante, agradable e interesante para muchos empresarios, aunque lo consideren un gran reto. Generar información novedosa para la toma de decisiones implica arte e ingeniería. La comprensión de los datos y su traducción en modelos eficaces es un proceso artístico. Transformar los datos en información es un proceso de ingeniería[1].

Los datos aumentan exponencialmente cada año, pero la información tiende a decrecer. Por lo tanto, Las organizaciones o empresas deben estar y tener la capacidad de crear y llevar a cabo planes de gestión de los datos (*data management*) y descubrimiento de conocimiento (KDD), con el objeto de realizar consultas de los mismos, generar informes/reportes y más específicamente, llevar a cabo todo el procesamiento para

RESUMEN

Ante la internacionalización de la economía, las organizaciones requieren basarse en la información y el conocimiento, apoyadas en tecnologías de la información y comunicación (TIC), pensar globalmente en políticas integrales y basarse en economías en red bajo esquemas asociativos que las fortalezcan. El creciendo de empresas en los últimos años, hace prioritario tratar de obtener conocimiento útil desde los propios datos y dar un paso más allá en el apoyo a la toma de decisiones más acertada. A tal fin, se ofrece en el documento información básica acerca de la minería de datos, se reconocen sus diferentes etapas y se determina su relación con otras disciplinas. Además se da a conocer el funcionamiento del tipo de algoritmo "árboles de decisión" y, se utiliza la herramienta "Weka" para ajustar modelos a conjuntos de datos.

PALABRAS CLAVE

Dato, información, descubrimiento de conocimiento, minería de datos, toma de decisiones, análisis de decisiones.

ABSTRACT

In view of the internationalization of economy, companies need to rely on information and knowledge. Also, they have to get support from information and communication technologies (TIC), think globally about comprehensive policies, and be based on network economies under associative schemes that strengthen them. The growth of companies in the last years prioritizes the achievement of getting useful knowledge from own data and go beyond when supporting the most appropriate decision making. In order to do this, we offer in the document basic information about data mining, its different stages, and its relationship with other disciplines. Besides, we present the functioning of an algorithm "decision trees" and the use of the tool "Weka" in order to adjust models to data groups.

KEY WORDS

Datum, information, knowledge discovery, data mining, decision making, decision analysis.

1 Turban, E. (2008). *Business intelligence: a managerial approach*. Upper Saddle River, N. J.: Pearson Prentice Hall.

traducir la lógica de los negocios a la lógica de sistemas empresariales y procesar información útil, válida y relevante para tomar decisiones[2].

En el ámbito empresarial todos los días se toman decisiones, pero ¿cuál es el grado de acierto de éstas? Los administrativos conocen como estratégica, táctica y operativamente se comporta cada una de sus organizaciones. Además, saben quiénes son sus empleados; se sienten confiados porque tienen buenos sistemas de información; están seguros porque poseen un buen parque tecnológico; pero la realidad es otra, en el mundo de los negocios actuales hay que estar preparado para descubrir conocimiento.

Las empresas se cierran y muchos culpan a las políticas gubernamentales, sin tener en cuenta que sus decisiones han sido consideradas bajo criterios pobres de extracción, depuración y transformación de los datos. No basta con invertir en nuevas plantas, lanzar nuevos productos, seleccionar e implantar tecnologías de punta, entre otras, sin haber realizado un estudio de las características comunes (reversibilidad, replicación, riesgo, impacto, futuro, etc.) de las decisiones que se toman día a día[3].

> Con todo y lo anterior y a pesar de su importancia, las decisiones no siempre se toman o evalúan utilizando métodos, herramientas y procedimientos apropiados. Aún más preocupante es que, en general, las empresas invierten cantidades considerables de dinero en publicidad y estudios de mercadeo, por ejemplo, y muy poco en métodos y herramientas de soporte a la toma de decisiones[4] .

Para finalizar este aparte, cabe resaltar que según los requerimientos de información y su funcionalidad, existen procesos que permiten el descubrimiento de conocimiento, que pueden ser aplicados en cada uno de los niveles de la organización. Las soluciones de gestión de los datos proporcionan un fácil acceso a los datos críticos dentro de la empresa, necesarios para el análisis, así como un medio para integrar los datos corporativos con los procesos de toma de decisiones estratégicas y tácticas; también permite a la empresa afinar la toma de decisiones cotidiana, asegurando que cada grupo operativo tenga acceso a la información necesaria para contestar preguntas específicas y distribuir dicha información a todos los niveles de la organización.

EL PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO (KDD)

KDD se ha definido como el "proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia, comprensibles a partir de los datos"[5].

Se trata de interpretar grandes cantidades de datos y encontrar relaciones o patrones. Para conseguirlo las

Johany Armando Carreño Gamboa.
Ingeniero de Sistemas de la Universidad Autónoma de Bucaramanga; consultor y asesor en implantación de proyectos de Gestión Tecnológica en empresas nacionales; analista y diseñador de *software* para gestión de datos y toma de decisiones. Docente investigador de tiempo completo adscrito a la Facultad de Ingeniería y Ciencias Básicas del Politécnico Grancolombiano Institución Universitaria. Coordinador de los semilleros de investigación en "Descubrimiento de Conocimiento y Minería de Datos" (http://si.kdmind.googlepages.com/home) y "Diseño y Desarrollo de Aplicaciones para Dispositivos Móviles" (http://sid.mobi.googlepages.com/home). Actualmente, hace parte del grupo de Expertos en Redes NGN Colombia, en calidad de representante del Politécnico Grancolombiano.
Para contactar al autor: jcarreno@poligran.edu.co.

2 Williams, Steve y Williams, Nancy (2006). *The profit impact of business intelligence.* Amsterdam, Boston: Morgan Kaufmann.
3 Castillo Hernández, Mario (2006). *Toma de decisiones en las empresas: entre el arte y la técnica: metodologías, modelos y herramientas.* Bogotá: Universidad de los Andes.
Vitt, E.; Luckevich, M.; y Misner, S. (2003). *Business intelligence: técnicas de análisis para la toma de decisiones. IT/Tecnología y Empresa.* Madrid: McGraw Hill Interamericana.
T. Moss, L., y Atre, S. (2003). *Business intelligence roadmap: The complete project lifecycle for decision-support applications.*Boston, MA: Addison-Wesley Professional.
4 Castillo Hernández, Mario (2006). Op. cit. 15.
5 Fayyad, U.; Piatesky-Shapiro, G. y Smyth, P. (1996). *Advances in knowledge discovery and data mining.* Masachussetts: MIT Press. Pág. 1-34.

investigaciones, en estos temas, incluyen técnicas de aprendizaje (*machine learning*), bases de datos, análisis estadístico de datos, técnicas de representación de conocimiento, razonamiento basado en casos (*case based reasoning* –CBR-), filtrado colaborativo, razonamiento aproximado, adquisición de conocimiento, redes neuronales, visualización de datos, algoritmos genéticos, recuperación de información y computación de altas prestaciones, entre otros. Tareas comunes en KDD son los problemas de clasificación y agrupamiento (*clustering*), el reconocimiento de patrones, estimación/regresión, el modelado predictivo, la detección de dependencias, análisis de *links*, análisis de canastas de mercado e inducción de reglas, entre otras[6].

Cuando se tratan temas de descubrimiento de conocimiento en el ámbito de los negocios, es importante entender que hay que focalizarse en aprender de lo que ha sucedido a través del tiempo y que aprender no puede hacerse de la nada. Por lo tanto, la aplicación de técnicas de KDD requiere la realización de una serie de actividades previas encaminadas a preparar los datos de entrada, debido a que en muchas ocasiones los datos provienen de fuentes heterogéneas, no tienen el formato adecuado o contienen ruido. Por otra parte, es necesario interpretar y evaluar los resultados obtenidos.

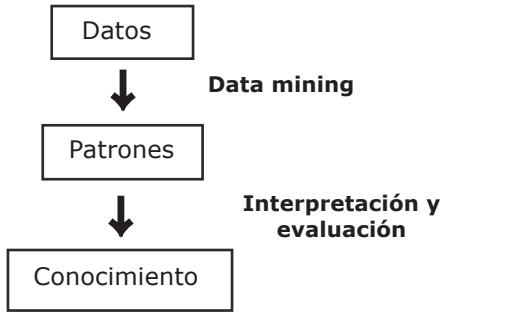
El proceso completo de KDD consta de las siguientes etapas[7] y se puede observar en la Figura 1.

1. Determinar las fuentes de información que pueden ser útiles y dónde conseguirlas. Para llevar a cabo esta tarea es necesario tener dominio del negocio y del problema. También es importante contar con sistemas de procesamiento de transacciones que permitan capturar la interacción de la empresa con sus clientes.

2. Diseñar el esquema de un almacén de datos (*data warehouse*) que almacene de manera operativa toda la información histórica exacta del comportamiento de los clientes, de modo que la aplicación de la minería de datos pueda hacer uso de ésta.

Cabe aclarar, que no es necesario siempre implementar un *data warehouse*, simplemente lo que se requiere es

Figura N°. 1. Proceso de extracción de conocimiento*



**Fuente*: Moreno, María y otros (2002). "Obtención y validación de modelos de estimación de *software* mediante técnicas de minería de datos". *Revista Colombiana de Computación*. Bogotá. Volumen 3, Número 1, p.p. 53–71.

preparar, estructurar y unificar los datos en el formato especializado, requerido e idóneo para la herramienta de minería de datos.

3. Implantación del conjunto de datos que permita la "exploración" y visualización previa de sus datos, para discernir qué aspectos pueden interesar para ser estudiados.

4. Fase de selección, limpieza y transformación de los datos que se van a analizar. En esta fase se eliminan o corrigen los datos incorrectos y se decide la estrategia a seguir con los datos incompletos. Además, con el objetivo de hacer más fácil la tarea propia de minería y para que los resultados de la misma sean más útiles, se proyectan los datos para considerar únicamente aquellas variables o atributos que van a ser relevantes. La selección incluye tanto una criba o fusión horizontal (filas/registros) como vertical, de columnas o atributos.

5. Seleccionar el método de minería de datos apropiado (clasificar, agrupar, etc.) y aplicarlo. La minería de datos tiene como objetivo primordial convertir la historia en planes de acción útiles para el futuro.

6 Berry, Michael J. A. y Linoff, Gordon (2004). *Data mining techniques: for marketing, saes, and customer relationship management.* Segunda edición. Indianapolis: Wiley.
7 Lavrac, N. "Selected techniques for data mining in medicine". En: *Artificial Intelligence in Medicine.* Tecklenburg, West Germany: Burgberlag. Vol. 16 (1), Pág. 3-23, 1999.

6. Interpretación, transformación y representación de los patrones extraídos. En esta etapa se evalúan los patrones y se analizan por los expertos, y si es necesario, se vuelve a las fases anteriores para una nueva iteración. El objetivo primordial es diseñar óptimos planes estratégicos de interacción y administración de relaciones con el cliente.

7. Fase de difusión y uso del nuevo conocimiento. En esta fase los empresarios deben hacer uso del nuevo conocimiento y hacer partícipes de él a todos los posibles usuarios.

Aunque las etapas anteriores se realizan en el orden en que aparecen, el proceso es altamente iterativo, y se establece retroalimentación entre los mismos. Además, no todos los pasos requieren el mismo esfuerzo; generalmente la etapa de preprocesamiento o preparación de los datos es la más costosa ya que representa aproximadamente el 60% del esfuerzo total, mientras que la etapa de minería solo representa el 20%.

Además de las fases descritas, frecuentemente se incluye una fase previa de análisis de las necesidades de la organización y definición del problema[8], en la que se establecen los objetivos de minería de datos. Por ejemplo, un objetivo de negocio de una empresa comercial sería encontrar patrones en los datos que le ayuden a conservar los buenos clientes; para ello, se podría tener varios objetivos de minería de datos: construir un modelo para predecir clientes rentables, y un segundo modelo para identificar los clientes que probablemente dejarán de hacerlo.

Dominios

Los dominios hacen referencia a las actividades del día a día en las organizaciones. Cuando se decide iniciar un proyecto orientado al descubrimiento de conocimiento hay que tener en cuenta que éste es un campo multidisciplinar que se ha desarrollado en paralelo con otras tecnologías. Por ello es importante nutrir los proyectos con bases de datos, técnicas de visualización, sistemas para la toma de decisiones, herramientas de recuperación de información, personal especializado o experto y todas las disciplinas que se requieran para profundizar en el área relacionada con el dominio de estudio.

Los negocios de la distribución y la publicidad dirigida han sido tradicionalmente las áreas en las que más se han empleado los métodos de descubrimiento de conocimiento. Pero éstas no son las únicas áreas a las que se pueden aplicar. De hecho, se pueden encontrar un gran número de dominios: financieros, científicos (medicina, farmacia, astronomía, psicología, etc.), políticas económicas, sanitarias o demográficas, educación, policiales, procesos industriales, turismo, tráfico, deportes, recursos humanos, *web*, entre otros.

A continuación se incluye una lista de ejemplos de algunas de las áreas antes referidas para exponer en qué dominios se pueden usar técnicas de KDD:

Medicina:

•Identificación de patologías. Diagnóstico de enfermedades.
•Detección de pacientes con riesgo de sufrir una patología concreta.
•Gestión hospitalaria y asistencial. Predicciones temporales de los centros asistenciales para el mejor uso de los recursos, consultas, salas y habitaciones.
•Recomendación priorizada de fármacos para una misma patología.

Análisis de mercado, distribución y, en general, comercio:

•Análisis de la cesta de compra (compras conjuntas, secuenciales, ventas cruzadas, señuelos, etc.).
•Evaluación de campañas publicitarias.
•Análisis de la fidelidad de los clientes. Reducción de fuga.
•Segmentación de los clientes.
•Estimación de existencias (*stocks*), de costes, de ventas, etc.

Aplicaciones financieras y de banca:

•Obtención de patrones de uso fraudulento de tarjetas de crédito.
•Determinación del gasto en tarjeta de crédito por grupos.
•Cálculo de correlaciones entre indicadores financieros.
•Identificación de reglas de mercado de valores a partir de históricos.
•Análisis de riesgos en créditos[9].

8 Hernández Orallo, José y otros (2004). *Introducción a la minería de datos*. Madrid: Pearson Educación S.A.
9 Ibíd.

Todos estos ejemplos muestran la gran variedad de aplicaciones donde el uso de técnicas de descubrimiento de conocimiento puede ayudar a entender mejor el entorno en el que se desenvuelve la organización y, en definitiva, la repercusión en calidad de servicio y de disminución costes que pueden ser altamente significativos. Sin embargo, el uso de estas técnicas todavía es reducido, especialmente en Colombia[10].

MINERIA DE DATOS COMO HERRAMIENTA DE APOYO EN LA TOMA DE DECISIONES

El término "minería de datos" está relacionado con la extracción de información relevante (no trivial, implícita, desconocida y con potencial utilidad) en grandes volúmenes de datos. El objetivo de la minería de datos es proporcionar el potencial de análisis necesario para explotar grandes volúmenes de información transaccional con el fin de obtener conocimiento de apoyo en la toma de decisiones (normas, regularidades, patrones, restricciones). Witten, Frank, C. y Boswell (2000) definen la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos[11].

La información transaccional de las organizaciones ha sido recogida a lo largo del tiempo por sistemas automáticos o de forma manual. Información que cuanto más precisa sea, más completo será el registro que proporcione de las interacciones de los distintos subsistemas de la empresa entre sí y de la empresa con el exterior: proveedores, clientes, aseguradoras, entidades promotoras de salud, centros de investigación, hoteles, jugadores, etc.

> La minería de datos es una forma de aprendizaje inductivo, que permite seleccionar las regularidades y reglas más plausibles soportadas por los datos. Los sistemas cognitivos intentan entender su ambiente usando una simplificación del mismo, llamado modelo, que consiste de clases que representan objetos similares en el ambiente y reglas que describen los cambios en el mismo[12].

El proceso de minería de datos implica ajustar modelos o determinar patrones a partir de datos. Este ajuste normalmente es de tipo estadístico, en el sentido en que se permite un cierto ruido o error dentro del modelo.

Los algoritmos de minería de datos realizan, en general, tareas de descripción (de datos y patrones), de predicción (de datos desconocidos) y de segmentación (de datos). Otras, como análisis de dependencias e identificación de anomalías, se pueden utilizar tanto para descripción como para predicción.

Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento de conocimiento[13]. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que denominamos variables objetivo o dependientes, usando otras variables o campos de la base de datos, a las que nos referiremos como variables independientes o predictivas. Por ejemplo, un modelo predictivo sería aquel que permite estimar la demanda de un nuevo producto en función del gasto de publicidad[14]. Esta forma de trabajar se conoce como aprendizaje supervisado y se desarrolla en dos fases: entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos).

Cuando una aplicación no es lo suficientemente madura, no tiene el potencial necesario para una solución predictiva, en ese caso hay que recurrir a los métodos no supervisados que descubren patrones y tendencias en los datos actuales (no utilizan datos históricos). Estos modelos descriptivos, en cambio, identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos. Por ejemplo, una agencia de viajes desea identificar grupos de personas con unos mismos gustos, con el objeto de organizar diferentes ofertas para cada grupo y poder así remitirles esta información; para ello analiza los viajes que han realizado sus clientes e infiere un modelo descriptivo que caracterize a estos grupos[15].

---

10 Ibíd.
11 Witten, I. H. y Frank, E. (1999). *Data mining: practical machine learning tools and techniques with Java implementations*. San Francisco: Morgan Kaufmann.
12 Hernández Orallo, J.; Ramírez Quintana, M. J. y Ferri Ramírez, C. (2004). *Introducción a la minería de datos*. Madrid: Pearson Educación. Pág. 3-18.
13. Witten, H., y Frank, E. Op. cit.
14. Hernández Orallo, José y otros. Op. cit.
15. Ídem.

Agregando a lo anterior, algunos problemas de interés intelectual, económico y de negocios, pueden ser expresados en términos de las siguientes tareas: clasificación, estimación, predicción, reglas de asociación, *clustering*, descripción y optimización de parámetros (*profiling*).

En la Tabla 1 se muestran algunas de las técnicas de minería en ambas categorías[16], que pueden ser utilizadas como estrategias de análisis de datos.

Hay que tener en cuenta una gran variedad de técnicas, combinaciones y nuevas variantes aparecidas recientemente, debido al interés del campo. Los sistemas de KDD incorporan la mayor cantidad de técnicas, así como la heurística para determinar o asesorar al usuario sobre qué método es mejor para distintos problemas. En varios autores (ver bibliografía) se puede encontrar información relevante y de interés sobre cada una de las técnicas de minería de datos.

Tabla N°. 1. Clasificación de las técnicas de minería de datos

Supervisados	No supervisados
Árboles de decisión	Detección de desviaciones
Inducción neuronal	Segmentación
Regresión	*Clustering*
Series temporales	Reglas de asociación
	Patrones secuenciales

Después de todo, a nadie se le escapa que la calidad de un trabajo de minería de datos está por demás condicionada al estado de la información sobre la que dicho trabajo se aplica: datos incompletos o distorsionados sólo producen resultados deficientes, por más sofisticadas que sean las herramientas con que se procesan. La naturaleza inicial de los datos del negocio no es, en este sentido, la más atractiva para un proyecto impecable de minería de datos, pues pueden presentar (y a menudo lo hacen) problemas en su captura, codificación, etc. No obstante, el uso de una metodología como CRISP-DM (CRoss Industry Standard Process for Data Mining) a la que nos referimos en el siguiente aparte hace que, en un momento determinado, la información alcance una "madurez" en términos de procesamiento informático a nuestro parecer suficiente para, ofrecer resultados de calidad.

UNA METODOLOGÍA PARA EL DESCUBRIMIENTO DE CONOCIMIENTO EN GRANDES CONJUNTOS DE DATOS

Existen varias metodologías de trabajo para la elaboración de un proyecto de minería de datos, que especifican diferentes etapas o fases, permitiendo tener una secuencia clara, ordenada y controlada de las mismas y facilitando la planeación y ejecución del proyecto. Un grupo de empresas europeas pioneras en enfrentar problemas de minería de datos son Teradata, SPSS, Daimler-Chrysler y OHRA y propusieron en 1999 la guía de referencia denominada CRISP-DM (Cross-Industry Standard Process for Data Mining)[17].

A grandes rasgos, se describe CRISP-DM una de las más relevantes y que puede ser incorporada, adaptada y especializada por las empresas, de tal manera que los gerentes y directivas en general puedan tener una mejor gestión y análisis de la información.

Dicha metodología consiste en un modelo estándar, no propietario y de libre distribución. Se describe como un modelo jerárquico, consistente en un conjunto de tareas con cuatro niveles de abstracción: 1. fases; 2. para cada fase un conjunto de tareas genéricas; 3. las tareas generan unas situaciones específicas (por ejemplo: limpieza de datos), y por último, 4. instancias o procesos (decisiones y resultados)[18].

Esta metodología, como se muestra en la Figura 2, brinda una ruta clara permitiendo determinar qué actividades desarrollar en qué fase, de tal manera que se logren cumplir los objetivos del proyecto de minería de datos. Las tareas pueden ejecutarse en diferente orden

16 MORENO, María y otros. "Obtención y validación de modelos de estimación de *software* mediante técnicas de minería de datos". Bogotá: *Revista Colombiana de Computación*. Volumen 3, número 1, pág. 53-71, 2002.
17 CHAPMAN, P. y otros (1999).*The crisp-dm process model.* Technical Report, CRISPDM Consortium.
CRISP-DM; "Cross Industry Standard Process for Data Mining", [en línea] <http://www.crisp-dm.org> [Consultado 05 de julio de 2007].
18 Ibíd.

e incluso pueden repetirse al conocer nuevos resultados obtenidos a partir de otras tareas (*backtrack*). Los procesos posteriores se benefician de las experiencias de los anteriores. El aprendizaje durante los procesos puede desencadenar nuevas tareas, más centradas en los objetivos de gestión.

A continuación se hace una breve descripción de las fases de la metodología[19]:

Fase 1. Comprensión del negocio

Incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial, con el fin de convertirlos en objetivos técnicos y en un plan para lograr éstos objetivos.

Fase 2. Comprensión de los datos

Abarca la recolección inicial de datos para identificar la calidad de los mismos estableciendo las relaciones más evidentes y permitiendo tener un acercamiento a las primeras hipótesis.

Fase 3. Preparación de los datos

Incluye las tareas de selección de los datos a los que se les va a aplicar la técnica de modelado (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. Esta fase se encuentra estrechamente relacionada con la fase de modelamiento ya que, de acuerdo con la técnica de modelado que se vaya a utilizar, los datos necesitarán ser procesados y formateados de maneras diferentes.

Fase 4. Modelamiento

En esta fase se seleccionan las técnicas de modelado más apropiadas para el proyecto específico de minería de datos. Dichas técnicas se seleccionan en función de los siguientes criterios:

1. Ser apropiado al problema.
2. Disponer de datos adecuados.
3. Cumplir los requerimientos del problema.
4. Conocimiento de la técnica.

Fase 5. Evaluacion de negocios

En esta fase se procede a la generación y evaluación del modelo, no desde el punto de vista de los datos, sino del cumplimiento adecuado de los objetivos empresariales. Se debe revisar el proceso seguido, teniendo en cuenta los resultados obtenidos, para poder repetir alguno de los pasos en los que se hayan podido cometer errores. Si el modelo generado es válido en función de los objetivos establecidos en la primera fase, se procederá al despliegue del modelo.

Fase 6. Despliegue

La generación y despliegue del modelo no determinan el final del proyecto. Incluso si el objetivo del modelo es "aumentar el conocimiento" de los datos, éstos deberán ser organizados y presentados de manera que la empresa pueda utilizarlos de manera adecuada. Dependiendo de los requisitos, la fase de despliegue puede ser tan simple como la generación de un informe de resultados o tan complejo como la aplicación y revisión nuevamente de todo el proceso de minería de datos.

En suma, CRISP-DM por no estar ligada a ninguna herramienta de software y por ser un estándar de amplia utilización, permite su libre aplicación como gestor de proyectos de minería de datos sobre diferentes herramientas orientadas a la inteligencia de negocios.

Conviene observar, sin embargo, que CRISP-DM no es la única guía que ha sido propuesta. También existen otras apropiadas o abiertas, como la desarrollada por la empresa SAS, denominada SEMMA (*sample*, *explore*, *modify*, m*odel*, *assess*)[20], DMAMC[21] o las cinco aes[22]. Todas estas metodologías, sin embargo, adolecen de métodos o técnicas que permitan tomar adecuadamente los requisitos del proyecto. Más concretamente, aún no existe un proceso maduro que pueda calificarse como una metodología sólida, pues si bien, por ejemplo, CRISP-DM que establece un conjunto de tareas y actividades que deben ser ejecutadas en el proyecto, no establece con qué técnicas o modelos se debe implementar cada

19 Ibíd.
20 Portal www.sas.com. "Descripción de la metodología SEMMA", [en línea] <http://www.sas.com/technologies/analytics/datamining/miner/semma.html> [Consulta: 19 de abril de 2006].
21 Portal www.isixsigma.com. "Consulta sobre metodología 6-Sigma" [en línea] < http://www.isixsigma.com/sixsigma/six_sigma.asp> [Consulta: 23 de junio de 2006].
22 Laudon, K. C. (2002). *Sistemas de información gerencial, organización y tecnología de la empresa conectada en red.* Mexico: Ed. Prentice Hall.

actividad. Por lo anterior, se recomienda que en cada proyecto se adapte una metodología y se complemente con los recursos (*hardware*, *software* y humanos), técnicas y actividades propias de cada tarea.

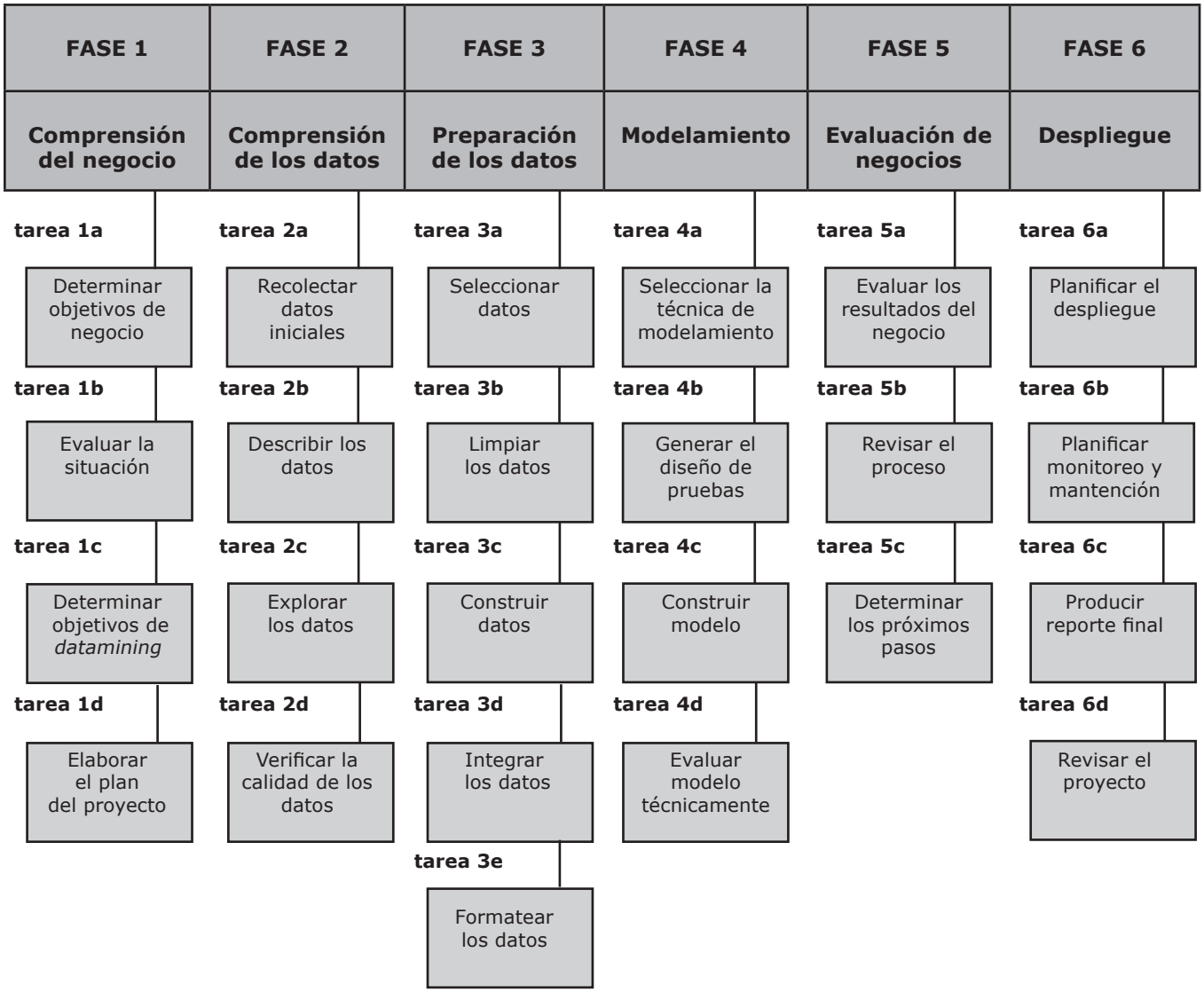
DESCUBRIMIENTO DE CONOCIMIENTO EN DATOS CLÍNICOS

A través de un ejemplo y siguiendo los siguientes numerales, se generaliza el inicio y desarrollo de un proyecto de minería de datos:

La integración de las técnicas de minería de datos en el dominio de la medicina se está convirtiendo en algo habitual.

Otra característica importante es que los usuarios de estos nuevos sistemas son profesionales de la medicina que, aunque tienen ciertos conocimientos de estadística obligatorios en su formación, no tienen conocimientos de aprendizaje de modelos ni de la mayoría de técnicas presentadas en el punto anterior. Por tanto, los sistemas deben ser sencillos de manejar; los patrones descubiertos deben ser fáciles de entender (ya sean simbólicos o

Figura N°. 2. Fases y tareas metodológicas CRISP-DM



visuales) y la interrelación con el resto de sistemas informáticos de adquisición de datos, visualización y gestión de los centros asistenciales, debe ser transparente para el usuario.

Al llegar a este punto los nuevos sistemas de una manera cómoda y eficaz deben permitir[23]:

1. Asociación de síntomas y clasificación diferencial de patologías.
2. Estudio de factores (genéticos, precedentes, hábitos alimenticios, etc.) de riesgo/salud en distintas patologías.
3. Segmentación de pacientes para una atención más inteligente según su grupo.
4. Predicciones temporales de los centros asistenciales para mejor uso de los recursos, consultas, salas y habitaciones.
5. Estudios epidemiológicos, análisis de rendimientos de campañas de información, prevención, sustitución de fármacos, entre otros.

Finalmente, el objetivo primordial de aplicar técnicas de minería de datos es la adquisición, el descubrimiento y el mantenimiento de gran parte del conocimiento de una manera automática. La inclusión manual del mismo a partir de expertos u otras fuentes debería minimizarse, siempre que hubiera alternativa automática. Para ello, como se ha visto anteriormente, surge el KDD.

Los nuevos objetivos del KDD en medicina son[24]:

1. La interpretación comprehensiva de los datos de los pacientes de una manera contextual y la presentación de tales interpretaciones de una manera visual o simbólica.
2. La extracción (descubrimiento) de conocimiento médico a partir del diagnóstico, revisiones médicas, seguimientos, terapias o tareas globales de gestión de los pacientes.

Dentro de la extracción de conocimiento se incluye el uso de conocimiento previo, ya sea para su refinamiento o particularización, o para ayudar al descubrimiento de nuevos patrones o modelos.

A. Metodología

El ejercicio consiste en realizar un tipo de análisis de detección de patrones de comportamiento asociados al dominio de hepatitis, sobre un conjunto de datos con un número de 20 atributos incluido el atributo clase o etiquetado y 155 instancias de datos nominales y numéricos normales recolectados por la Carnegie-Mellon University. Cabe aclarar que las instancias poseen valores perdidos de algunos atributos.

Antes de seguir adelante es importante nombrar algunas herramientas de *software* comercial existentes en el mercado, tales como: SPSS Clementine, Salford CART/MARS/TreeNet/R, SAS, Angoss, KXEN, entre otras, y algunas herramientas de *software* libre como Yale, Weka, R, Orange y Knime. El *software* utilizado para el análisis de los datos clínicos en este documento es WEKA (Waikato Enviroment for Knowledge Analisis), que contiene una biblioteca de clases de aprendizaje en Java, con interfaces gráficas muy sencillas de utilizar y con las que se pueden aplicar y evaluar un gran número de algoritmos de minería a grandes conjuntos de datos. En *http://www.cs.waikato.ac.nz/ml/weka/*[25] se puede encontrar toda la documentación necesaria y relevante sobre esta herramienta.

El formato de los datos en WEKA es como se describe en un archivo con extensión ARFF (Atribute-Relation File Format) donde se especifican los atributos y los datos, los cuales pueden tomar valores nominales y numéricos[26].

Por otra parte, también se encuentran disponibles en internet un gran número de archivos ARFF para los investigadores que estén interesados en el estudio de datos médicos. Estos conjuntos de datos son el resultado de evaluaciones de datos clínicos realizadas por la Universidad de Waikato y la Universidad de California[27].

Como estrategias de análisis de descubrimiento de conocimiento en datos médicos, se pueden utilizar métodos estadísticos, métodos basados en conocimiento (sistemas expertos) y métodos de aprendizaje automático supervisado (redes neuronales, reglas de asocia-

23 Hernández, O. José y otros (2004). Op. cit.
24 Lavrac, N. Op. cit.
25 Machine Learning Project. "Minería de Datos" 2006. [en línea] <http://www.cs.waikato.ac.nz/ml/weka/> [Consultado 02 de septiembre 2006].
26 Portal <http://www.cs.waikato.ac.nz/ml/weka/>, "Consulta sobre la herramienta WEKA, archivos ARFF, algoritmos y técnicas de minería de datos" [en línea], <http://www.cs.waikato.ac.nz/~ml/publications.html> [Consulta: 23 de junio de 2006].
27 Ibíd.

ción, sistemas inmunes, árboles de decisión) y no supervisado (aprendizaje bayesiano, vecino más cercano, *fuzzy logic*, *support vector machines* y redes neuronales). Para el ejemplo descrito en el modelo de estudio se escogieron árboles de decisión, que son utilizados fundamentalmente para clasificación; además es uno de los más sencillos y fáciles de implementar y a su vez, de los más poderosos[28].

Resumiendo y para entender mejor lo descrito en el modelo de estudio, a continuación se describe brevemente la inducción de árboles de decisión:

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas. Una de las ventajas de los árboles de decisión es que, en su forma más general, las opciones posibles a partir de una determinada condición, son excluyentes. Esto permite analizar una situación y, siguiendo el árbol de decisión apropiadamente, llegar a una sola acción o decisión. Estos algoritmos se llaman algoritmos de partición o algoritmos de "divide y vencerás".

Un árbol de decisión toma de entrada un objeto o situación descrita por un conjunto de atributos y regresa una decisión de verdadero / falso. Cada nodo interno corresponde a una prueba en el valor de los atributos y las ramas están etiquetadas con los posibles valores de la prueba. Cada hoja especifica el valor de la clase.

Los árboles de decisión están limitados a hablar de un solo objeto; es decir, son esencialmente proposicionales, siendo cada prueba de atributo una proposición. Por lo tanto, cualquier función booleana (función matemática cuyas variables son binarias y están unidas mediante los operadores del álgebra de Boole suma lógica (+), producto lógico (•) o negación (')) puede ser descrita por un árbol de decisión.

Para muchas funciones, los árboles son relativamente pequeños. Sin embargo, para otras funciones se puede requerir de un árbol exponencialmente grande. Para n atributos, hay $2n$ filas. La salida se puede considerar como una función definida por $2n$ bits. Con esto hay ($2^{2 \wedge n}$) posibles funciones diferentes para n atributos (para 6 atributos, hay 2×10^{19}).

Cuando se realiza inducción de árboles de decisión a partir de ejemplos, un ejemplo es descrito por los valores de los atributos y el valor del predicado meta. Al valor del predicado meta se le llama "la clasificación del ejemplo". Si el predicado es verdadero, entonces el ejemplo es positivo, si no, el ejemplo es negativo. En caso de existir más clases, los ejemplos de una sola clase son positivos y el resto de los ejemplos se consideran negativos.

Para elegir qué atributos y en qué orden aparecen en el árbol, se utiliza una función de evaluación: ganancia de entropía. La entropía es la cantidad de bits, en promedio, que harían falta para codificar los ejemplos; es equivalente a la medida de la "cantidad de información" representada. Si se tienen p ejemplos positivos y n ejemplos negativos, entonces la entropía o medida de la "cantidad de información" representada en el conjunto (S) es:

$$Et\ (S) = -p\log_2(p) - n\log_2(n) \quad (1)$$

Caso binario donde: p y n = proporciones de ejemplos positivos/negativos.

Un atributo normalmente no proporciona toda esta información, pero se puede estimar viendo cuánta información se necesita después de utilizarlo; cada atributo A divide los ejemplos del conjunto de entrenamiento en subconjuntos E_1, E_2,⋯, E_v de acuerdo con los valores del mismo.

Cada subconjunto E_i tiene p_i ejemplos positivos y n_i negativos, por lo que para cada rama se necesita:

$$I\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right) \quad (2)$$

La fórmula (2) describe la cantidad de información para responder a una pregunta.

28 Ibíd.

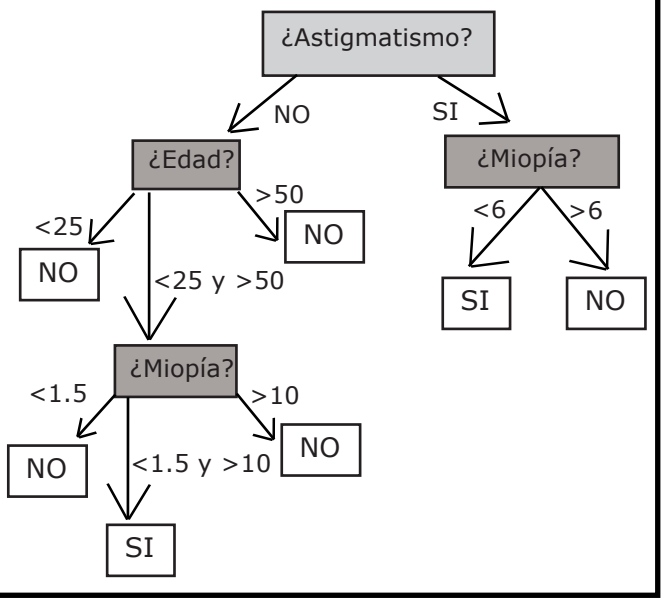
Si todos los ejemplos son positivos o negativos, por ejemplo, pertenecen todos a la misma clase, la entropía será 0. Una posible interpretación de esto, es considerar la entropía como una medida de ruido o desorden en los ejemplos. Se define la ganancia de información como la reducción de la entropía causada por particionar un conjunto de entrenamiento S, con respecto a un atributo A:

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} \, I\left(\frac{p_i}{p_i + n_i} , \frac{n_i}{p_i + n_i} \right) \quad (3)$$

La cantidad de información que se gana al seleccionar un atributo está dada por:

$$Ganancia\ (A) = I\left(\frac{p}{p+n} , \frac{n}{p+n} \right) - E(A) \quad (4)$$

Figura N°. 3. Árbol de decisión para determinar recomendación de cirujía ocular



La ganancia de A dice el número de bits que se ahorran para responder a la pregunta clase de un ejemplo, dado que se conoce el valor del atributo A.

Dicho de otra forma, mide qué tan bien un atributo separa los ejemplos de entrenamiento de acuerdo con la clase meta o etiquetada.

La función evaluadora escoge el atributo con mayor ganancia de entropía en cada iteración.

Por ejemplo[29], un hospital público en el que se realizan operaciones de cirugía refractiva (LASIK) a los miopes que lo soliciten. Evidentemente, dichas operaciones no están indicadas en muchos casos, y algunos podrían ser excluidos en una primera fase, con el objetivo de evitar riesgos potenciales o efectos secundarios. Aunque la indicación o no de dicha cirugía requiere un examen minucioso por parte del servicio de oftalmología del hospital, existen algunas condiciones claras por las cuales se puede determinar si, en principio, una persona está indicada para el estudio detallado (tensión ocular y paquimetría) y, finalmente, para la operación. En la figura 3 se muestra un ejemplo del árbol de decisión que se utiliza para admitir solicitudes.

Como se puede observar en la figura, es sencillo aplicar el árbol de decisión a un nuevo paciente para saber si se le ha de recomendar o no para dicha operación. Basta realizar las preguntas y seguir las respuestas hasta alguna de las hojas del árbol, catalogadas con un "no" o un "sí". Este árbol de decisión en concreto funciona como un "clasificador"; es decir, dado un nuevo individuo, lo clasifica en una de las dos clases posibles: "no" o "sí".

Los árboles de decisión se pueden expresar como un conjunto de reglas, con respecto al ejemplo de la cirugía refractiva. En las siguientes líneas se muestra un ejemplo del árbol de decisión anterior expresado en forma de reglas:

¿Operación?

1. Sí astigmatismo = No y 25 < edad < = 50 y 1.5<miopía<=10 Entonces sí
2. Sí astigmatismo=sí y miopía<=6 entonces sí
3. En otro caso no

29 Hernández, O. José y otros (2004). Op. cit.

La representación en forma de reglas suele ser, en general, más concisa que la de los árboles, ya que permite incluir condiciones y permite el uso de reglas por defecto, como la que comienza por "en otro caso", en el ejemplo anterior.

Para terminar, cabe recalcar que uno de los aspectos más importantes en los sistemas de aprendizaje de árboles de decisión es el denominado criterio de partición, ya que una mala elección de la partición (especialmente en las partes superiores del árbol) generará un peor árbol.

B. Modelo de estudio

A continuación se hace una breve descripción del trabajo realizado en el área de análisis de datos clínicos a través de técnicas y del uso de herramientas de minería de datos. La limitación de espacio con la que se cuenta obliga a dedicar estas líneas sólo a las aportaciones más relevantes.

Una de las formas de usar WEKA es aplicar un método de aprendizaje (clasificador), inducción de árboles de decisión al conjunto de datos hepatitis.arff y analizar la salida para extraer información sobre los datos. Lo primero que se debe hacer es estructurar los datos y normalizarlos, con el objeto de hacer la carga del archivo en el *software* WEKA. Los datos estructurados y normalizados para utilizar en el ejercicio son (atributo: instancias o valores que puede tomar el atributo en un momento determinado)(ver columna siguiente):

Antes de continuar se sugiere que el lector esté familiarizado con el dominio que se evalúa aquí en este aparte, para tal fin, se recomienda revisar: Hepatitis - Guías y Revisiones de la Fundación Ginebrina para la Formación y la Investigación Médica, Falla Hepática Aguda, Hepatitis en Urgencias, Encefalopatía Hepática y Health-Cares.net, que son temas, documentos y enlaces que enriquecen el conocimiento en cuanto a la evaluación y manejo de la hepatitis, así como presentan una gran cantidad de lecturas recomendadas al respecto[30].

1. Class: die, live
2. Age: 10, 20, 30, 40, 50, 60, 70, 80
3. Sex: male, female
4. Steroid: no, yes
5. Antivirals: no, yes
6. Fatigue: no, yes
7. Malaise: no, yes
8. Anorexia: no, yes
9. Liver big: no, yes
10. Liver firm: no, yes
11. Spleen palpable: no, yes
12. Spiders: no, yes
13. Ascites: no, yes
14. Varices: no, yes
15. Bilirubin: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00
16. Alk phosphate: 33, 80, 120, 160, 200, 250
17. Sgot: 13, 100, 200, 300, 400, 500
18. Albumin: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19. Protime: 10, 20, 30, 40, 50, 60, 70, 80, 90
20. Histology: no, yes

Weka permite la aplicación de diversos algoritmos de clasificación, y para este ejercicio se prestó especial atención en el algoritmo J4.8, que presenta los mejores resultados de precisión, cobertura y coste computacional para el ejemplo descrito en este aparte. También se ha escogido este algoritmo por ser quizás el más representativo y uno de los algoritmos más utilizados en la práctica[31].

La herramienta Weka también permite aplicar cuatro tipos de pruebas distintas:

Use training set: mide la calidad del clasificador para predecir la clase de las instancias en las que ha sido entrenado. Útil cuando se tienen pocas muestras en el conjunto.

Supplied test set: evalúa la calidad del clasificador para predecir la clase de un conjunto de instancias cargadas desde un archivo.

Cross-validation: evalúa la calidad del clasificador mediante validación cruzada, usando el número de grupos que se especifiquen. Este tipo de prueba es la utilizada para la evaluación del caso de estudio.

30 Valcarcel Asencios, Violeta (2004). *Data mining y el descubrimiento del conocimiento*. Lima: Universidad Nacional Mayor de San Marcos. Cios, K. J.(2001). *Medical data mining and knowledge discovery*. New York: Physica-Verlag Heidelberg.
31 Portal <http://www.cs.waikato.ac.nz/ml/weka/>. Op. cit.

Percentage split: evalúa la calidad del clasificador según lo bien que clasifique un porcentaje de los datos que se reservan para "*test*"[32].

Realizado el corrimiento del clasificador J4.8 se observa la siguiente informació[33]:

Run Information: muestra el esquema utilizado para tratar los datos (el nombre de la clase empleada y los parámetros usados), el número de instancias, una lista de los atributos presentes y el modo de "test" para las cuatro anteriores.

=== Run information ===

Scheme: weka.classifiers.trees.j48 -c 0.25 -m 2

Relation: hepatitis

Instances: 155

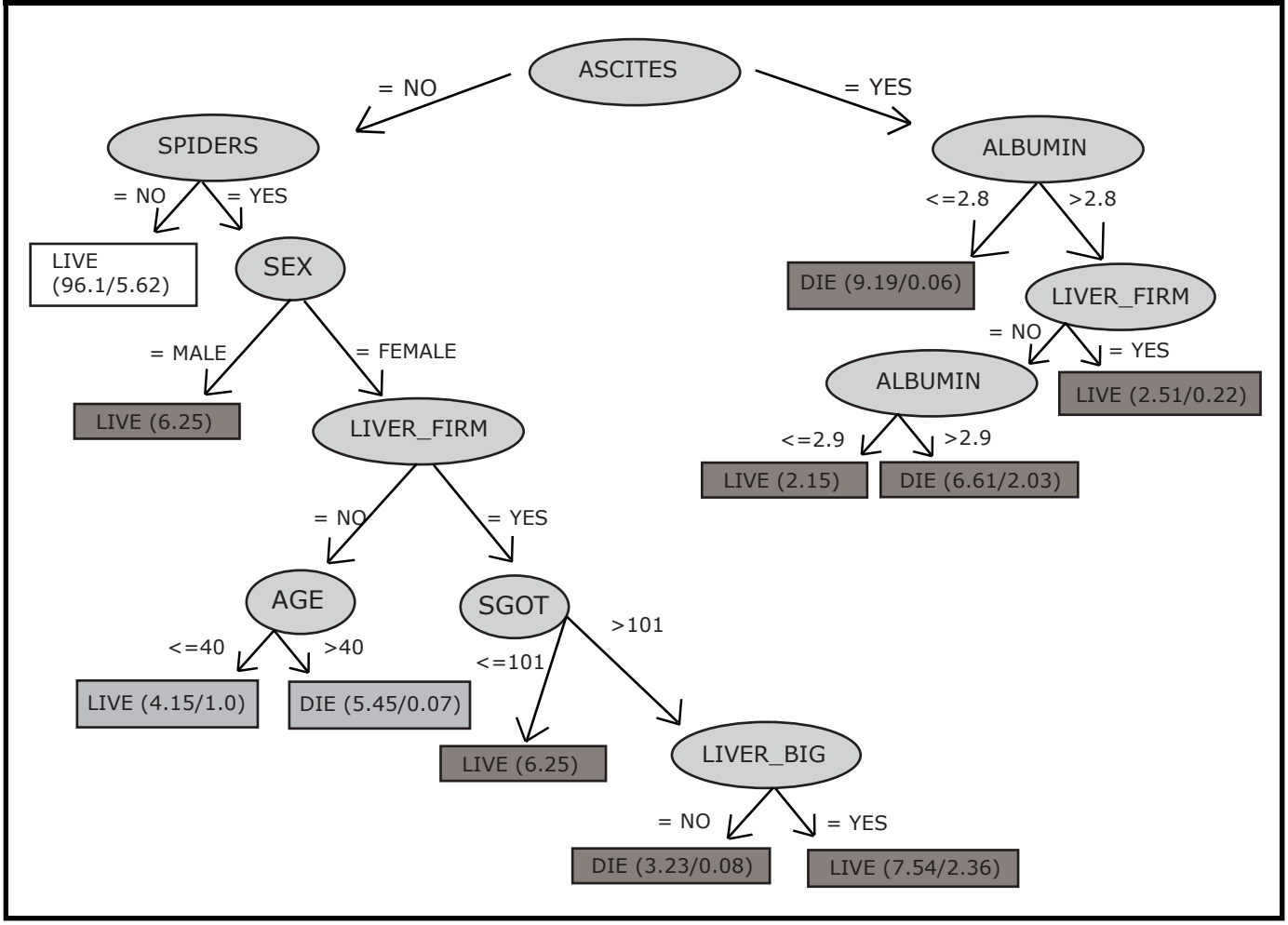
Attributes: 20

[list of attributes omitted]

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Figura N°. 4. Árbol de decisión para determinar esperanza de vida en pacientes con hepatitis



32 Witten, I. H., & Frank, E. (1999). Op. cit.
33 Machine Learning Project. Op. cit.

Descripción del modelo del clasificador: depende del algoritmo utilizado; aquí se presenta la estructura del árbol de clasificación resultante (ver Figura 4). En cada hoja y rama se especifica el criterio de división, y en las hojas finales aparece para la clase que se especifica el número de casos correctamente clasificados y el número de casos mal clasificados. Si todos los datos pertenecen a la clase correcta, solo aparece dicho número. Las ramas que aparecen son aquellas que clasifican el mayor número de casos con el menor error posible (menor número de casos erróneos).

En este caso, el primer atributo es "ASCITES" porque es el que produce la división de los datos con entropía mínima en ese nivel, y análogamente, con el resto. Para clasificar un ejemplo nuevo se sigue el árbol de arriba abajo; la hoja final es la categoría inferida. Los caminos, desde la raíz hasta los nodos hoja, se pueden ver como reglas, donde el antecedente está formado por la intersección de los pares atributo-valor de los caminos.

Para todos los casos se han realizado validaciones cruzadas para crear los conjuntos de pruebas y entrenamiento, tomando los datos y dividiéndolos de forma aleatoria en diez subconjuntos mutuamente excluyentes del mismo tamaño aproximadamente, y utilizando nueve de ellos para la inducción del modelo y uno para la prueba. Cuando estamos en el primer tipo de prueba se hace una evaluación de los datos, *evaluation on training set* y *stratified cross-validation* en el resto. El proceso de inducción se repite diez veces, de manera que en cada iteración se elige un subconjunto distinto como conjunto de prueba, utilizando los restantes para el entrenamiento. Posteriormente se han realizado, también de forma aleatoria, particiones diferentes y se ha vuelto a inducir el modelo de forma iterativa.

A continuación se presentan las estadísticas: porcentaje de instancias clasificadas correctamente (precisión), estadísticas kappa (mide lo que se ajusta la predicción a la clase real, 1.0 significa ajuste total), error medio, error cuadrático medio, error relativo y error cuadrático relativo (más útiles cuando se hace regresión que en clasificación). Es necesario recalcar que Kappa es siempre menor o igual a 1. Un valor de 1 implica que la predicción se ajusta a la clase real; un valor menor de 1 representa un pobre ajuste. Una posible interpretación de la estadística Kappa puede ser[34]:

- Ajuste pobre = Menor de 0.20
- Acuerdo justo = 0.20 a 0.40
- Ajuste moderado = 0.40 a 0.60
- Buen ajuste = 0.60 a 0.80
- Muy de acuerdo = 0.80 a 1.00

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	130	83.871 %
Incorrectly Classified Instances	25	16.129 %
Kappa statistic	0.436	
Mean absolute error	0.2029	
Root mean squared error	0.363	
Relative absolute error	61.4384 %	
Root relative squared error	89.6358 %	
Total Number of Instances	155	

C. Evaluación

Los procesos y medidas de evaluación son los mismos para todos los experimentos: dada la colección de datos, una parte de la misma es considerada como conjunto de entrenamiento y el resto como conjunto de examen. Así, los modelos aprenden del conjunto de entrenamiento y tratan de inferir las categorías de los ejemplos del conjunto de prueba. Puesto que las categorías de éstos últimos son conocidas, se pueden validar las inferencias de los modelos. Así pues, esta validación se realiza para cada categoría mediante tres medidas típicas: precisión, cobertura y medida-F[35].

Los parámetros de exactitud para cada clase son los siguientes:

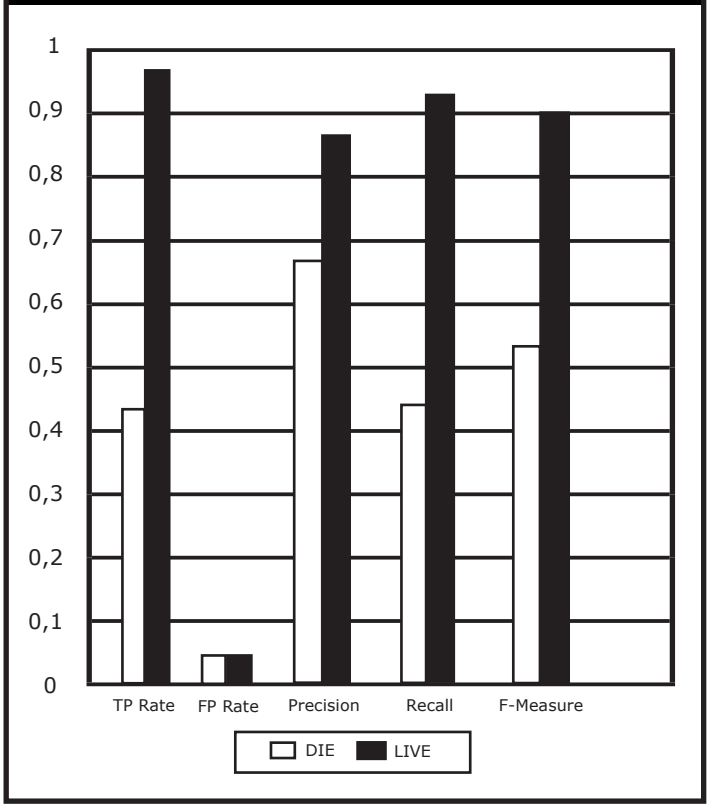
=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.438	0.057	0.667	0.438	0.528	DIE
0.943	0.563	0.866	0.943	0.903	LIVE

34 Hernández, O. José y otros (2004).

35 Machine Learning Project. Op. cit.

Figura N°. 5. Gráfica de comparación de medidas de exactitud para las clases "die" y "live"



A continuación se presentan los detalles de exactitud (ver Figura 5) para cada clase (*detailed accuracy* y *class*) y finalmente se procede a realizar la evaluación del modelo inducido.

True positive (TP) rate: es la proporción de ejemplos que fueron clasificados como clase x ("die", "live"), de entre todos los ejemplos que realmente tienen clase x; es decir, la cantidad de la clase que ha sido capturada. En la matriz de confusión, es el valor del elemento de la diagonal dividido por la suma de la fila relevante.

False positive (FP) rate: es la proporción de ejemplos que fueron clasificados como clase x, pero en realidad pertenecen a otra clase de entre todos los ejemplos que no tienen clase x. En la matriz de confusión, es la suma de la columna menos el valor del elemento de la diagonal dividido por la suma de las filas de las otras clases.

Precision: precisión es la proporción de ejemplos que de veras tienen clase x entre todos los que fueron clasificados como tal. En otras palabras el indicador de precisión es el resultado que se presenta en forma de clasificación o estimación, medido a través del porcentaje de predicciones que son correctas. Cuando se habla de clasificación se emplea el porcentaje de casos bien clasificados, y para la estimación se emplea el porcentaje de registros con una estimación que el clasificador considere correcta. En la matriz de confusión, la precisión es el elemento de la diagonal dividido por la suma de la columna relevante.

Recall: cobertura es el resultado de dividir el número de ejemplos recuperados que son relevantes, sobre el total de elementos que son relevantes. Es el coeficiente que mide el porcentaje de registros en los cuales se puede aplicar una regla. Se corresponde con el denominador que se ha empleado para el cálculo del coeficiente de confianza y/o precisión. Por ejemplo, si en el conjunto de datos que se ha utilizado en el desarrollo de este ejercicio se dispone de 15 registros, de los cuales el atributo "albumin": albúmina se presenta en 10 ocasiones con un valor menor de 2.8, que es precisamente la condición impuesta por la regla "*if (albumin <= 2.8) then die*". Así, el coeficiente de cobertura se sitúa para dicha regla en el 0.66 (10/15).

F-measure: es simplemente el resultado de la ecuación:

$$\frac{2*Precision*Recall}{Precision + Recall} \qquad (5)$$

Una medida combinada de "precision" y "recall". Representa, en cierto modo, la intersección entre los ejemplos implicados en la precisión y la cobertura, normalizada mediante la suma de ambas.

D. Validación del modelo inducido

Para realizar la validación del modelo inducido con todos los datos de entrenamiento, se aplica la técnica de

las matrices de confusión, o también la tabla de contingencia que muestran el tipo de las predicciones correctas e incorrectas cuando se aplica el modelo sobre el conjunto de prueba. La matriz de confusión ayuda a conocer la bondad del modelo para predecir y también para descubrir en qué casos se producen errores[36].

=== Confusion Matrix ===

a	*b*	<-- Classified as
14	*18*	*a = DIE*
7	*116*	*b = LIVE*

Una tabla de contingencia está formada por tantas filas y columnas como clases hay. El número de instancias clasificadas correctamente es la suma de la diagonal principal de la matriz; el resto están clasificadas de forma incorrecta e indican el tipo de error cometido (qué valor ha predicho el modelo y cuál es el valor verdadero). En el ejemplo que se está tratando no se clasifican correctamente todos los registros, aunque el peso de las clasificaciones incorrectas es muy pequeño; además, éstas se encuentran próximas a la diagonal, lo que indica que el error cometido no es demasiado elevado.

Como se ha visto, la información que se da aquí y en el apartado anterior es la misma expresada de otra forma. Por tanto, se presentaron solamente los resultados de la matriz de confusión con 130 instancias bien clasificadas equivalentes al 83.871% del número total de instancias del modelo, ya que comprobando el número de elementos no nulos (25 instancias mal clasificadas equivalentes al 16.129% del número total de instancias, y que se pueden visualizar sumando los valores que se encuentran en la diagonal secundaria de la matriz de confusión,) se puede demostrar que J4.8 es un buen algoritmo para este caso. El número de personas que en este análisis tienen posibilidades de vivir es de 116 y de morir, un total de 14.

¿LIVE?
1. **IF** (ASCITES =no) **THEN** LIVE
2. **IF** (ASCITES=no) **AND** (SPIDERS=no) **THEN** LIVE
3. **IF** (ASCITES=no) **AND** (SPIDERS=yes) **AND** (SEX=MALE) **THEN** LIVE
4. **IF** (ASCITES=no) **AND** (SPIDERS=yes) **AND** (SEX=FEMALE) **AND** (LIVER_FIRM=no) AND (AGE<=40) THEN LIVE
5. **IF** (ASCITES=no) **AND** (SPIDERS=yes) **AND** (SEX=FEMALE) **AND** (LIVER_FIRM=yes) AND (SGOT<=101) **THEN** LIVE
6. **IF** (ASCITES=no) **AND** (SPIDERS=yes) **AND** (SEX=FEMALE) **AND** (LIVER_FIRM=yes) **AND** (SGOT>101) **AND** (LIVER_BIG=yes) **THEN** LIVE
7. **IF** (ASCITES=yes) **AND** (2.8<ALBUMIN<=2.9) **THEN** LIVE
8. **IF** (ASCITES=yes) **AND** (ALBUMIN>2.9) **AND** (PROTIME>52) **THEN** LIFE
9. **IN ANOTHER CASE** DIE

E. Interpretación

Los patrones asociados al comportamiento de la hepatitis se pueden representar mediante reglas, como se muestra en el siguiente recuadro.

Interpretando el árbol de decisión en el modelo de estudio y las reglas descritas anteriormente, lo que resalta, desde luego, es que el mayor problema asociado a las personas que fueron diagnosticadas es la ascitis "ascites", que es una de las complicaciones de la cirrosis y consiste en la acumulación de líquido en la cavidad abdominal. La presencia de ascitis en una persona enferma del hígado se considera una indicación de trasplante hepático[37]. Las causas de ascitis son muy variadas, desde infecciones hasta insuficiencia cardiaca. Sin embargo, la causa más frecuente es la cirrosis hepática. Asociados a la ascitis existen otros atributos que pueden complicar o no la valoración del paciente, como se puede observar si un paciente presenta cuadro de ascitis y ausencia de proteínas específicas (albúmina: "albumin" y protrombina: "protime") o enzimas ("sgot": enzima glutamato oxaloacetato deshidrogenada; un aumento de la actividad de esta enzima es indicador de muerte celular o daño severo en el hígado) para meta-

36 Cabena, P. y otros (1997). Op. cit.
37 Chen, H. (2005). *Medical informatics: knowledge management and data mining in biomedicine. Integrated series in information systems.* New York: Springer.
Aseervatham, S. y Osmani A. *Mining short sequential patterns for hepatitis type detection.* Université de Paris-Nord, Laboratoire LIPN-CNRS UMR 7030F-93430 Villetaneuse Cedex, France: ECML/PKDD Discovery Challenge, 2005.

bolizar diferentes substancias en el hígado, las posibilidades que se tienen de sobrevivir son muy pocas.
En el conjunto de datos la mayoría de los pacientes son diagnosticados entre los 40 y los 60 años y se presenta un comportamiento asociado muy interesante como lo es el atributo "spiders" (arañas vasculares: dilatación de capilares sanguíneos a modo de patas de araña), que como caso típico se encuentra fuertemente coligado al género femenino, porque aunque no es exclusivo de las mujeres, sí es cierto que a ellas, las pequeñas lesiones en las venas suelen afectarlas más que a los hombres.

Evidentemente, muchas personas desarrollan síntomas tales como pérdida de apetito, malestar general, fiebre, náuseas y vómitos, dolor en los músculos, fatiga, dolor de cabeza, agrandamiento del hígado, etc., dependiendo de la enfermedad, antes de ser diagnosticados y asistidos por especialistas. Como aporte a la medicina, las técnicas de minería de datos, ayudan a los médicos en la detección de patrones de comportamiento de enfermedades, infecciones, posibles complicaciones, entre otros, que son relevantes en el momento de tomar una decisión inherente al tratamiento de los mismos[38].

Para finalizar, es necesario resaltar que el número de veces que se ejecutan los procesos y por el cual está enmarcada la métrica, debe ser determinado por un grupo de especialistas en medicina. Estas métricas son diseñadas dependiendo de los criterios relevantes y asociados a los patrones de comportamiento de la gran variedad de enfermedades existentes actualmente, así como hay que tener en cuenta las actividades desarrolladas por el organismo de salud y la infraestructura tecnológica, entre otros.

CONCLUSIONES

La aplicación de técnicas de minería de datos en las empresas ofrece una excelente oportunidad a las directivas en general de tener una visión amplia del negocio para hacer proyecciones y descubrimientos de patrones en el comportamiento de sus dominios, y, en definitiva, garantizar más competitividad y lograr así su permanencia y crecimiento en el mercado. Además existe un buen número de herramientas de despliegue de los modelos de minería basados en *software* libre, lo cual rompe con la limitación de muchas de ellas en cuanto a la inversión de programas costosos, con requerimientos e infraestructura que muchas veces están fuera de su alcance.

Las técnicas estadísticas son fundamentales a la hora de validar hipótesis y analizar datos, por lo cual la estadística desempeña un papel importante en KDD. Asimismo, destacamos que, en particular, los sistemas de KDD pretenden automatizar el proceso completo de análisis de datos y descubrir modelos que permitan generar nuevas posibilidades de descubrimiento de conocimiento en grandes conjuntos de datos.

Con la aplicación de técnicas de descubrimiento de conocimiento, los estudios realizados y validados pueden ser enfocados a la mejora de cualquier organización de la que se pueda extraer un gran número de datos estructurados.

Queda pendiente de momento emprender más estudios reales y verídicos en conjuntos de datos aportados por los sistemas transaccionales de diferentes organizaciones que mediante la aplicación de las diferentes técnicas de minería de datos generen resultados predictivos útiles para prever el comportamiento futuro de algún tipo de comportamiento o dominio, con el único objeto de mejorar la calidad de los sistemas organizacionales. Este documento ofrece información general sobre el descubrimiento de patrones de comportamiento asociados al dominio de la Medicina, específicamente a un conjunto de datos de hepatitis. No quiere esto decir que puede ser considerado un consultor, y se sugiere comprobar siempre con un especialista cuando se tenga una pregunta sobre la forma de gestionar y administrar la información.

Se pretende motivar a los administradores y expertos del comercio a usar herramientas especializadas de descubrimiento de conocimiento a partir de conjuntos de datos estructurados de sus sistemas, para apoyar las tareas de toma de decisiones. Además de inducirlos a que se vayan familiarizando con las nuevas herramientas y sistemas que, bien serán cada día más potentes y posibles de manejar.

38 Pizzi, L. C., Ribeiro, M. X., Vieira, M. T. P. (2005). *Analysis of hepatitis dataset using multirelational association rules.* Department of Computer Science, Federal University of São Carlos, São Carlos, Brazil: ECML/PKDD Discovery Challenge, 2005.

BIBLIOGRAFÍA

• TURBAN, E. (2008). *Business intelligence: A ma-nagerial approach*. Upper Saddle River, N.J.: Pearson Prentice Hall.

• WILLIAMS, S., y WILLIAMS, N. (2007). *The profit impact of business intelligence*. Amsterdam: Elsevier/ Morgan Kaufmann.

• CASTILLO HERNÁNDEZ, Mario (2006). *Toma de decisiones en las empresas: entre el arte y la técnica: metodologías, modelos y herramientas*. Bogotá: Ediciones Uniandes.

• VITT, E.; LUCKEVICH, M. y MISNER, S. (2003). *Business intelligence: técnicas de análisis para la toma de decisiones*. IT/Tecnología y Empresa. Madrid: McGraw-Hill Interamericana.

• MOSS, L. T. y ATRE, S. (2003). *Business intelligence roadmap: the complete project lifecycle for decision-support applications*. Boston, MA: Addison-Wesley.

• FAYYAD, U., PIATESKY-SHAPIRO, G. y SMYTH, P. (1996). *Advances in knowledge discovery and data mining*. Masachussetts: MIT Press.

• BERRY, M. J. A., y Linoff, G. (2004). *Data mining techniques for marketing, sales, and customer relationship management*. Indianapolis: Wiley.

• LAVRAC, N. "Selected techniques for data mining in medicine". *Artificial Intelligence in Medicine*. Vol. 16 (1), pp. 3-23, 1999.

• MORENO, María y otros. "Obtención y validación de modelos de estimación de software mediante técnicas de minería de datos". Bogotá: *Revista Colombiana de Computación*. Volumen 3, número 1, pág. 53-71, 2002.

• HERNÁNDEZ, O. José y otros (2004). *Introducción a la minería de datos*. Madrid: Pearson Educación S.A.

• WITTEN, I. H. y FRANK, E. (1999). *Data mining: Practical machine learning tools and techniques with Java implementations*. San Francisco: Morgan Kaufmann.

• WITTEN, I. H., & FRANK, E. (2005) *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann, Segunda edicion.

• PÉREZ, César y SANTÍN, Daniel (2006). *Data mi-ning: soluciones con enterprise miner*. Madrid: Alfaomega Ra-Ma.

• CHAPMAN, P. y otros (1999).*The crisp-dm process model*. Technical Report, CRISPDM Consortium.

• CRISP-DM; "Cross Industry Standard Process for Data Mining", [en línea] <http://www.crisp-dm.org> [Consultado 05 de julio de 2007].

• Portal www.sas.com, "Descripción de la metodología SEMMA", [en línea] <http://www.sas.com/technologies/analytics/datamining/miner/semma.html> [Consulta: 19 de abril de 2006].

• Portal www.isixsigma.com, "consulta sobre metodología 6-Sigma" [en línea] < http://www.isixsigma.com/sixsigma/six_sigma.asp> [Consulta: 23 de junio de 2006].

• LAUDON, K. C. (2002). *Sistemas de información gerencial, organización y tecnología de la empresa conectada en red*. Mexico: Ed. Prentice Hall.

• Machine Learning Project. "Minería de Datos" 2006. [en línea] <http://www.cs.waikato.ac.nz/ml/weka/> [Consultado 02 de septiembre 2006].

• Portal http://www.cs.waikato.ac.nz/ml/weka/, "Consulta sobre la herramienta WEKA, archivos ARFF, algoritmos y técnicas de minería de datos" [en línea], <http://www.cs.waikato.ac.nz/~ml/publications.html> [Consulta: 23 de junio de 2006].

• CABENA, P. y otros (1997). *Discovering data mining: from concept to implementation*. New York: Editorial Prentice Hall, Upper Saddle River.

• VALCARCEL ASENCIOS, Violeta (2004). *Data mining y el descubrimiento del conocimiento*. Lima: Universidad Nacional Mayor de San Marcos.

• CIOS, K. J.(2001). *Medical data mining and knowledge discovery*. New York: Physica-Verlag Heidelberg. Two Crows Corporation. (1999). *Introduction to data mining and knowledge discovery*. Potomac, MD: Two Crows Corp.

• CHEN, H. (2005). *Medical informatics: knowledge management and data mining in biomedicine. Integrated series in information systems*. New York: Springer.

• ASEERVATHAM, S. y OSMANI A. *Mining short sequential patterns for hepatitis type detection*. Université de Paris-Nord, Laboratoire LIPN-CNRS UMR 7030F-93430 Villetaneuse Cedex, France: ECML/PKDD Discovery Challenge, 2005.

• PIZZI, L. C.; RIBEIRO, M. X. yVIEIRA, M. T. P. *Analysis of hepatitis dataset using multirelational association rules.* Department of Computer Science, Federal University of São Carlos, São Carlos, Brazil: ECML/PKDD Discovery Challenge, 2005.