Universidad de La Habana.
Facultad de Matemática Computación.

ESTRATEGIA DE TRABAJO PARA EL DESARROLLO DEL MÓDULO DE MINERÍA DE DATOS DE UN CALL CENTER, APLICANDO LA METODOLOGÍA CRISP-DM.

TESIS DE GRADO PARA OPTAR POR EL TITULO DE LICENCIADO EN CIENCIAS DE LA COMPUTACION.

Autores:

Isidro Manuel Corría Ramírez
Ronald Shelton Nadal

Tutores:

Lic. Gabriel Zerquera Guerra.
Ing. Gladys Nieto Zurbano.
Ing. Yamilet Sosa Remón.

Cubatel s.a.
Universidad de las Ciencias Informáticas.
Ciudad de La Habana. Cuba
Junio, 2004

RESUMEN

La necesidad de la informatización de la sociedad en la actualidad ha impulsado a la búsqueda de alternativas que permitan satisfacer esta gran demanda de parte de las empresas y la población. La integración de tecnologías de telecomunicaciones e informática, indiscutiblemente propone una solución elegante y eficiente. Una Plataforma Call Center (CC), la cual permita el acceso a la información, vía telefónica, correo electrónico y navegación Web, no cabe duda que cumpliría con estas expectativas.

Teniendo en cuenta que el Call Center sería un punto de intersección entre los usuarios y proveedores de productos y servicios, es fácil percatarse de la cantidad de información sobre mercadeo que este puede almacenar. Para brindar un mejor servicio y darle un aprovechamiento económico a esta información, se pretende agregar un modulo de Minería de Datos al CC, de manera que se le extraiga información muy valiosa (vetas de oro) a estas montañas de datos; muy bien vista por los proveedores.

La complejidad de la búsqueda de conocimiento en bases de datos (KDD) ha hecho que surjan metodologías de trabajo, que en cierta medida intentan un estándar para este proceso. La selección de la **CRISP-DM** como metodología, que nos permite un planeamiento de una estrategia de trabajo para la implementación del módulo de minería de datos del CC, dándole cumplimiento al objetivo fundamental del presente proyecto. Para lograr esto se desplegó cada una de las fases de esta metodología extrapolada a nuestro problema en específico. Como se mencionó anteriormente la búsqueda de conocimiento, en este caso está dirigida a una mejora de ofertas, importante para proveedores. Esta condición hace que los objetivos de la minería se encaminen a responder los intereses de estos. Para ello se formularon los objetivos apoyándose en la Gestión de Relación con Clientes (**CRM**).

ABSTRACT

The necessity of information of the society, at the present time it has impelled to the search of alternatives that allow to satisfy this great demand on behalf of the companies and the population. The integration of technologies of telecommunications and computer science, unquestionably propose an elegant and efficient solution. A Call Center Platform (CC), which allows the access to the information, via phone, e-mail and Web, doesn't fit doubt that it fulfilled these expectations.

Keeping in mind that the Call Center would be an intersection point between the users and suppliers of products and services, it is easy to notice of the quantity of information it has more than enough marketing that this it can store. To offer a better service and to give an economic use to this information it is sought to add an I modulate from Data Mining to the CC, so that is extracted very valuable information (veto of gold) to these mountains of data; very well seen by the suppliers.

The complexity of the Knowledge Discovered in Databases (KDD) has made that work methodologies arise that attempts a standard for this process in certain measure. The selection of the CRISP-DM like methodology that it allows us a planning of a work strategy for the implementation of the module of data mining of the CC, giving execution to the fundamental objective of the present project. To achieve this each one of the phases of this methodology spread extrapolated to our problem in specific. As it was mentioned the search of knowledge previously, in this case it is directed to an improvement of offers, important for suppliers. This condition makes that the objectives of the mining head to respond the interests of these. For they were formulated it the objectives leaning on the Customer Relation Management (CRM).

INDICE

Resumen II
Abstract III
Indice IV
Introducción ..... - 1 -
1 Capítulo 1. Estado del Arte. ..... - 5 -
1.1 Introducción ..... - 5 -
1.2 ¿Qué es Minería de Datos? ..... - 5 -
1.2.1 Definición. ..... - 6 -
1.2.2 Cronología. ..... - 9 -
1.2.3 Fases en un proceso clásico Minería de Datos. ..... - 11 -
1.2.3.1 Definición del alcance y objetivos. ..... - 11 -
1.2.3.2 Selección de datos relevantes. ..... - 13 -
1.2.3.3 Preprocesado y Limpieza de datos. ..... - 13 -
1.2.3.4 Uso de los Algoritmos de Minería de Datos. ..... - 16 -
1.2.3.5 Interpretación de los resultados. ..... - 18 -
1.2.4 Herramientas de Minería de datos. ..... - 19 -
1.2.5 Aplicaciones de la Minería de Datos y Tendencias. ..... - 21 -
1.2.6 Dificultades en la aplicación de la MD. ..... - 23 -
1.3 Metodologías de aplicación de MD ..... - 24 -
1.3.1 Metodología CRISP-DM ..... - 24 -
1.3.1.1 Contexto del Proyecto ..... - 25 -
1.3.1.2 Proyección. ..... - 26 -
1.3.1.3 Como proyectar. ..... - 26 -
1.3.2 Metodología SEMMA. ..... - 28 -
1.3.2.1 Muestreo. ..... - 29 -
1.3.2.2 Exploración. ..... - 30 -
1.3.2.3 Manipulación. ..... - 30 -
1.3.2.4 Modelización. ..... - 31 -
1.3.3 Metodología CRITIKAL. ..... - 31 -
1.3.4 Metodología de las "5 A'S" ..... - 31 -
1.3.5 Metodología de DM: Conclusiones ..... - 32 -
2 Capítulo 2. Aplicación de CRISP-DM. ..... - 33 -
2.1 Introducción ..... - 33 -
2.2 Comprensión del problema (Fase I CRISP-MD) ..... - 33 -
2.2.1 Objetivos del negocio ..... - 33 -
2.2.1.1 Antecedentes ..... - 33 -
2.2.1.2 Determinación de los Objetivos del negocio ..... - 34 -
2.2.2 Evaluar la situación ..... - 35 -
2.2.2.1 Recursos Disponibles ..... - 35 -
2.2.2.2 Fuentes de Datos ..... - 35 -
2.2.2.3 Requerimientos, suposiciones y restricciones: ..... - 36 -
2.2.2.4 Riesgos y contingencias ..... - 36 -

2.2.2.5 Costes y beneficios ..... - 36 -
2.2.3 Determinar las metas de Minería de Datos ..... - 37 -
2.2.3.1 Metas de minería de datos. ..... - 37 -
2.2.3.2 Criterios de éxito (perspectiva de minería de datos) ..... - 38 -
2.2.4 Producir un plan de proyecto ..... - 38 -
2.2.4.1 Planificación y técnicas previstas ..... - 38 -
2.3 Comprensión de los datos (Fase II CRISP-MD) ..... - 40 -
2.3.1 Conseguir el conjunto inicial de datos. ..... - 41 -
2.3.1.1 Informe inicial sobre los datos ..... - 41 -
2.3.1.2 Se definen los siguientes pasos: ..... - 41 -
2.3.2 Describir los datos ..... - 41 -
2.3.2.1 Informe con la descripción de los datos ..... - 41 -
2.3.3 Explorar los datos ..... - 42 -
2.3.4 Verificar la Calidad de los datos ..... - 43 -
2.3.5 Comprensión de datos ..... - 43 -
2.3.5.1 Selección de las fuentes ..... - 43 -
2.3.5.2 Estudio de los datos ..... - 44 -
2.3.5.3 Establecer el tipo de las variables: ..... - 44 -
2.3.5.4 Establecer la caducidad de cada dato (vida de las variables). ..... - 44 -
2.4 Preparación de los datos (Fase III CRISP-MD) ..... - 45 -
2.4.1 Introducción ..... - 45 -
2.4.2 Selección de los Datos ..... - 46 -
2.4.3 Limpieza de los Datos. ..... - 47 -
2.4.3.1 Valores nulos. ..... - 48 -
2.4.4 Generación de variables adicionales. ..... - 49 -
2.4.5 Integración de Orígenes de Datos. ..... - 49 -
2.4.6 Cambio de Formato en los Datos. ..... - 50 -
2.5 Modelado (Fase IV CRISP-MD) ..... - 50 -
2.5.1 Selección de la técnica de modelado ..... - 51 -
2.5.2 Diseño del Método de Evaluación ..... - 51 -
2.5.3 Generar un diseño de prueba ..... - 51 -
2.5.4 Evaluar el modelo ..... - 51 -
2.6 Análisis del descubrimiento directo. ..... - 52 -
2.6.1 Descubrimiento directo (predictivo) ..... - 52 -
2.6.2 Procesos del descubrimiento directo. ..... - 53 -
2.6.2.1 Clasificación y Estimación ..... - 54 -
2.6.2.2 Predicción de valores ..... - 63 -
2.7 Análisis del descubrimiento indirecto. ..... - 68 -
2.7.1 Descubrimiento indirecto (Descriptivo): ..... - 68 -
2.7.2 El proceso del descubrimiento indirecto. ..... - 69 -
2.7.2.1 Segmentación de bases de datos ..... - 70 -
2.7.2.2 Análisis de Asociaciones. ..... - 77 -
2.8 Evaluación (Fase V CRISP-MD) ..... - 81 -
2.8.1 Fases y Salidas ..... - 81 -
2.9 Implantación (Fase VI CRISP-MD) ..... - 82 -
2.9.1 Fases y salidas Implantación. ..... - 82 -

Conclusiones ..... - 83 -
Recomendaciones ..... - 84 -
Bibliografía ..... - 85 -
Anexos ..... - 90 -
Glosario ..... - 93 -

INTRODUCCIÓN

En la actualidad existe una férrea voluntad del Gobierno por lograr la Informatización de la Sociedad Cubana. A tales efectos se han dado pasos importantes para materializar estos aspectos. Por mencionar algunos:

- La creación y extensión de los Joven Club de Computación en todos los municipios.
- La creación de Laboratorios de Computación en todas las Escuelas del país desde la enseñanza primaria.
- La creación de la Universidad de Ciencias de la Información (UCI).
- La creación de la Oficina Nacional de Informatización de la Sociedad (ONIS).
- La celebración del Evento Informática con periodicidad anual.
- La digitalización telefónica y la conectividad del país.

Estos son algunos aspectos que han permitido lograr una infraestructura tecnológica adecuada para la utilización masiva e intensiva de la teleinformática en todas las esferas de la Sociedad y la formación de los recursos humanos capaces de utilizar la informática de forma eficiente en función del desarrollo socio-económico del país.

Por otra parte, un elemento fundamental para la elevación de la satisfacción de los clientes es la utilización de las nuevas tecnologías en función de la sociedad y la introducción de nuevos servicios teleinformáticos. El desarrollo empresarial y las necesidades en la vida cotidiana de cada persona han sido fieles demandantes de este proceso revolucionario. A pesar de los grandes avances obtenidos en el desarrollo de esta tarea, a la media poblacional le resulta más asequible un teléfono que un ordenador online. De aquí la necesidad de búsqueda de vías alternativas utilizando este medio de más fácil acceso entre todos nosotros.

Luego no nos cabe duda alguna que lograr la informatización vía telefónica sería una buena opción. Valorando lo antes expuesto y haciendo uso de las tecnologías que brinda la telefonía en la

actualidad, el desarrollo de un Centro de Atención a Llamadas o **Call Center** (CC) como se le suele llamar, resolvería el problema fundamental del "acceso a la información".

CUBATEL S.A., Sociedad Cubana para las Telecomunicaciones, tiene por objeto la proyección, diseño, ingeniería, producción, instalación, montaje, supervisión, consultoría, capacitación, activación, configuración, puesta en marcha, garantía y asistencia técnica de sistemas, redes y/o equipos de telecomunicaciones e informáticos de todo tipo que estén asociados a las comunicaciones, incluyendo el software vinculado a los mismos, entre otros servicios. Para lo cual también podrá realizar por sí misma la importación de los sistemas, equipos, componentes, accesorios, piezas, materias primas, insumos y software necesarios así como la exportación de los mismos. Por otra parte, es de destacar que CUBATEL S.A. comercializa Sistemas CC incluyendo los diferentes elementos que los componen, tales como ACD, IVR, CTI, entre otros.

Actualmente existe una alianza estratégica entre CUBATEL S.A. y la Universidad de Ciencias Informáticas (UCI), para el desarrollo de productos informáticos propios de la cartera de negocios de CUBATEL S.A., para lo cual está prevista la participación de grupos multidisciplinarios conformados por especialistas de CUBATEL S.A., diplomantes de especialidades afines y estudiantes de la UCI, así como el uso de las capacidades técnicas de ambas entidades.

Bajo esta alianza se desarrolla un proyecto para poner en funcionamiento un CC en el país que sea accedido tanto por personas naturales como jurídicas, con una plataforma teleinformática que pueda ser un Portal Web. Este proyecto permitirá que el cliente pueda conocer la existencia de productos y servicios con que cuente una empresa, en remoto, ya sea a través de llamada a un CC donde reciba atención por los agentes, a través de navegación por el portal o por e-mail. Una perspectiva interesante para el futuro sería añadir un servicio que prepare un paquete con los productos solicitados por el cliente (Televenta).

Se darán estadísticas de solicitudes que permitirán a las empresas conocer los índices de consumo de los diferentes productos y servicios, tanto por tipo de cliente, precios, etc. Como por zonas geográficas de la ciudad. Todo esto dado a que el mayor valor agregado que proporciona un CC, bien equipado, es registrar la historia de los contactos, independiente de la vía de comunicación, potenciando una mejor atención a sus clientes.

Además se construirá otro módulo de vital importancia, principalmente para los proveedores de servicios y productos, Gestión de Relación con Clientes (Customer Relation Management, CRM). Módulo que mirándole desde una perspectiva de negocio y a tono con la actualidad mundial podríamos llamarle un Software de CRM Analítico.

No es difícil percatarse de la cantidad de información que puede acopiarse en un CC que relaciona a los usuarios con los proveedores. Las dimensiones de la base de datos y sus velocidades de crecimiento, hacen muy difícil para un humano su análisis y la extracción de alguna que otra información importante. Estos datos con frecuencia contienen valiosa información que puede resultar muy útil y ser vista como vetas de oro por los ojos de los proveedores.

Aún con el uso de herramientas estadísticas clásicas esta tarea es casi imposible. El Descubrimiento de Conocimiento en Base de Datos (*Knowledge Discovered in Databases*, KDD) combina las técnicas tradicionales con numerosos recursos desarrollados en el área de la inteligencia artificial, aunque dentro de este proceso la Minería de Datos es solo una fase, este término "Minería de Datos" MD(*Data Mining, DM*) ha tenido mucha más aceptación.

Es entonces cuando, dadas las condiciones se decide diseñar e implementar utilizando técnicas de MD un módulo que le permita al proveedor obtener información para desarrollar un perfecto proceso de CRM.

Partiendo que en el KDD el peso mayor del proceso radica en las acciones derivadas del descubrimiento de conocimiento, no en el mecanismo de descubrimiento en si. Aunque no se puede obviar el papel fundamental de los algoritmos, la solución es más que un conjunto de técnicas y herramientas; es preciso aplicarlas en el caso correcto y a los datos correctos. Considerando lo antes expuesto no cabe duda que para el descubrimiento de conocimiento en bases de datos es de suma importancia seguir un patrón que oriente y estructure las fases del proceso haciendo viables los objetivos que se tracen.

Cumpliendo con este requerimiento se utilizara el Proceso Estándar Industrial para Minería de Datos (*CRoss Industry Standard Process for Data Mining,* CRISP-DM).

El CRISP-DM es un estándar industrial utilizado por más de 160 empresas e instituciones de todo el mundo, que surge en respuesta a la falta de estandarización y propone un modelo de proceso general para proyectos de minería de datos:

- Neutral respecto a industria y herramientas.
- Aplicable en cualquier sector de negocio.

En el presente trabajo se pretende organizar una estrategia para la implementación del módulo de Minería de Datos del CC, ver los objetivos y requerimientos del proyecto desde una perspectiva de negocio. Para ello se plantean los objetivos:

1. Análisis de las Metodologías usadas para resolver problemas de Minería de Datos.
2. Estudio en detalle de cada una de las fases de la metodología CRISP-DM para el CC.
3. Planeamiento de una estrategia de trabajo para la confección del modulo de Minería de Datos del CC, sobre las premisas de la metodología CRISP-DM.

Como objetivo general se pretende desarrollar el estándar industrial para minería de datos CRISP-DM en el modulo de minería de datos del CC.

La estructura de este trabajo está conformada por una introducción y dos capítulos. El primer capítulo "Estado del Arte" se ofrece una revisión del concepto del *Data Mining*, las metodologías y técnicas más comunes, organizadas de una nueva forma según las últimas tendencias. Se describe brevemente, la metodología escogida: CRISP-DM [11].

En el segundo capítulo desarrolla cada una de las fases de la metodología, hasta el punto que nos permite el estado de implementación en el que se encuentra el CC. Finalmente conclusiones así como recomendaciones.

1 CAPÍTULO 1. ESTADO DEL ARTE.

1.1 INTRODUCCIÓN

Gracias al desarrollo de las comunicaciones, la implementación y mejora de las redes informáticas; cada vez es más la cantidad de información que fluye en las empresas y la capacidad de acceso a la misma ha aumentado considerablemente, sin embargo, cada vez se tiene menos tiempo y capacidad para asimilarla y analizarla. Se ha estimado que cada 20 meses se duplica la cantidad de información en el mundo [42]. Muchas veces, las empresas no saben obtener información valiosa de la cantidad colosal de datos que tienen almacenados, a pesar de intuir que el conocimiento que se podría extraer de ellos sería de gran ayuda en muchas de las áreas y facetas en que se desenvuelven (toma de decisiones, mejoras en la producción, mejor conocimiento de los gustos del cliente, etc.). Se estima que solo entre el 5% y 10% de las bases de datos comerciales han sido analizadas [19]. La competencia entre empresas, y por lo tanto su estabilidad, depende de que este conocimiento pueda salvaguardarse y utilizarse de forma eficaz.

Evidentemente, este interés solo aparece cuando la empresa tiene un volumen de históricos realmente importante del proceso y una cultura de mejora arraigada. Las herramientas de Minería de Datos y estadística multivariante son útiles en este momento, cuando ya tenemos un volumen de información importante y de buena calidad. Los campos de aplicación de estas nuevas técnicas dentro de la informática son numerosos: obtención de reglas y patrones de comportamiento, búsqueda de causas y relaciones entre variables, los de mayor relevancia en nuestro caso.

1.2 ¿QUÉ ES MINERÍA DE DATOS?

Como se comentaba anteriormente, para que las empresas trabajen eficientemente en su competitividad, se necesita contar con el mínimo de información necesaria y presentarla de la forma más fácil de dilucidar y manipular. La Gestión del Conocimiento (*Knowledge Discovery*, KD) abarca todas aquellas tecnologías relativamente nuevas que surgen de esta necesidad de procesar, analizar y aprovechar esta información encubierta en grandes volúmenes de datos. Esta capacidad de captación, estructuración y transmisión de conocimiento es lo que se requiere de este conjunto de

tecnologías. Definitivamente, la Gestión del Conocimiento es una ciencia de joven aparición que al desarrollar múltiples herramientas, permite tratar datos garantizando la obtención de información útil de la forma lo más eficiente posible a partir de los mismos.

Dentro de las múltiples áreas que se agrupan alrededor a la gestión del conocimiento, surge la Minería de Datos o Data Mining como una de las disciplinas que más influyen en nuestros días dentro del medio del análisis de datos. A grandes rasgos se puede decir que la minería de datos es un conjunto de metodologías y herramientas que permiten extraer el conocimiento útil (patrones de comportamiento, modos de operación, información útil para descubrir fallos, tendencias, etc.) para la ayuda en la toma de decisiones, comprensión y mejora de procesos o sistemas, etc.; partiendo de grandes cantidades de datos.

Para alcanzar el éxito es necesario advertir que la minería de datos no se basa en una metodología estándar y genérica que resuelve todo tipo de problemas, sino que gravita en una metodología dinámica e iterativa que va a estribar del problema planteado, de la disponibilidad de la fuentes de datos, del conocimiento de las herramientas precisas y de los requerimientos y recursos de la empresa.

Claramente se identifican numerosos campos de aplicación de estas nuevas técnicas: control, optimización y supervisión de procesos industriales, control de calidad, modelado e identificación de sistemas, obtención de tendencias económicas, correlación entre indicadores financieros, diagnóstico de enfermedades, determinación de los efectos de un medicamento, clasificación de señales biomédicas, predicción de ventas, planificación de campañas publicitarias, gestión de relaciones con clientes, detección de fraudes, detección de evasión de impuestos, hallazgos de patrones de comportamientos criminales, etc.

1.2.1 Definición.

El nombre de Minería de Datos o Data Mining, se deriva de la semejanza que se encuentra entre buscar valiosa información de negocios en grandes bases de datos y la búsqueda de vetas de metales preciosos en una montaña. Ambos procesos requieren inspeccionar inteligentemente una monstruosidad de material hasta encontrar algo que pueda resultar útil.

La definición del concepto de Minería de Datos (MD) puede cambiar entre unos estudiosos y otros. Por ejemplo, los estadísticos, analistas de datos y la comunidad de sistemas de gestión de la información adoptan mayoritariamente este término para referirse al **proceso genérico correspondiente a las técnicas y herramientas de investigación usadas para extraer información útil de una base de datos.** Dentro de estas técnicas podemos suponer todos aquellos métodos matemáticos y software para el análisis inteligente de los datos y búsqueda de patrones o tendencias en los mismos aplicados de forma iterativa e interactiva.

Dentro de las definiciones que se pueden encontrar en la bibliografía relacionada se muestran algunas de las más reveladoras:

- "Data Mining es la exploración y análisis, mediante métodos automáticos o semiautomáticos, de grandes cantidades de datos para descubrir reglas o patrones significativos" [5].
- "Data Mining es el proceso analítico diseñado para explorar grandes cantidades de datos (típicamente relacionados con el mercado o los negocios) con el fin de investigar patrones consistentes y/o relaciones sistemáticas entre variables y, a continuación, validar los resultados aplicando modelos detectados para nuevos subgrupos de datos" [52].
- "Data Mining es el conjunto de técnicas y herramientas aplicadas al proceso trivial de extraer y presentar el conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con el objeto de predecir de forma automatizada tendencias y comportamientos y/o descubrir de forma automatizada modelos previamente desconocidos" [44].
- "Data Mining es el descubrimiento eficiente de información valiosa, no obvia, de una gran colección de datos" [6].
- "Data Mining es la extracción de información implícita, previamente desconocida y potencialmente útil de una base de datos" [56].

- "Data Mining combina técnicas de la estadística, inteligencia artificial, Bases de Datos, Visualización y otras áreas, para descubrir, de forma automática o semiautomática, modelos de (algunas veces enormes) series de datos." [49].
- "Data Mining es el proceso de plantear varias preguntas y extraer información útil, patrones y tendencias de grandes cantidades de datos generalmente almacenados en bases de datos." [53].
- "Data Mining es el análisis de, habitualmente grandes, series de datos (observaciones) para encontrar relaciones inesperadas y resumir la información de nuevas maneras que sean entendibles y útiles por el propietario de los datos."[24].

Sin embargo, en la esfera del Descubrimiento de Conocimiento en Bases de Datos, Knowledge Discovery in Databases, KDD o la Minería de Datos tiene otro significado. Efectivamente, el término KDD se empezó a utilizar en 1989 [44] popularizándose por lo expertos en Inteligencia Artificial (IA) y aprendizaje de ordenadores (*Machine Learning,* **ML**) para referirse al amplio proceso de búsqueda de conocimiento en bases de datos y para enfatizar de que este "conocimiento" es el producto final del proceso del KDD. La definición de KDD más representativa surge de diversos autores especialistas en ese campo:

- "Descubrimiento de Conocimiento en Bases de Datos (KDD): es el proceso no trivial de identificar patrones en datos que sean válidos, novedosos, potencialmente útiles y por último comprensibles" [19] [20].

Las fases en que se divide el KDD según [16] [42] son: exploración del dominio, recolección de los datos, extracción de patrones en los datos, inducir generalizaciones, verificación del conocimiento, transformación del conocimiento.

- De esta forma, y según los autores provenientes del campo de la IA o del ML, el Data Mining corresponde a un paso del KDD y se define en la literatura especialista de las siguientes formas:
- "Data Mining consiste en obtener modelos comprensibles o patrones de una base de datos" [49].

- "Data Mining: búsqueda de patrones de interés mediante árboles o reglas de clasificación, técnicas de regresión, clusterizado, modelizado secuencial, dependencias, etc." [54].
- "El proceso de extraer patrones o modelos a partir de los datos" [19].

En esta última definición coinciden la mayor parte de los autores que se dedican al Data Mining, el KDD, la IA o el ML, aunque también otros autores, como se ve en definiciones anteriores, describen el DM como el proceso completo.

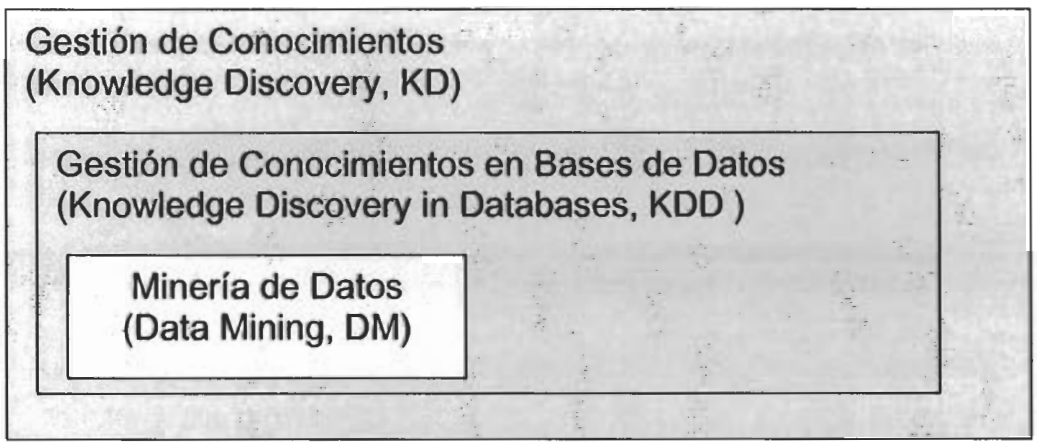


Figura 1. La Minería de Datos frente al KD y KDD.

En la Figura 1.se resume cómo el DM está incluido en el KDD, y cómo éste se incluye a su vez en el KD.

1.2.2 Cronología.

Aunque los elementos claves de la MD existen desde hace décadas en áreas de investigación como inteligencia artificial, estadística o el aprendizaje automático, se puede alegar que es ahora cuando se presenta al reconocimiento de la madurez de estas técnicas.

Las raíces del MD se remontan a los años 50. En dicha época, los departamentos de informática preparaban resúmenes de la información, primordialmente de tipo comercial, que se encontraba en los ficheros de una computadora central, con la intención de facilitar la labor directiva. Así nacieron

los sistemas de información para la dirección (EIS), que sin embargo, eran voluminosos, poco flexibles, y embarazosos de leer para los ajenos a la informática.

En los años 60 nacen los sistemas gestores de base de datos que aún se exponían estrictos y carecían de maleabilidad para realizar consultas. Posteriormente aparecieron los motores relacionales resolviendo estas dificultades, aunque los informes resultaban muy laboriosos de preparar y depurar, perdiéndose relevancia por su bajo nivel de actualización. Otro grave problema era la diversidad de bases de datos no integradas, establecidas por los diferentes departamentos de una organización.

El término "Descubrimiento de Conocimiento en Bases de Datos" (Knowledge Discovered in Data Bases o KDD para abreviar) empezó a usarse entre los especialistas de inteligencia artificial y aprendizaje de ordenadores, para referirse al extenso proceso de búsqueda de conocimiento en bases de datos y para enfatizar el hecho de que "el conocimiento" es el producto del incremento del ritmo de adquisición de datos. El aumento de la cantidad de datos almacenados se ve favorecido no sólo por la depreciación de los discos y sistemas de almacenamiento masivo, sino también por la automatización de muchos trabajos y técnicas de recogida de datos.

De esta forma, surge el término Minería de Datos a finales de la década de los 80, de las similitudes que existen entre buscar valiosa información de negocio en grandes bases de datos y minar una montaña, para encontrar una veta de metales preciosos. Fundamentalmente, el avance de la Minería de Datos en las empresas se debe fundamentalmente a estos aspectos:

- Uso de la información para la búsqueda de la mejora de la competitividad en aspectos como: mejora de la calidad, reducción de costos, control de la producción, optimización de los recursos, etc.
- Incremento de la potencia de las computadoras y abaratamiento de las mismas.
- Incremento del ritmo de adquisición de datos. El crecimiento de la cantidad de datos almacenados se ve beneficiado no sólo por el abaratamiento de los discos y sistemas de almacenamiento masivo, sino también por la automatización de muchos trabajos y técnicas de recogida de datos.
- Aparición de nuevos métodos de técnicas de aprendizaje y almacenamiento de datos.

1.2.3 Fases en un proceso clásico Minería de Datos.

Las fases en el proceso global de Minería de Datos no están notoriamente diferenciadas lo que hace que sea un proceso iterativo e interactivo con el usuario experto. Las interacciones entre las decisiones tomadas en diferentes fases, así como los parámetros de los métodos utilizados y la forma de representar el problema suelen ser considerablemente complejos. Pequeños cambios en una parte pueden afectar fuertemente al resultado final. La estructura del proceso consta de seis fases (tal como se muestra en la Figura.1) que serán desarrolladas con detalle durante este capítulo.

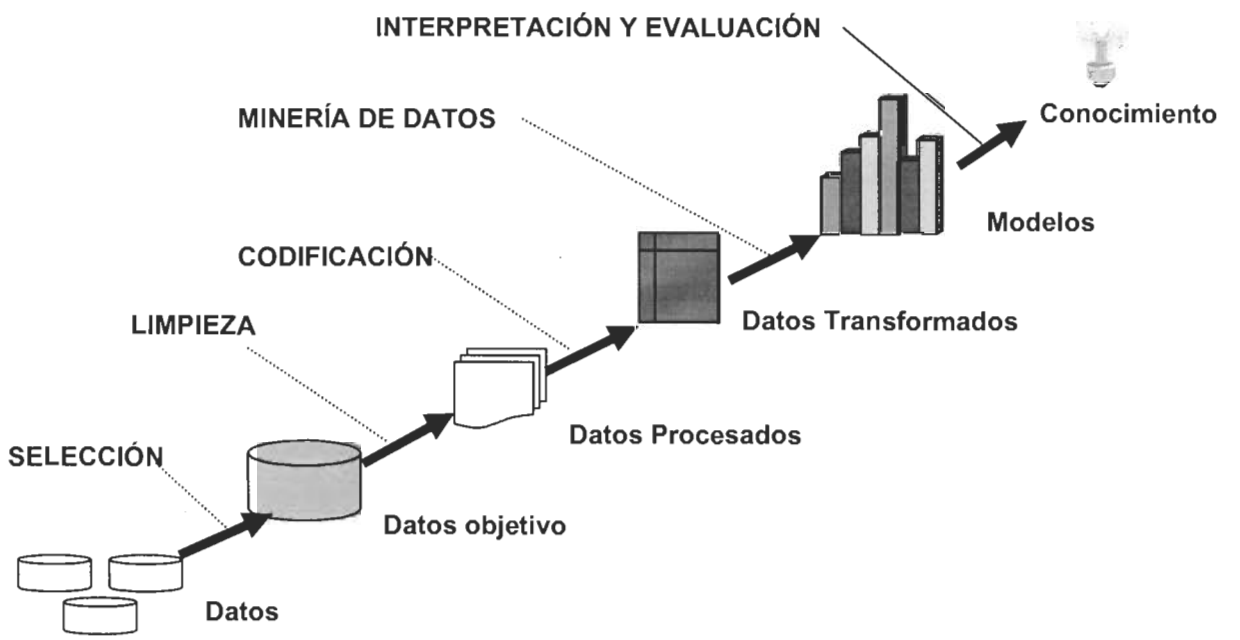


Figura 2. Fases en un proceso clásico Minería de Datos.

1.2.3.1 Definición del alcance y objetivos.

El primer paso de un proyecto de *Minería de Datos* radica en conocer el desarrollo y dominio de la aplicación, determinar el conocimiento relevante a usar, así como establecer los objetivos del usuario final. El progreso de esta primera fase, establece las bases para la realización de las posteriores fases del proceso, por lo que el éxito o fracaso del proceso va a depender en gran medida

de las decisiones que se tomen en esta etapa. En esta fase se establecen los factores que son susceptibles de un procesado automático, los cuellos de botella del dominio, los conocimientos a priori que se tienen del proceso, así como cuáles son los objetivos finales que se intentan lograr y cuáles van a ser los criterios de rendimiento exigibles. Por tanto, esta fase requiere cierta dependencia usuario-analista. Siendo necesario el establecimiento de unos canales de comunicación entre ambas partes.

Otro factor clave corresponde al conocimiento que se tiene del sistema. Muchas veces, lagunas de conocimiento del proceso a analizar, pueden involucrar pérdidas significativas de tiempo en fases ulteriores. Por lo general, es conveniente volcar todos los esfuerzos iniciales en comprender el sistema en todos sus pormenores, para evitar situarse en posiciones incomodas (callejones sin salida) debido a una falta de conocimiento de algunas de las partes del proceso.

La importancia de este último factor queda manifestada en una de las metodologías de implantación de minería de datos más usada actualmente, el método CRISP-DM (Cross-Industry Standard Process for Data Mining) que se estudiará con más detalle en epígrafes posteriores.

Se considera que un 80% de la importancia para llegar al éxito proviene en la forma de abordar el problema, definir cuales pueden ser las pautas para llegar a la solución y la forma de implementarlas para solucionar el problema con éxito, (Tabla 1.).

Tarea	Tiempo dedicado (%)	Importancia para llegar al éxito final (%)
1. Definir problema.	10	15
2. Explorar solución.	9	14
3. Implementación de los resultados.	1	51
4.1. Data Mining: Preparación de Datos.	60	15
4.2. Data Mining: Procesamiento de Datos.	15	3
4.3. Data Mining: Modelizado.	5	2

Tabla 1. Porcentajes de tiempo e importancia en las fases de DM.

1.2.3.2 Selección de datos relevantes.

La identificación de los datos relevantes para una operación de *minería de datos* es una tarea que no puede ser automatizada y que por lo tanto debe ser realizada por personal humano (analista). Esta tarea consiste en la creación del conjunto de datos objetivo, enfocando la búsqueda en subconjuntos de variables y/o muestras de datos en donde realizar el proceso de análisis.

En esta fase deben de ser seleccionados, de forma coordinada por el analista y el usuario, los datos más relevantes del proceso, así como su disponibilidad. Esto implica consideraciones sobre la homogeneidad y variación a lo largo del tiempo de los datos, los grados de libertad o la estrategia de muestreo.

1.2.3.3 Preprocesado y Limpieza de datos.

El objetivo del preprocesado de datos es la transformación del conjunto original de datos en un nuevo conjunto de datos más significativo y manejable. El preproceso es una transformación T que transforma la matriz que contiene los datos reales del proceso, X, en una nueva matriz Y tal que:

- Y conserva la información de X.
- Y elimina al menos uno de los problemas contenidos en X.
- Y es más útil que X

El preproceso de los datos incluye cuatro etapas principales: Identificación y conversión de tipos, imputación (rellenar los datos inexistentes), identificación de espurios (outliers), eliminación de ruido y datos incompletos.

Identificación y Conversión de atributos.

Las primeras tareas de preprocesado, son las más espinosas ya que, generalmente, deben consistir en identificar, casi manualmente, los diferentes tipos de variables existentes en la base de datos y convertirlos a otro tipo dependiendo de las necesidades posteriores. Fundamentalmente, podemos clasificarlos en los siguientes dos grupos [56]:

- Numéricos o Cuantitativos. También algunas veces llamados "continuos".

- Nominales o Cualitativos. También algunas veces llamados "discretos". Aunque la literatura estadística introduce unos "niveles de medida" clasificados en los siguientes subgrupos (aunque fundamentalmente se usan solamente los dos primeros):
  - *Nominales*. Que corresponden a valores que tienen distintos símbolos generalmente denominados etiquetas o nombres. Por ejemplo: colores. Un caso especial son los datos binarios (que solo pueden tener dos valores).
  - *Ordinales*. Que determinan un cierto ranking en las categorías. Por ejemplo: frío < templado < caliente, bajo < medio < alto, etc.
  - *Intervalos*. Que son valores que no solo están ordenados sino también medidos en unidades iguales con un cero arbitrario. Por ejemplo: temperaturas, años, etc.
  - *Ratios*. Que corresponden con medidas donde está definido un punto cero inherente en si mismo. Por ejemplo: la distancia de un objeto a otro, tiene como cero la distancia del objeto a si mismo, temperatura en grados absolutos, edad desde el *Big Bang*, etc.

Las variables, según el tipo que sean, deben ser acomodadas a los algoritmos que se vayan a utilizar. De esta forma, muchas veces resulta necesaria la conversión de los datos para que puedan ser tratados convenientemente. Este proceso de conversión depende en gran manera de los esquemas utilizados. Por ejemplo, algunos de los esquemas utilizan valores formados por escalas ordinales y solamente usan comparaciones *mayor-que*, *menor-que* para compararlos. Otros en cambio, usan escalas tipo ratios y usan distancias entre ellas. Es decir, **es necesario comprender cómo trabajan los algoritmos de minería de datos para saber como preparar los datos.**

Conversión de Tipos de Variables

En dependencia de los algoritmos a utilizar deberán transformarse los datos de un tipo a otro. Por ejemplo, una variable nominal no puede ser tratada por una red neuronal o un clasificador basado en árboles puede necesitar que los datos sean nominales. En [56] se profundiza con detalle en los diferentes tipos de transformaciones.

Muchas veces un atributo nominal puede ser convertido a un atributo ordinal simplemente indicándole al sistema unas reglas que relaciones estos. Por ejemplo, una serie de la forma: {bajo,

alto, medio}, cómodamente puede ser convertida en una serie ordinal simplemente mediante la regla: *bajo* < *medio* < *alto*, o *alto* > *medio* > *bajo*, aunque otras veces las reglas no son tan claras.

La conversión de datos nominales a numéricos dependerá del conocimiento que tengamos sobre el grado de proximidad o alejamiento de unos con otros. Por ejemplo, si tenemos una serie de datos del tipo: {*error insignificante, error medio, error peligroso*}, y queremos alimentar con ellos una variable numérica de un modelo matemático, será necesario desarrollar una escala de medidas numéricas que se adapten convenientemente.

Un caso más interesante, es la conversión de una serie de valores numéricos a una serie de datos nominales. Esta metamorfosis consistirá fundamentalmente en la creación de clases agrupando los conjuntos de datos según algún criterio preestablecido: distancia, similitud, en relación a otra variable, etc.

Otro tipo de transformaciones más avanzadas se basan en reglas *fuzzy* o difusas, capaces de tratar las indecisiones mediante funciones aplicadas a cada valor del campo.

Una vez que tenemos los tipos de atributos adaptados a nuestras necesidades, será conveniente realizar las siguientes fases:

- Detectar los espurios y eliminarlos.
- Rellenar los datos inexistentes.
- Eliminar el ruido.

Transformación de los datos.

La fase de transformación y reducción de los datos, es otra de las fases críticas dentro del proceso global que necesita de un buen conocimiento y una buena intuición que determinará el éxito o el fracaso del proceso de minería.

Se busca, por un lado, preparar la información que se tiene para que pueda ser procesada por los algoritmos de minería de datos y además, reducir la cantidad de información redundante para simplificar las tareas posteriores.

Se busca por lo tanto:

- Extracción de las características (o atributos) útiles de los datos (reducción de dimensionalidad).
- Transformación de los datos con el objetivo de proporcionar una representación de los datos más intuitiva y manejable.
- Fundamentalmente podemos destacar tres tareas específicas:
- Reducción de los Datos.
- Creación de Datos Derivados.
- Transformación de la distribución de los Datos.

1.2.3.4 Uso de los Algoritmos de Minería de Datos.

Una vez que se tienen los datos transformados y preparados en una base de datos normalizada, con variables poco correlacionadas entre si, con los espurios y el ruido eliminados, y con una dimensión adecuada; sería el momento del uso de los algoritmos de minería de datos.

Las herramientas de MD empleadas en el proceso de extracción de conocimiento se pueden clasificar en dos grandes grupos:

- Técnicas de verificación (en las que el sistema se limita a comprobar hipótesis suministradas por el usuario).
- Métodos de descubrimiento (en los que se han de encontrar patrones potencialmente interesantes de forma automática, incluyendo en este grupo todas las técnicas de predicción). El resultado obtenido con la aplicación de algoritmos de descubrimiento **puede ser de carácter descriptivo o predictivo**. Las predicciones sirven para prever el comportamiento futuro de algún tipo de entidad mientras que una descripción puede ayudar a su comprensión.

Antes de poder utilizar los datos, casi siempre es necesario preprocesarlos para adecuarlos a las necesidades de las técnicas que se van a utilizar sobre ellos. Las técnicas de visualización son muy útiles en este momento, para aumentar el conocimiento previo de los datos y como paso previo a procesos posteriores. También ayudan a descubrir la estructura de clusters de los datos y posibles correlaciones entre ellos, así como en la detección de espurios.

Los algoritmos de minería de datos pueden ser utilizados para alguna de las siguientes tareas:

- Agrupamiento o segmentación: Se busca la identificación de tipologías o grupos en los cuales los elementos guardan similitud entre sí y se diferencias de los otros grupos. Esto permite el tratamiento particularizado de cada una de estas agrupaciones.
- Asociación: Consiste en establecer las posibles relaciones entre acciones o sucesos aparentemente independientes. Así, se puede reconocer cómo la ocurrencia de un determinado suceso puede inducir la aparición de otro u otros.
- Secuenciamiento: Es un concepto similar al anterior, pero en el que se incluye el factor tiempo. Es decir, permite reconocer el tiempo que transcurre o suele transcurrir entre el suceso inductor y los sucesos inducidos.
- Reconocimiento de patrones: Se trata de analizar la asociación de una señal o información de entrada con aquella o aquellas con las que guarda mayor similitud, y que están ya catalogadas en el sistema. Generalmente se usan para identificar las causas de problemas o incidencias y buscar las posibles soluciones, siempre y cuando se disponga de la base de información necesaria en la que buscar.

Previsión: Se busca establecer el comportamiento futuro más probable de una variable o una serie de variables a partir de la evolución pasada y presente de las mismas o de otras de las cuales dependan. Las técnicas asociadas a estas herramientas tienen ya un elevado grado de madurez.

- Simulación: Comparan la situación actual de una variable y su posible evolución futura según la variación probable de las que depende.
- Optimización: resuelve el problema de la minimización o maximización de una función que depende de una serie de variables, encontrando los valores de éstas que satisfacen la condición de máximo (típicamente beneficios), o mínimo (típicamente costes). Normalmente suele haber unas restricciones, que hacen que no todas las posibles soluciones sean aceptables, de modo que el universo de búsqueda se reduce a aquellas soluciones que satisfagan las restricciones.

- Clasificación: Agrupa a todas las herramientas que permiten asignar a un elemento la pertenencia a un determinado grupo o clase. Esto se lleva a cabo a través de la dependencia de la pertenencia a cada clase en los valores de una sede de atributos o variables. Se establece un perfil característico de cada clase y su expresión, en términos de un algoritmo o reglas, en función de las distintas variables. Se establece también el grado de discriminación o influencia de estas últimas. Con ello, es posible clasificar un nuevo elemento una vez conocidos los valores de las variables presentes en él.

Para desarrollar todos estos procesos, se dispone de una extensa gama de técnicas que le pueden ayudar en cada una de las fases de dicho proceso.

1.2.3.5 Interpretación de los resultados.

La interpretación y verificación de resultados es un proceso complejo. La obtención de resultados aceptables dependerá de factores como: definición de medidas del interés del conocimiento (de tipo estadístico. en función de su sencillez) que permitan filtrarlo de forma automática, existencia de técnicas de visualización para facilitar la valoración de los resultados o búsqueda manual de conocimiento útil entre los resultados obtenidos.

Un factor muy importante en esta fase, es el grado de experiencia y conocimiento del analista. La cantidad de información extraída depende en gran medida del grado de conocimiento que el analista tenga del problema, así como de sus experiencias en la resolución de problemas similares.

Las decisiones tomadas durante esta fase irán encaminadas en dos direcciones:

- Verificación de resultados: La verificación de resultados incluye determinar el grado de cumplimiento de los objetivos finales establecidos durante la primera fase del proceso de MD, así como la validación de la información extraída. Durante esta fase se debe verificar la coherencia de la información obtenida con otros tipos de conocimiento ya previamente asentado y aceptado, resolviendo las posibles inconsistencias existentes. Si los objetivos finales han sido alcanzados, se procederá a la consolidación del conocimiento descubierto, incorporándolo al sistema, o simplemente documentándolo y enviándolo a la parte interesada. En caso contrario se procederá a la obtención de más información.

- Obtención de más información: La información extraída se utilizará como información a priori para la extracción de más información. Para ello será necesario retornar a alguna de las fases anteriores del proceso de MD y modificar algunas de las decisiones tomadas durante esas fases, haciendo para ello uso de la nueva información obtenida. De esta forma el proceso de MD se convierte en un proceso potencialmente iterativo. Alguna de las decisiones que pueden ser tomadas para la obtención de más información son, por ejemplo: recolección de nuevos datos, separación de datos en clases, transformaciones de las variables, eliminación de datos, selección de otros algoritmos de MD, cambio en los parámetros introducidos en los algoritmos, delimitación del campo de búsqueda, etc.

1.2.4 Herramientas de Minería de datos.

En el **Anexo 1** (Encuesta sobre herramientas de Minería de datos usadas regularmente. [1324 votos]) Se puede apreciar, el resultado de una encuesta hecha en el conocido portal sobre Minería de Datos y Gestión del Conocimiento, KDnuggets [30], donde se pregunta al encuestado sobre la herramienta de MD que habitualmente usa. Este tipo de encuesta es particularmente importante, porque refleja una idea de las aplicaciones que más están usando los profesionales y puede ayudar a decidir correctamente a la hora de adquirir uno de estos programas.

La lista que aparece en la encuesta es una pequeña muestra de las múltiples aplicaciones que existen en el mercado. De ella destacan programas comerciales que forman parte de familias de aplicaciones estadísticas como por ejemplo: SAS (SAS, SAS EnterpriseMiner), o SPSS (SPPS Clementine, SPSS AnswerTree) y que son preferencia de aquellos que habitualmente trabajan con estos paquetes.

Por otro lado, este tipo de aplicaciones comerciales contrastan con otras desarrolladas íntegramente en el campo de la Minería de Datos como por ejemplo: CART/MARS, IBM-I-Miner, Angoss, Megaputer PolyAnalyst, KXEN, etc.; y que fundamentalmente abarcan métodos estadísticos y de visualización combinados con algoritmos, bastante eficientes, más propios de MD (clasificadores, generadores de reglas, clusterizado, etc.).

Habitualmente, estas herramientas disponen de sus propios entornos gráficos y suelen permitir al usuario hacer múltiples tareas, pero siempre acotados a las especificaciones de cada aplicación [23]. El grado de eficiencia de cada herramienta depende de múltiples factores: tipos de algoritmos,

funciones de tratamiento de la información, eficiencia de los algoritmos, generadores de informes, formas de pasar la información, etc.; aunque generalmente, los primeros de la lista cubren bastante bien las expectativas que se espera de ellos. Algunos de ellos pueden ser descargados de internet y evaluados durante un corto periodo de tiempo.

Por otro lado, en la lista, se alza la herramienta WEKA [55].

Esta aplicación es de libre distribución (licencia GPL) y destaca por la cantidad de algoritmos que presenta así como por la eficiencia de los mismos. Esta aplicación está desarrollada por miembros de la Universidad de Waikato (Nueva Zelanda) y es una muy buena opción, tal y como muestra la encuesta, frente a las costosas distribuciones comerciales.

Se han obtenido excelentes resultados con las herramientas de libre distribución siguientes:

- R: Herramienta excelente para el análisis de datos basada en el conocido programa estadístico S-Plus y con un manejo de las matrices y variables equivalente a MATLAB. Este programa es muy útil para el análisis estadístico, transformación y manipulación de los datos. Está compuesto de múltiples librerías para realizar: gráficos y análisis estadísticos de todo tipo, regresiones lineales y no lineales, modelizado, clusterizado, etc.; y sigue en continua evolución. Cabe destacar la excelente asesoría técnica (responden las preguntas en pocas horas) llevada a cabo principalmente por algunos de los principales profesores e investigadores en estadística del mundo.
- WEKA: Programa de libre distribución que abarca algoritmos clasificadores de todo tipo, generadores de reglas, herramientas de clusterizado, etc. Esta aplicación proporciona gran cantidad de herramientas para la realización de tareas propias de minería de datos y permite la programación en JAVA de algoritmos más sofisticados.
- XELOPES: Otra librería de libre distribución con cantidad de funciones para minería de datos. Permite la implementación en JAVA o C++.
- SNNS: Aplicación de libre distribución para el desarrollo, entrenamiento y testeo de multitud de tipos diferentes de redes neuronales. Muy útil para desarrollar clasificadores sofisticados y modelos basados en redes neuronales.

- XmdvTool, Xgobi, IBM-OpenDX, Visipoint: Otras herramientas con licencia GPL que tienen diferentes funciones de visualización muy útiles para encontrar patrones ocultos en los datos.

Hoy en día, existen herramientas de libre distribución, realmente sorprendentes. Las que se acaban de enumerar, y muchas otras, permiten múltiples posibilidades. Los programas R y WEKA, usados conjuntamente, no solo se pueden utilizar como herramientas de aplicación, sino también, como auténticos entornos de programación. Esta característica, como es lógico, unido a que su coste es cero por ser programas con licencia GPL, aporta múltiples ventajas para los campos de investigación y docencia en el aprendizaje y desarrollo de la Minería de Datos [34]. La experiencia obtenida con estas últimas herramientas, nos ha demostrado que las ventajas en el campo de la investigación son muchas...

1.2.5 Aplicaciones de la Minería de Datos y Tendencias.

Hasta ahora esta tecnología ha sido de gran ayuda en áreas como la banca (detección de fraudes, análisis de morosidad o segmentación del mercado), telecomunicaciones (control de fugas de clientes, control de redes, ventas cruzadas), seguros (riesgos, mercadeo) y comercial.

Actualmente hay un número creciente de organizaciones inmersas en proyectos de MD. La tecnología se puede aplicar a cualquier organización que disponga de una gran cantidad de datos y que se plantee explotarlos para obtener reglas de negocio o mejorar el servicio que presta.

Se presentan a continuación algunos ejemplos [29]:

- Predicción automática de tendencias y comportamientos [9]:
  - Marketing dirigido: analizar datos sobre envíos por correo publicitarios para identificar el segmento más apropiado para realizar un nuevo mailing.
  - Comportamiento del cliente en supermercados en ciertos días de la semana, de forma que se puedan promocionar ciertos productos en fechas determinadas.
  - Análisis de las ventas de una compañía farmacéutica para reforzar las acciones de marketing en los hospitales y médicos de mayor impacto.

- Identificación de mejores clientes para el lanzamiento de una nueva tarjeta de crédito.
- Detección de fraudes en distribuidores de una empresa multinacional.

- Descubrimiento automático de patrones ocultos:
  - Análisis de datos de ventas de productos para identificar aquellos que sin estar relacionados entre sí, se compran juntos a menudo.
  - Detección de transacciones fraudulentas realizadas con tarjeta de crédito.
  - Detección de errores de grabación de datos.
  - Búsqueda de tendencias en la bolsa.
  - Determinación de las causas que producen los fallos en sistemas de producción.
  - Descriptores que "expliquen" los fallos de calidad en el producto.
- Prospectiva:
  - Conseguir modelos aplicables a bases de datos para la selección priorizada de nuevos clientes.
  - Estudios de respuesta ante un posible cambio de precios.
  - Generar nuevos modelos de control de un proceso.
- Segmentación y Clustering:
  - Dividir la base de datos de clientes en segmentos relativamente homogéneos basados en conductas estudiadas.
  - Una organización bancaria puede estudiar qué grupo de usuarios tiene una alta probabilidad de cancelar su cuenta en función de determinados parámetros y a continuación realizar acciones específicas para evitar que ocurra.
  - Clasificar los tipos de clientes de una empresa de seguros.
- Aplicaciones científicas:

- Análisis de los datos obtenidos a partir de instrumental científico. Esto permite el análisis de los datos para investigación, la formación de hipótesis y teorías.
- Aplicaciones en Biología y Medicina, bioinformación que se traduce en minería de bases de datos distribuidas (por ejemplo el proyecto Genoma).

Ver Anexo 3. (Campos donde actualmente de aplica la MD. Agosto 2003 [279 votes total].)

1.2.6 Dificultades en la aplicación de la MD.

Se enumeran a continuación algunos de los problemas más habituales a los que se enfrenta cualquier proyecto de MD:

- Uno de los mayores problemas es que el número de posibles relaciones es demasiado grande, y resulta prácticamente imposible validar cada una de ellas. Para resolver este problema, se utilizan estrategias de búsqueda, extraídas del área de aprendizaje automático ML [5].
- Además todas estas herramientas siguen funcionando mejor fijándoles objetivos de búsqueda concretos. Si bien la minería de datos da la impresión de que se puede simplemente aplicar como herramienta a los datos, se debe tener un objetivo, o al menos una idea general de lo que busca.
- El coste de esta prospección de datos debe ser coherente con el beneficio esperado. Si bien las herramientas (hardware y software) han bajado su precio, el coste en tiempo, personal y consultoría se ha incrementado, llegando en algunos casos a hacer inviable el proyecto.
- Suele funcionar mejor en problemas ligados a empresas de éxito que en otros casos, debido a la gran dependencia que estas herramientas tienen respecto a todos los estamentos de la empresa, desde mantenimiento a compras.
- Es necesario trabajar en estrecha colaboración con expertos en el negocio para definir modelos. Su ausencia y/o disponibilidad marca el proyecto.
- Otro problema es que la información muchas veces está corrompida, tiene ruido, o simplemente le faltan partes. Para esto, se aplican técnicas estadísticas que ayudan a estimar la confiabilidad de las relaciones halladas.

1.3 METODOLOGÍAS DE APLICACIÓN DE MD.

A la vista de las dificultades anteriores y para utilizar estas técnicas de forma eficiente y ordenada, es preciso aplicar una metodología estructurada. A este respecto se proponen las siguientes metodologías, siempre adaptables a la situación a la que se aplique.

1.3.1 Metodología CRISP-DM.

CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) [11], es una metodología para el desarrollo de proyectos de MD que se ha convertido en un estándar de facto.

El consorcio CRISP-DM, responsable de esta metodología, está integrado por importantes empresas europeas y estadounidenses que poseen una amplia experiencia en proyectos de análisis de datos relacionados con muy diversos campos de la industria.

La metodología para minería de datos CRISP-DM, está definida como un proceso jerárquico, que consiste en un conjunto de tareas descritas en cuatro niveles de abstracción, desde el general hasta el específico: fase, tareas generales, tareas específicas e instancias de proceso.

Al nivel más alto, el proceso está organizado en un número de fases; cada fase consiste en varias tareas generales de segundo nivel. Este segundo nivel se denomina genérico porque se pretende que sea lo suficientemente general como para cubrir todas las posibles situaciones. Las tareas generales deben ser lo más completas y estables posibles. Se entenderán tareas completas a aquellas que cubran completamente el proceso de análisis y sus posibles aplicaciones. Por otro lado, se entiende como estables aquellas tareas que cubran incluso desarrollos aún no conocidos.

El tercer nivel, el nivel de tareas especializadas, es el lugar en el que se describe cómo las acciones de las tareas generales (nivel 2) se deberían desarrollar en ciertas situaciones específicas. Por ejemplo, en el segundo nivel puede existir una tarea general llamada "limpieza de datos". El tercer nivel describe cómo difiere esta tarea de unas situaciones a otras, por ejemplo la limpieza de valores numéricos y la limpieza de valores categóricos o si el tipo de problema es un clusterizado a un modelo predictivo.

La descripción de fases y tareas en pasos discretos desarrollados en un orden específico representa una secuencia idealizada de eventos. En la práctica, muchas de estas tareas pueden ser desarrolladas

en un orden diferente y frecuentemente será necesario volver atrás a tareas previas y repetir ciertas acciones. El modelo de procedimiento no pretende abarcar todas estas posibles rutas a lo largo del proyecto porque esto requeriría un modelo enormemente complejo.

El cuarto nivel, el nivel de instancias de proceso, es un conjunto de acciones, decisiones y resultados sobre el proceso de MD en curso. Una instancia de proceso se organiza de acuerdo con las tareas definidas en los niveles superiores, pero representa lo que pasa en realidad en un proceso particular, más que lo que pasa en general.

Horizontalmente, la metodología CRISP-DM distingue entre el modelo de referencia y la guía del usuario. El modelo de referencia presenta una vista rápida de las fases, tareas y sus salidas y describe lo que hay que hacer en un proyecto de MD. La guía del usuario da consejos y trucos mucho más detallados para cada fase y para cada tarea dentro de una fase y describe cómo desarrollar un proyecto de análisis de datos.

1.3.1.1 Contexto del Proyecto

El contexto del proyecto dirige el paso entre el nivel general y el especializado en el CRISPDM.

Actualmente se distinguen cuatro dimensiones diferentes de contextos:

- El dominio de aplicación: es el área específica en la cuál el proyecto tiene lugar.
- El tipo de problema: describe la clase(s) específica(s) de objetivo(s) que el proyecto va a abarcar.
- El aspecto técnico: cubre temas específicos que describen diferentes desafíos técnicos que puedan ocurrir durante el proceso.
- La dimensión de herramientas y técnica: especifica qué herramientas y/o qué técnicas se van a aplicar durante el proyecto.

Un contexto específico de minería de datos es un valor concreto para una o más de estas dimensiones. Por ejemplo, un proyecto que abarca un problema de clasificación en estimación de producción constituye un contexto específico. Cuantos más valores de dimensiones de diferentes contextos se cubran, más concreto es el contexto.

1.3.1.2 Proyección.

Se distinguen 2 tipos diferentes de proyecciones entre los niveles genérico y especializado en CRISP-DM:

- Proyección para el presente: si sólo se está aplicando el modelo genérico del proceso para llevar a cabo un solo proyecto y se pretende proyectar las tareas generales y sus descripciones para ese proyecto concreto, se habla entonces de una proyección sencilla para (probablemente) un solo uso.
- Proyección para el futuro: si se especializa el modelo genérico del proceso de acuerdo a un contexto predefinido encaminándolo a un modelo de proceso especializado para usar en el futuro en contextos similares.

Siendo evidente que el tipo de proyección apropiado depende del contexto específico y de las necesidades de cada organización.

1.3.1.3 Como proyectar.

La estrategia básica para proyectar el modelo genérico de proceso al nivel especializado es la misma para todos los tipos de proyecciones:

- Analizar el contexto específico.
- Eliminar cualquier detalle que no sea aplicable en dicho contexto.
- Añadir detalles específicos al contexto.
- Especializar (o instanciar) contenidos genéricos de acuerdo a características concretas del contexto.
- Posiblemente, y para una mayor claridad, renombrar contenidos genéricos.

CRISP-DIM define las diferentes fases de las que consta un proyecto, las tareas correspondientes y las relaciones entre ellas.

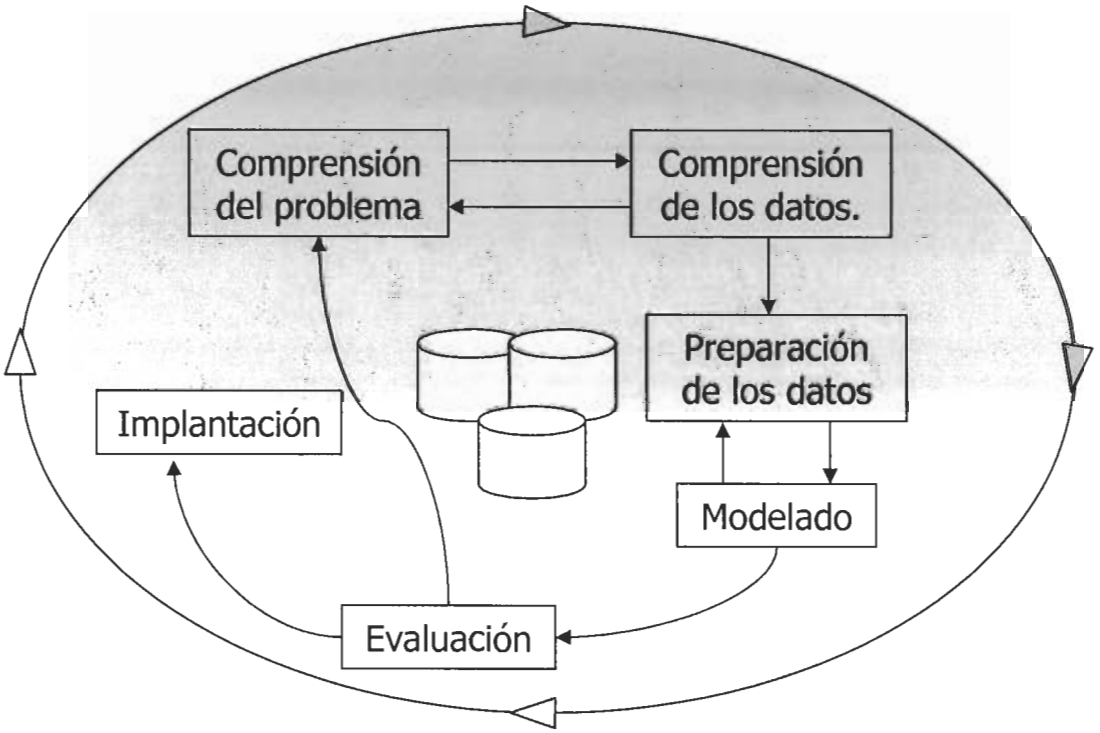


Figura 3. Fases del modelo de referencia CRISP-DM.

En la Figura 3. Se muestran las 6 fases definidas. El orden de las mismas no es estricto, ya que frecuentemente a lo largo del desarrollo del proyecto, es necesario volver atrás en numerosas ocasiones, dependiendo de los resultados obtenidos en las fases previas. Las flechas indican las relaciones más habituales entre las fases. El círculo exterior simboliza la naturaleza cíclica de la MD, ya que la solución a la que finalmente se llega puede conducir al planteamiento de nuevas cuestiones que den origen a otros proyectos.

A continuación, se resumen las tareas genéricas en las que se desglosan cada una de las fases y las salidas generadas por cada una de ellas. A continuación se describen detalladamente.

El ciclo de vida de un proyecto de MD consiste en 6 fases:

- **Análisis del Problema**: Fase inicial que incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva de negocio, con el fin de convertirlos en objetivos y en una planificación.

- **Análisis de los Datos**: Recolección inicial de datos para familiarizarse con ellos, identificar su calidad y descubrir las relaciones entre los más evidentes para las primeras hipótesis de relaciones ocultas entre ellos.
- **Preparación de los Datos**: Construcción de la base de datos a partir de los datos primarios. Estas tareas se desarrollan en numerosas ocasiones y no de una forma muy estructurada. Incluye la selección de tablas, registros y atributos, así como su transformación y preparación para las herramientas de modelizado.
- **Modelizado**: Se seleccionan y aplican varias técnicas de modelizado. Normalmente existen varias técnicas para el mismo problema y cada una exige una entrada de datos particular por ello es necesario interactuar con la fase anterior para adecuar la base de datos de trabajo. Los parámetros son calibrados.
- **Evaluación**: Una vez creado un buen modelo se debe evaluar el rendimiento del mismo y la integridad de todos los pasos sobre todo teniendo en cuenta que se han introducido todos los criterios de negocio. Se debe dar el visto bueno final a la aplicación del modelo de DM.
- **Desarrollo o Implantación**: Normalmente los proyectos de DM no terminan en la implantación del modelo sino en el incremento de conocimiento obtenido de los datos. Para ello es imprescindible documentar y presentar los resultados de manera comprensible. Además debe asegurarse el mantenimiento de la aplicación y la posible difusión de estos resultado

1.3.2 Metodología SEMMA.

SAS Institute desarrollador de esta metodología [45], la define como el proceso de selección, exploración y modelizado de grandes cantidades de datos para descubrir patrones de negocio desconocidos.

El nombre de esta terminología es el acrónimo correspondiente a los cinco pasos básicos del proceso "*Sample, Explore, Modify, Model and Assess*". El esquema siguiente presenta la dinámica del sistema.

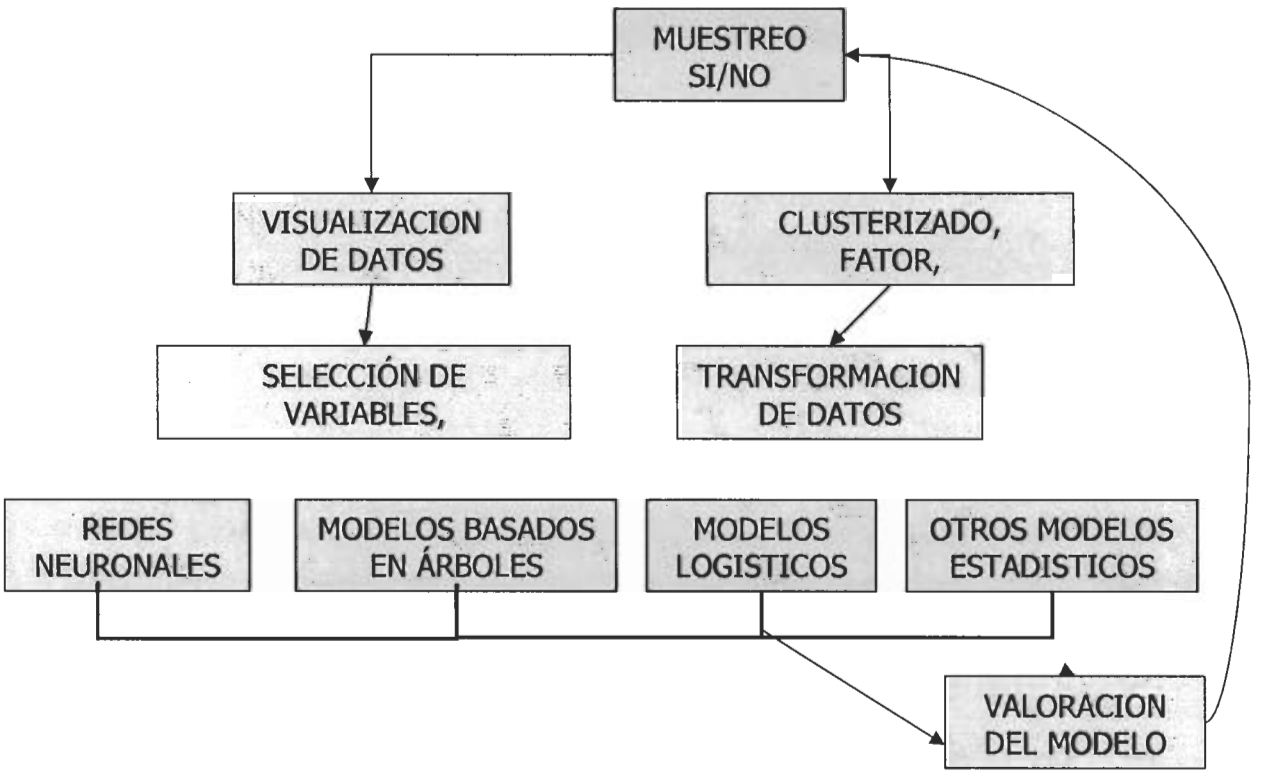


Figura 4. Metodología de SEMMA.

1.3.2.1 Muestreo.

Extracción de la población muestral sobre la que se va a aplicar el análisis. En ocasiones se trata de una muestra aleatoria, pero puede ser también un subconjunto de datos del *data warehouse* que cumplan unas condiciones determinadas. El objeto de trabajar con una muestra de la población en lugar de con toda ella, es la simplificación del estudio y la disminución de la carga de proceso. La muestra más óptima será aquella que, teniendo un error asumible, **contenga el número mínimo de observaciones.**

En el caso de que se recurra a un muestreo aleatorio, se debería tener la opción de elegir entre:

- El nivel de confianza de la muestra (usualmente el 95% o el 99%).

- El tamaño máximo de la muestra (número máximo de registros), en cuyo caso el sistema deberá informar del error cometido y la representatividad de la muestra sobre la población original.
- El error muestral que está dispuesto a cometer, en cuyo caso el sistema informará del número de observaciones que debe contener la muestra y su representatividad sobre la población original.
- Para facilitar este paso se debe disponer de herramientas de extracción dinámica de información con o sin muestreo (simple o estratificado). En el caso del muestreo, dichas herramientas deben tener la opción de, dado un nivel de confianza, fijar el tamaño de la muestra y obtener el error o bien fijar el error y obtener el tamaño mínimo de la muestra que proporcione este grado de error.

1.3.2.2 Exploración.

Una vez determinada la población que sirve para la obtención del modelo se deberá determinar cuáles son las variables explicativas que van a servir como entradas al modelo. Para ello es importante hacer una exploración de la información disponible de la población que permita eliminar variables que no influyen y agrupar aquellas que presentan efectos similares.

El objetivo es simplificar en lo posible el problema con el fin de optimizar la eficiencia del modelo. En este paso se pueden emplear herramientas que permitan visualizar de forma gráfica la información, utilizando las variables explicativas como dimensiones.

También se pueden emplear técnicas estadísticas que ayuden a poner de manifiesto relaciones entre variables. A este respecto resultará imprescindible una herramienta que permita la visualización y el análisis estadístico integrado.

1.3.2.3 Manipulación.

Tratamiento realizado sobre los datos de forma previa a la modelización, en base a la exploración realizada, de forma que se definan claramente las entradas del modelo a realizar (selección de variables explicativas, agrupación de variables similares, etc.).

1.3.2.4 Modelización.

Permite establecer una relación entre las variables explicativas y las variables objeto del estudio, que posibilitan inferir el valor de las mismas con un nivel de confianza determinado.

1.3.3 Metodología CRITIKAL.

Desarrollada en el marco de un proyecto ESPRIT 22700 [28] [47] se caracteriza por su fuerte integración con el desarrollo del *data warehouse* y no es de completa distribución libre. Los pasos que plantea son similares a los del CRISP-DM. Su principal fortaleza radica en la extensa valoración de otras herramientas antes de comenzar uno de los cuatro proyectos de DM en los que clasifica todos los problemas.

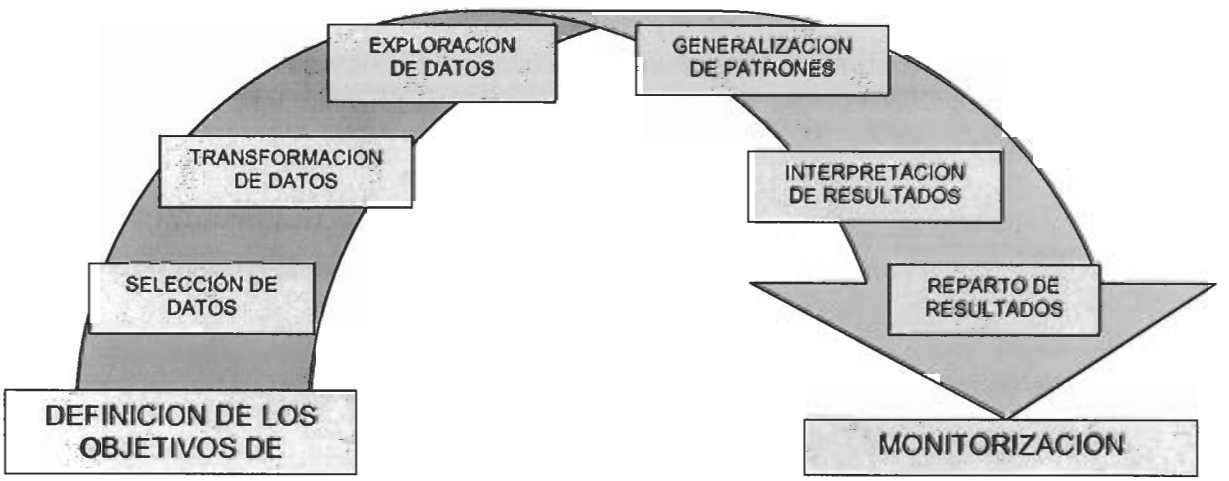


Figura 5. Pasos de la metodología CRITIKAL.

1.3.4 Metodología de las "5 A'S"

A título anecdótico, la metodología "5A's" la definió SPSS [51] antes de desarrollar junto con otras empresas la metodología "CRISP-DM". Su nombre viene de las cinco palabras siguientes: "*Assess, Access, Analyze, Act* y *Automate*". El significado de estas cinco palabras al contexto de la MD se explica a continuación:

- Asesorar (*Assess*): La clave de una minería de datos no solo está en la tecnología que se va a usar, sino también en la forma en que se va a manejar y transmitir la información, ya que al final el objetivo **es asesorar en la toma de decisiones**. El *data warehouse*, la tecnología de

manejo de la información, las herramientas a utilizar, etc.; deben estar orientadas a los procesos, estrategias y objetivos de la organización.

- Acceso (*Access*): Una vez que el contexto está planteado, entramos en la parte del proceso donde la tecnología puede ayudarnos. Es indispensable un sistema que nos ayude a recolectar la información de la mejor forma y con la mayor calidad posible.
- Analizar (*Analyze*): En esta fase, se hace uso de las diferentes herramientas de data mining para analizar la información y extraer el conocimiento deseado.
- Actuar (*Act*): Una vez se extraen conclusiones importantes, es conveniente plantear las posibles soluciones o aplicaciones. Generalmente hay que hacer uso de informes y gráficos que puedan ser interpretados por los agentes que toman las decisiones de actuar.
- Automatizar (Automate): La actividad del DM no termina cuando las decisiones se han tomado, sino que es necesario monitorizar los efectos de las mismas. Como esto necesita ser validado continuamente, muchas veces es necesario desarrollar un sistema automático que permita con "una simple pulsación de un botón", monitorizar los resultados de las decisiones adoptadas.

1.3.5 Metodología de DM: Conclusiones.

Últimamente la metodología CRISP-DM es la más aplicada a nivel mundial (**Anexo 2).**

Las razones fundamentales son debidas a su generalización y practicidad, además de su libre utilización.

Un punto en común que se puede observar en todas estas metodologías es que todas se basan en un modelo espiral, de forma que se retorna a las primeras fases del proceso pero aun nivel superior ya que la comprensión alcanzada es mayor. Además, en todas vemos la importancia del análisis inicial del contexto y de la validación final de la toma de decisiones.

2 CAPÍTULO 2. APLICACIÓN DE CRISP-DM.

2.1 INTRODUCCIÓN

Este capitulo contiene el despliegue de cada una de las fases de la metodología CRISP-MD aplicada a nuestro problema en específico, módulo de minería de datos del CC. Es preciso aclarar que algunas de estas fases no son desarrollarlas a plenitud, esto se debe a que el CC esta en trance de implementación lo cual imposibilita el empleo de algunas de las restricciones de las fases. Teniendo en cuenta la importancia del Modelado, este capítulo cuenta además con el análisis del Descubrimiento Directo, el cual contiene a este como parte de su proceso. Para dar cumplimiento a los objetivos que quedan fuera del los modelos se trata también el análisis del Descubrimiento Indirecto.

2.2 COMPRENSIÓN DEL PROBLEMA (FASE I CRISP-MD)

En esta fase se realiza un estudio detallado de los requerimientos y objetivos del proyecto en vista de que estos respondan aportando resultados sustanciales desde una perspectiva de negocio. Es de suma importancia transformar todo este conocimiento en una definición de un problema de Minería de Datos y crear un plan antecesor diseñado para lograr los objetivos.

2.2.1 Objetivos del negocio

2.2.1.1 Antecedentes

Debido a los avances en las organizaciones de negocios y en la tecnología, ahora se tiene una mayor habilidad en la captura de información sobre los clientes, que aquella que se tenía hace unos años. Existen dos tipos de información que se deben recolectar sobre los clientes, información sobre comportamiento e información sobre preferencias. La información sobre comportamiento es la información transaccional que se observa cuando el cliente interactúa con la compañía. La información de preferencias es aquella que los clientes suministran sobre si mismos, a través de sondeos y perfiles [61].

Partiendo de que estos objetivos estarán encaminados a los intereses de los proveedores y puestos en función de responder a la nueva tecnología CRM (*Customer Relationship Manager*), automatizando

los procesos de negocio horizontalmente integrados que involucran a todas aquellas áreas de la empresa que constituyen punto de contacto con el cliente, ventas (gestión de contactos, configuración de productos), marketing (gestión de campañas, tele marketing), servicio al cliente, a través de múltiples e interconectados canales de comunicación y servicio. Estas herramientas tecnológicas existentes hoy día, nos permite un mercadeo uno a uno, es decir, dejan a la empresa comunicarse con los consumidores a un nivel personal. Ahora bien, toda esta tecnología tiene que ver directamente con tres aspectos o áreas de una empresa: Ventas, Servicios y Marketing, en sí las referentes al trato con el cliente y que manejadas correctamente nos permite:

- Captar Clientes.
- Reforzar la Lealtad del Cliente.
- Mejorar Relaciones con el Cliente

Ahora, todo esto es viable gracias a la Tecnología Habilitadora para la Venta (TES) del CC. La implementación de un TES no se refiere únicamente a técnica pura, sino requiere la creación de nuevos procesos soportados por tecnología que integren la información del cliente, los datos de las transacciones de manera tal que generen información estratégica para la compañía [61].

2.2.1.2 Determinación de los Objetivos del negocio

Teniendo en cuenta los antecedentes y la particularidad que este modulo de minería pertenece a un CC integrado el cual servirá a proveedores de diferentes índoles, es necesario formular los objetivos de manera que estos compongan la intersección de sus intereses:

- Minimizar los costos y maximizar ganancias.
- Captar Clientes
- Reforzar la Lealtad del Cliente
- Mejorar Relaciones con el Cliente
- Elevar las ventas

Una vez definidos los objetivos es fácil percatarse que definir criterios de éxitos para el proyecto pudiera ser una tarea compleja, debido que para cualquier objetivo que se seleccione podemos encontrar un conjunto de acciones que viabilicen este. Por lo que tomaremos como criterio de éxito del proyecto cualquier resultado que de alguna forma garantice el cumplimiento de cualquiera de los objetivos.

2.2.2 Evaluar la situación

Para esta fase se realiza un estudio en detalles de todos los factores que pudieran influir en el proceso de extracción de conocimiento, fuese positiva o negativamente.

2.2.2.1 Recursos Disponibles

- Hardware
  - Dos equipos de mesa PC (ofimática) Pentium 4 arquitectura Intel con 256 MB de memoria RAM, sistema operativo Windows XP profesional.
- Software
  - Paquete Office completo Access, Excel, Word… etc.
  - Herramientas GNU de Minería de Datos Weka.
  - Herramienta de análisis estadístico Staditicas.
  - Matlab 6.5.
  - Microsoft Visual Studio .Net 2003.
  - Herramienta de manejo con base de datos Microsoft SQL Server.

2.2.2.2 Fuentes de Datos

Los datos con los cuales se trabaja son el resultado de la interacción de los clientes con la información que brindan los proveedores respecto a sus productos y servicios, estos son recopilados en el servidor de datos del CC, relacionando directamente a los clientes con los proveedores; conteniendo suma importancia en su totalidad para el proceso de extracción de conocimiento de los datos. Contando con un total de 221 variables.

2.2.2.3 Requerimientos, suposiciones y restricciones:

- Análisis de la caducidad de los datos para extracción de sus conocimientos esto requiere de personal especializado, en este caso de parte de los proveedores.
- Un alto porcentaje de fiabilidad de los datos.
- Estricto análisis con los datos, capas de garantizar que estos estén libres de errores.
- Garantizar una buena cantidad de datos para tener una mayor fiabilidad en los resultados del empleo de técnicas que requieran entrenamiento, como redes neuronales.
- Teniendo en cuenta que este proceso responderá a diferentes entidades pudiera arrojar resultados de poca relevancia para alguna entidad.

2.2.2.4 Riesgos y contingencias

- El soporte de software y hardware debe ser capaz de tratar gran volúmenes de datos en poco tiempo de procesado.
- Se debe garantizar una correcta comunicación entre las diferentes fases de proceso debido a que cualquier deficiencia en el almacenado o lectura de datos puede acarrear distorsiones en proceso de análisis.
- Como este modulo corresponde a un proyecto en fase de implementación, será necesario para recopilar datos esperar un tiempo prudencial en el cual permita contar con el suficiente volumen de datos para su explotación.

2.2.2.5 Costes y beneficios

Debido a que el proyecto es desarrollado en conjunto con la Universidad de Ciencias Informáticas esto conlleva a que los costos sean reducidos.

En cuanto a los beneficios es necesario aclarar que estos responden al CC y por otro lado a los proveedores.

- Beneficios CC
  - Brindar una mejor prestación de sus servicios y aprovechamiento de los datos almacenados.

- Sacarle provecho al conocimiento extraído mediante los proveedores.
- Detectar problemas de funcionamiento, morosidad en el CC.

- Beneficios Proveedores
  - Información de gran valor para trazarse estrategias de negocios exitosas.
  - Potencia la competitividad con respuestas rápidas y eficaces.
  - Reestructurar y detectar mal funcionamiento dentro de la entidad.

2.2.3 Determinar las metas de Minería de Datos

Una vez formalizados los objetivos respondiendo a las perspectivas de negocio del proyecto estamos en la fase de la metodología CRISP-MD donde estos son enfocados desde un ángulo de minería de datos, permitiendo el empleo de las técnicas pertinentes para la extracción de conocimiento.

2.2.3.1 Metas de minería de datos.

Emplear técnicas de minería de datos y análisis estadístico para detectar las variables de mayor peso en el proceso. Podemos aclarar que el empleo de la tecnología CRM nos puede aportar valiosos criterios para esta selección. Para dar cumplimiento a esta meta es necesario cumplir:

- Eliminación de basura en los datos, estudio visual, empleo de técnicas de visualización multivariantes para la eliminación de ruido, transformación de información.
- Empleo de gráficos multivariantes apoyados en índices estadísticos para detectar variables de mayor peso en el proceso.
- Extracción de conocimiento mediante técnicas basadas en detectar patrones de comportamiento.
  - Búsqueda de clases mediante algoritmos de clusterizado.
  - Clasificación de las clases basados en su comportamiento.
- Extracción de reglas de asociación que permitan construir modelos categóricos de variables.
  - Preparación de la base de datos anterior para el tratamiento correcto de los clasificadores.

- Empleo de clasificadores que permitan obtener reglas o modelos que expliquen los diferentes grupos obtenidos.

2.2.3.2 Criterios de éxito (perspectiva de minería de datos)

Para corroborar los objetivos definidos anteriormente es necesario tener de alguna manera un criterio que permita saber en que medida fueron logrados cada uno de ellos (criterio de éxito). Partiendo de esto para cada uno tenemos:

- Detectar variables de mayor peso en el proceso
  - Nivel de validez de los modelos obtenidos con dichas variables.
  - Acciones realizadas empleando el conocimiento obtenido.
- Extracción de conocimiento
  - Valoración de los resultados obtenidos por expertos.
  - Modo de empleo de los conocimientos, verificar su eficacia.
- Extracción de reglas de asociación
  - Valor de uso de los expertos con estas.
  - Eventualidades y nuevas estrategias sugeridas por estas reglas.

2.2.4 Producir un plan de proyecto

Dándole cumplimiento a la primera fase de la metodología **CRISP-MD** es preciso una planificación teórica de las tareas a realizar y las herramientas que intervendrán en el proceso. Es preciso en los casos que se pueda definir la duración estimada, recursos, entradas, salidas y riesgos.

2.2.4.1 Planificación y técnicas previstas

Las tareas definidas dentro de la planificación son:

1. Obtener el conjunto inicial de variables a tratar a partir del conocimiento de los expertos.
   - Tiempo previsto 6 semanas.
   - Métodos y técnicas: Plataforma .NET 2003, confrontación de opiniones.

- Recursos previstos: Expertos, documentación sobre la tecnología **CRM**.
- Entrada: 221 variables.
- Salida: Primer conjunto de variables

2. Recogida de los datos del servidor.
   - Tiempo previsto 1 año.
   - Métodos y técnicas: Herramientas para el manejo de bases de datos **SQL-Server**.
   - Recursos necesarios previstos: Una buena divulgación de la puesta en explotación del CC, para garantizar el uso del mismo por la mayor cantidad posible de usuarios.
   - Entradas: registro de acceso de los usuarios al CC.
   - Salidas: Base de datos formada por la interacción de los usuarios con la información de los proveedores con 221 variables.
3. Estudio exploratorio de los datos mediante técnicas estadísticas y visualización multivariable, selección de las variables mas significativas, filtrado de los datos, eliminación de los ruidos en los datos.
   - Tiempo previsto 6 meses.
   - Métodos y técnicas: Técnicas de visualización estadísticas y multivariantes.
   - Recursos necesarios previstos:
     - Herramientas de análisis estadísticos **Stadisticas**, índices estadísticos, clusterizado.
     - Herramientas de visualización multivariantes.
     - **Matlab 6.5**.
     - **Microsoft Access**.
     - **PC** (ofimática) **Pentium-4**, 256 Megabyte **RAM**.
   - Entradas: Base de datos obtenida en el paso anterior.

- Salidas: Base de Datos libre de ruidos, con las variables más significativas y con los datos transformados.

En caso de que los datos no sean adecuados será preciso volver a 2. Si se cree que se dejó de considerar alguna variable necesaria en el proceso será necesario retornar a 1.

4. Extracción de reglas de asociación y patrones de comportamiento.

- Tiempo previsto 4 meses.
- Métodos y técnicas: Técnicas inductivas (árboles clasificadores), Herramientas de clusterizado, reglas asociativas y de decisión.
- Recursos necesarios previstos:
  - Herramientas GNU **WEKA 3-4-1**.
  - Herramienta de análisis estadístico **Stadistica**.
  - **Matlab 6.5**.
  - **Microsoft Access**.
  - **PC** (ofimática) **Pentium-4**, 256 Megabyte **RAM**.
- Entradas: Base de datos libres de ruidos (filtrada), con las variables más significativas.
- Salidas: Reglas y árboles.

En este caso si los datos arrojados no son adecuados será necesario retornar a 3.

2.3 COMPRENSIÓN DE LOS DATOS (FASE II CRISP-MD)

La fase de comprensión de los datos comienza con una colección de datos inicial y realiza actividades para familiarizarse con los datos, identificar problemas de calidad para descubrir las primeras características de los datos o detectar subconjuntos para realizar las primeras hipótesis sobre la información oculta [38]. Podemos decir que el objetivo fundamental de la segunda fase consiste en analizar la información que se tiene verificando la calidad final. Si ésta no es aceptable, será necesario realizar nuevas adquisiciones de datos hasta que la calidad de éstos sea la adecuada.

Una vez conseguidos, se definen los objetivos de la tercera fase que consisten en preparar la información, seleccionando primeramente las variables más importantes, filtrar y eliminar el ruido y transformar los datos para la posterior etapa de modelizado (fase IV) [35].

2.3.1 Conseguir el conjunto inicial de datos.

Este primer paso consiste en la obtención de los datos y fuentes de donde son extraídos. Es necesario unificar las bases de datos en un número reducido y con estas las variables de mayor peso en el proceso.

2.3.1.1 Informe inicial sobre los datos

- Generar el conjunto inicial de datos.
- Elaborar un informe que describa la forma de adquirir los datos de las fuentes.

2.3.1.2 Se definen los siguientes pasos:

- Planificar requerimientos: Se analiza el tipo de información requerida.
  - Variables necesarias, tipos de rangos de cada variable.
  - Se comprueba si posible adquirir las variables y si están disponibles.
- Criterio de selección: Se definen los siguientes pasos:
  - Criterio de selección de las variables.
  - Seleccionar las tablas de mayor interés.
  - Caducidad o periodo de los datos (mes, días, años).

2.3.2 Describir los datos

En esta fase el objetivo fundamental es lograr total comprensión de la forma de los datos, para lo cual se describen las características fundamentales: Tablas, variables individuales (tipos de datos, cantidad de registros), etc.

2.3.2.1 Informe con la descripción de los datos

- **Análisis volumétrico de los datos**: Para acceder a los datos contamos con las herramientas que nos brinda la **plataforma .NET** (ADO.NET). Partiendo de que el servidor de base de

datos esta diseñado específicamente para el CC esto potencia la capacidad de almacenar solo lo pertinente con este. Todas las bases de datos usadas pertenecen al servidor, el volumen de datos contenido en estas depende primeramente de la cantidad de proveedores que se le preste servicios y del tiempo de funcionamiento de CC.

- **Tipos y valores de las variables**: Los datos contenidos respecto a tipo se pueden clasificar en numéricos y categóricos, en cuanto a rango de sus valores es necesario contar con los datos físicos y los proveedores de estos, como se ha comentado anteriormente el estado de implementación del CC no hace posible esto.
- **Claves**: El diseño del servidor de datos responde además del modulo de minería a las funcionalidades del CC, por lo que se realizó un estudio de cada una de las bases de datos donde estas se llevaron a tercera forma normal. Sobre la hipótesis de la información oculta que pudiera aportar acciones concretas a los objetivos definidos es evidente que por su naturaleza esta debe contenerse en la base de datos encargada de relacionar a los clientes con los proveedores (**BD Estadísticas&CRM**).
- **Revisión de Objetivos**: Como ya se a plantado la fuente de los datos responde directamente a los objetivos del problema, debido a que está diseñada para la colección de datos relacionados con el CC.

2.3.3 Explorar los datos

Una vez descritos los datos se realiza un primer análisis superficial de las particularidades de los datos:

- Relación entre variables: Como ya se ha comentado la particularidad de la fuente de los datos nos garantiza que las variables involucradas en el proceso estén relacionas directamente y en específico las de mayor peso.
- Tipo de distribución de los datos: Este punto queda indicado en el proceso debido a la ausencia física de los datos.
- Agrupamiento: Al este punto depender del anterior queda indicado.

Para lo cual se puede utilizar:

- Técnicas de visualización.
- Análisis de correlación.
- Técnicas estadísticas.

2.3.4 Verificar la Calidad de los datos

Una vez concluida esta fase se debe verificar la calidad le los datos disponibles valorando si estos son lo suficientemente buenos como para que aporten los conocimientos esperados o si no es así será necesario repetir los procesos anteriores.

Para lograr este propósito se analizará: Si los datos contienen errores, si describen realmente la realidad.

Para corroborar la calidad de los datos, revisar las variables teniendo en cuenta:

- Si representan la realidad y son consistentes.
- La cantidad de campos vacíos y el por que de los mismos.
- Variables innecesarias.
- Si existe espurios y su causa.

2.3.5 Comprensión de datos

2.3.5.1 Selección de las fuentes

En esta primera etapa, se parte de los datos obtenidos en la fase anterior y de la descripción de los mismos. A partir de toda esta información, se realiza una selección de las variables más importantes según los siguientes requerimientos:

- Que sean lo más independientes entre si. Se eliminarán las variables muy dependientes de otras.
- Que describan casi completamente el sistema a estudiar.
- Que tengan una relevancia individual destacada. Se descartarán aquellas cuya influencia en el sistema sea nula o muy escasa.
- Que estén exentas de ruido y con datos fiables.

2.3.5.2 Estudio de los datos

Como se ha mencionado anteriormente los datos utilizados son extraídos del servidor de bases de datos del CC, esto nos garantiza que toda la información aquí almacenada puede ser empleada para la extracción de conocimiento. Ahora teniendo en cuenta que los objetivos del módulo responden fundamentalmente a la tecnología **CRM** en casi su totalidad y partiendo de los requerimientos de esta, es fácil deducir que los datos propensos a brindar una mayor información son los que relacionan directamente a los clientes con los servicios o productos solicitados. De ahí que las bases de datos de mayor interés son las que nos aportan datos directamente de los clientes (a) y las que relacionan la interacción de estos con la información de los proveedores (b).

a) Base de datos Clasificadores: Contiene toda la información de un cliente.

b) Base de datos Estadísticas&CRM: Contiene todas las relaciones entre clientes y la información que brindan los proveedores (productos, servicios, etc.).

2.3.5.3 Establecer el tipo de las variables:

- Cuantitativas se distinguen a su vez en:
  - Discretas (Número de clientes, Cantidad de un producto, etc.).
  - Continuas (Precio de un producto, Tasa de cambio de una moneda, etc.).
- Cualitativas se distinguen a su vez en:
  - Nominales nombran el objeto al que se refieren (estado civil, sexo, etc).
  - Ordinales se puede establecer un orden en sus valores (alto, medio, bajo).

2.3.5.4 Establecer la caducidad de cada dato (vida de las variables).

Para realizar este análisis no basta con la presencia física de las variables es necesario que este estudio sea llevado a cabo por alguien experto en la materia que se relacione con esta variable, única persona capacitada para dar juicio del tiempo de vida de la variable. Esta fase es de suma importancia tanto para la extracción de conocimiento como para tomar criterios de poda de los datos.

2.4 PREPARACIÓN DE LOS DATOS (FASE III CRISP-MD)

2.4.1 Introducción.

El preprocesado de los datos es una de las tareas más importantes dentro de todo proceso que pretenda extraer conocimiento, modelar un sistema o evaluarlo. Efectivamente, varios autores subrayan la importancia, para la consecución con éxito del trabajo de DM, de las tareas iniciales de definición del problema, análisis de los datos y preparación de los datos [54].

La primera de estas fases, desarrollada apartados anteriores corresponde con el **análisis del problema**, donde se buscaba garantizar la perfecta comprensión del problema planteado y poder llegar así a una definición lo más completa posible de los objetivos finales. La segunda, corresponde con el **Comprensión de los Datos**, donde se buscaba garantizar la perfecta comprensión del problema planteado y poder llegar así a una definición lo más completa posible de los objetivos finales.

En este apartado se describen los pasos realizados en la tercera fase del proceso de *CRISP-DM:* la **Preparación de los Datos**.

Las fases dos y tres de la metodología CRISP-DM, tratan de preparar la información de la mejor forma posible para la posterior etapa correspondiente al modelado.

Se puede decir que el objetivo fundamental de la segunda fase consiste en analizar la información que se tiene verificando la calidad final. Si ésta no es aceptable, será necesario realizar nuevas adquisiciones de datos hasta que la calidad de éstos sea la adecuada. Una vez conseguidos, se definen los objetivos de la tercera fase que consisten en preparar la información, seleccionando primeramente las variables más importantes, filtrar y eliminar el ruido y transformar los datos, generando una base de datos óptima para la posterior etapa de modelado, cuarta fase.

Esta fase cubre todas las actividades de construcción del conjunto final de datos (datos entrada de los algoritmos de Minería de Datos), desde el conjunto inicial de datos. Es posible que estas actividades se tengan que realizar múltiples veces y sin orden determinado. Entre las tareas se destacan las de selección de tablas, atributos, registros, así como la de transformación y limpieza de los datos.

Esta fase consta de las siguientes etapas:

- Selección de los datos.
- Limpieza de los datos.
- Generación de variables adicionales.
- Integración de datos.
- Cambios de formato de los mismos.

Estas etapas **se realizarán repetidamente hasta que se obtenga una base de datos lo suficientemente adecuada para las posteriores fases de la metodología *CRISP-DM*.**

2.4.2 Selección de los Datos.

En esta primera etapa, se parte de los datos obtenidos en la fase anterior y de la descripción de los mismos. A partir de toda esta información, se realiza una selección de las variables más importantes según los siguientes requerimientos:

- Que sean lo más independientes entre si. Se eliminarán las variables muy dependientes de otras.
- Que describan casi completamente el sistema a describir o analizar.
- Que tengan una relevancia individual destacada. Se descartarán aquellas cuya influencia en el sistema sea nulo o muy escaso.
- Que estén exentas de ruido y con datos fiables.

Así mismo, si el volumen de datos es suficientemente grande, se decidirá si es necesario reducir el número de muestras.

Por ejemplo, un caso muy destacado aparece cuando decidimos dividir los datos en grupos, ya que habrá que homogeneizar la densidad de cada uno de ellos reduciendo los datos en aquellos grupos numerosos. Esto es necesario cuando se quiere modelar varios grupos y en uno de ellos tenemos una densidad mucho más elevada que en los demás, lo que puede influir negativamente en la creación de

un modelo que los clasifique, ya que éste dará más peso a los grupos con mayor densidad de individuos.

Ahora bien, recordando que los datos a analizar se encontrarán almacenados en la base de datos del CC, se puede concluir que serán de fácil acceso, además de que ya es conocido el formato y la organización que tendrán. Teniendo en cuenta los criterios de selección del proceso y los objetivos del negocio no sería difícil hacer una predicción. Es de intuir que lo más importante podría encontrarse en las relaciones que puedan surgir entre los clientes y los productos o servicios solicitados. De aquí que la selección estará basada en variables que nos brinden información sobre estas relaciones.

La base de datos de Clasificadores del servidor central del CC contiene toda la información necesaria en especial la tabla **Tbl_Usuario,** de los clientes. Como también en la base de datos de **Estadísticas&CRM** está almacenado todo aquello que relacione a los clientes y la información proporcionadas por proveedores de productos y servicios solicitados. El resto de los datos deberá ser analizado por métodos estadísticos en busca de variables de relevancia. A razón de la fase de implementación en la que se encuentra el CC necesitaremos especialistas en cada rama para que valoren que variables podría arrojar información importante. En fin, la selección estará basada en las pautas ya marcadas, siempre cumpliendo con los criterios relacionados.

2.4.3 Limpieza de los Datos.

Como se ha comentado en el punto anterior, las variables deben ser lo más fiables posibles.

Para ello habrá que:

- Tratar el ruido en los datos:
  - Corrigiendo, ignorando o eliminando aquellos datos con ruido.
  - Se estudiarán las posibles causas que generan ese ruido y la forma de resolverlo.
  - Se usarán técnicas de filtrado para mejorar la calidad de los datos.
- Tratar los espurios:
  - Analizándolos por separado y determinando las causas que los generaron.

- Eliminándolos.
- Clasificándolos en otros grupos.

- Tratar los valores incompletos:
  - Eliminándolos.
  - Ignorándolos.
  - Completándolos con técnicas estadísticas.
  - Rellenándolos con otras técnicas.

Dentro de los relacionados anteriormente podríamos mencionar tareas mucho más específicas

- Rellenar los valores nulos.
- Identificar los outliers (valores atípicos).
- Corregir los datos inconsistentes.

Dado el caso en cuestión debemos recordar la procedencia de los datos a reprocesar. Toda la información a preparar de una forma u otra tendrá como fuente el servidor central del CC. Esta condición hace bien bajas las probabilidades de aparición de outliers, valores inconsistentes, valores incompletos, ruido. No así pues con los valores nulos que por el contrario tendrán una alta probabilidad de aparición, por ello se habla un tanto más de su tratamiento.

2.4.3.1 Valores nulos.

El fenómeno de valores nulos puede aparecer con gran probabilidad. Los datos no siempre están totalmente disponibles. Los valores nulos se pueden deber a:

- Mal funcionamiento del equipo
- No se insertan por no entender el significado
- No se consideraron importantes en el momento de la recogida

Existen varias técnicas para el tratamiento de valores nulos.

- Ignorar la tupla. Esta opción tiene ciertas limitaciones. Si el porcentaje de valores nulos por atributo varia considerablemente se hace poco efectivo.
- Rellenar el valor manualmente. Es obvio que estamos tratando gran cantidad de información, el volumen de datos podrá ser lo suficientemente grande como para hacer esta tarea tediosa. Además de dudosa fiabilidad.
- Utilizar la media para rellenar todos los valores.
- Utilizar la media dentro de la clase.
- Utilizar el valor más probable mediante un árbol de decisión Bayes.

2.4.4 Generación de variables adicionales.

Se generarán nuevas variables a partir de las ya existentes siempre que permitan agilizar los estudios posteriores.

Dentro este proceso, se pueden incluir tareas de transformación de los datos como:

- Estandarizar o normalizar variables.
- Asignar pesos según la importancia de cada variable.
- Cambiar la codificación de alguna variable.
- Uso de transformadas.
- Uso de proyectores.
- Adición de nuevas variables a partir de otras.
- Uso de indicadores estadísticos.
- Uso de otros indicadores.

2.4.5 Integración de Orígenes de Datos.

Se combinarán los datos procedentes de diferentes orígenes, siempre que no se haya hecho ya, para obtener una base de datos más compacta y útil.

- Integración de datos:

- Combina datos de fuentes diversas

- Integración de esquemas
  - Integra metadatos de distintas fuentes
  - Problema de la identificación de entidades.
- Detección y resolución de los conflictos
  - Para la misma entidad los valores de diferentes fuentes son diferentes
  - Razones: distintas representaciones, métricas, escalas.

Esta etapa es obviada en nuestra fase puesto que la fuente de todos los datos es común.

2.4.6 Cambio de Formato en los Datos.

Se adecuará el formato de los datos para que puedan ser usados por las herramientas que se vayan a utilizar en fases posteriores. Por ejemplo:

- Cambiando el orden de las variables de un registro.
- Cambiando el tipo de variables. Convirtiendo variables numéricas a categóricas, o viceversa.
- Reordenando los datos.
- Etc.

2.5 MODELADO (FASE IV CRISP-MD)

En esta fase se seleccionan técnicas de minería y se aplican calibrando sus parámetros para conseguir los valores óptimos. Hay distintas técnicas para el mismo tipo de problema la diferencia muchas veces radica en los requisitos que han de cumplir los datos de entrada por ello a menudo es necesario volver a la fase de preparación de datos [38].

Las tareas propias de esta fase, constan de los siguientes pasos:

- Selección de las técnicas de modelado.
- Diseño del método de evaluación.
- Generación del modelo.

- Evaluación del modelo

2.5.1 Selección de la técnica de modelado

- El tipo de problema.
- Los datos a manejar.
- El tiempo necesario para obtener el modelo
- El conocimiento de la técnica.
- Las herramientas de que se disponen.

2.5.2 Diseño del Método de Evaluación

Es preciso antes de generar el modelo, definir el mecanismo de validación del mismo.

- Función que determine el error cometido y umbral de calidad estimado: error cuadrático medio.
- Tipo de método de evaluación de los modelos generados: validación cruzada.
- Si es necesario el tamaño de los grupos de validación y el tiempo de entrenamiento.

2.5.3 Generar un diseño de prueba

Una vez que se han precisado los detalles relativos a la generación de los modelos, se aplicaran las técnicas de modelado a los datos preparados.

- Parámetros elegidos, incluyendo su importancia, como son capaces estos de influir en los resultados del modelo y valores iniciales asignados.
- Modelo derivados, graficas de entrenamiento y validaciones, resultados numéricos.
- Descripción en detalle tanto de los modelos como de sus parámetros, verificar la exactitud y sensibilidad.

2.5.4 Evaluar el modelo

En este punto se verifica cuán eficiente fue cada modelo analizado, teniendo en cuenta:

- Nivel de eficacia de la predicción.

- Potencial de predicción de la información no conocida.
- Velocidad de procesado.
- Perfeccionamiento de los resultados con nuevos datos.
- Influencia de los parámetros en el modelo.

En caso de que los resultados obtenidos no sean consistentes, se repetirán todos los pasos hasta lograr una solución lo más óptima posible.

2.6 ANÁLISIS DEL DESCUBRIMIENTO DIRECTO.

Teniendo en cuenta que casi todos los objetivos definidos en el proyecto requieren de la confección de modelos para viabilizarse, es de suma importancia dejar claro que tipo de problemas requiere de este mecanismo de solvencia.

Estos tipos de problemas se ubican dentro del descubrimiento directo (predictivo). Su solución requiere explicar el valor de algún campo (precios, respuestas, valor) en términos del resto de los campos disponibles. Se selecciona el campo y se le pide al ordenador que prediga, estime, o clasifique el campo. Esta orientado a la meta de descubrir.

2.6.1 Descubrimiento directo (predictivo)

Su meta es automatizar el proceso de toma de decisión por medio de la construcción de un modelo que sea capaz de realizar una predicción ya sea asignando un elemento a una clase o realizando una estimación de un valor.

Ejemplos de descubrimiento directo:

- Estimar las ventas de la próxima semana.
- Que zona geográfica solicita más un servicio o producto.
- ¿Cual es el valor potencial de este nuevo cliente?
- Que clientes permanecerán fieles.

2.6.2 Procesos del descubrimiento directo.

1. Identificar las fuentes con datos pre-clasificados. (Se encuentra la clasificación basándose en casos ya clasificados)

2. Preparar los datos para el análisis

3. Seleccionar las técnicas dependiendo del tipo de datos y de la meta a encontrar

4. Dividir los datos en: entrenamiento, prueba y evaluación

5. Utilizar el conjunto de entrenamiento para construir **el modelo**:

- Elegir el algoritmo
- Elegir cuales son las variables de entrada
- Elegir la(s) variable(s) objetivo (salida)
- Determinar las variables que se ignoran.
- Establecer los parámetros del algoritmo

6. Mejorar el modelo mediante el conjunto de test.

7. Comprobar la exactitud del modelo aplicándolo al conjunto de evaluación.

8. Tomar una acción basándose en los resultados.

9. Medir el efecto de la acción tomada.

10. Usar los resultados para futuras acciones.

Para dar cumplimiento a la fase de **Modelado** de la metodología CRISP-MD es preciso trazarse objetivos, sobre los cuales se basará el descubrimiento directo. Una vez definidos estos objetivos en la fase de comprensión del problema solo resta definir los modelos.

Modelos formalizados:

- Clasificar clientes según criterios especificados (Clasificación).
- Clasificar la solicitud de un servicio o producto (Clasificación).
- Estimar las solicitudes de un producto en un intervalo de tiempo futuro dado (Estimación).

- Predecir una región que solicitará más un producto o servicio (Predicción).

Para el descubrimiento directo contamos con las herramientas siguientes:

- Clasificación y Estimación
  - Árboles de inducción:
    - ID3, C4.5.
  - Redes neuronales
    - Propagación hacia atrás.
- Predicción
  - Regresión lineal y múltiple.
  - Regresión no lineal.

2.6.2.1 Clasificación y Estimación

Muy similar al proceso de aprendizaje humano, utiliza observaciones para elaborar modelos. Consiste en analizar un conjunto de datos para determinar las características de los mismos, basado en un conjunto de restricciones. Esta técnica se ubica dentro del aprendizaje supervisado. El modelo se construye a partir de datos correctamente clasificados de antemano, estos son desarrollados en dos fases: **entrenamiento** y **prueba** [38].

Objetivos

Obtener modelos que discrimine las instancias de entrada en diferentes clases de equivalencia por medio de los valores de diferentes atributos.

Requisitos

- Suministrar el atributo decisión o clase (**label**). El conjunto de valores de este atributo debe ser finito y no excesivamente grande.
- Suministrar los atributos **condición**.

- Podría requerir datos que no sean numéricos pero existen variedades que tratan con datos numéricos.
- Número máximo de precondiciones.
- Soporte mínimo de las reglas.

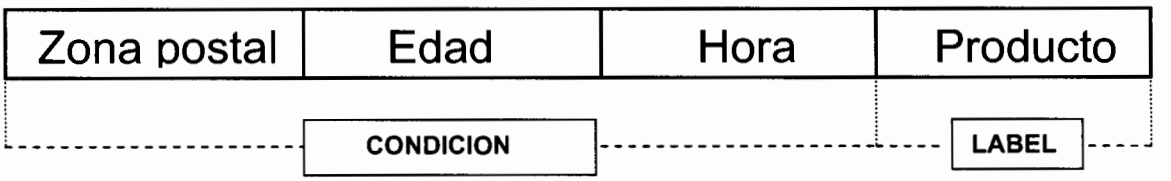


Figura 6.

Como se ha mencionado antes la particularidad del CC de brindar servicios a proveedores diferentes no nos permite establecer los atributos **label** y **condición**, aunque los intereses de los proveedores fuesen los mismos. Necesitando para cada uno de ellos criterios de expertos para concretar estos atributos.

Entrada de los algoritmos

- Atributos decisión o label: Atributos usados para construir las clases de equivalencia en los métodos supervisados (una clase por cada valor o combinación de valores de dichos atributos).
- Atributos condición: Atributos usados para describir por medio del proceso de inducción las clases de equivalencia.

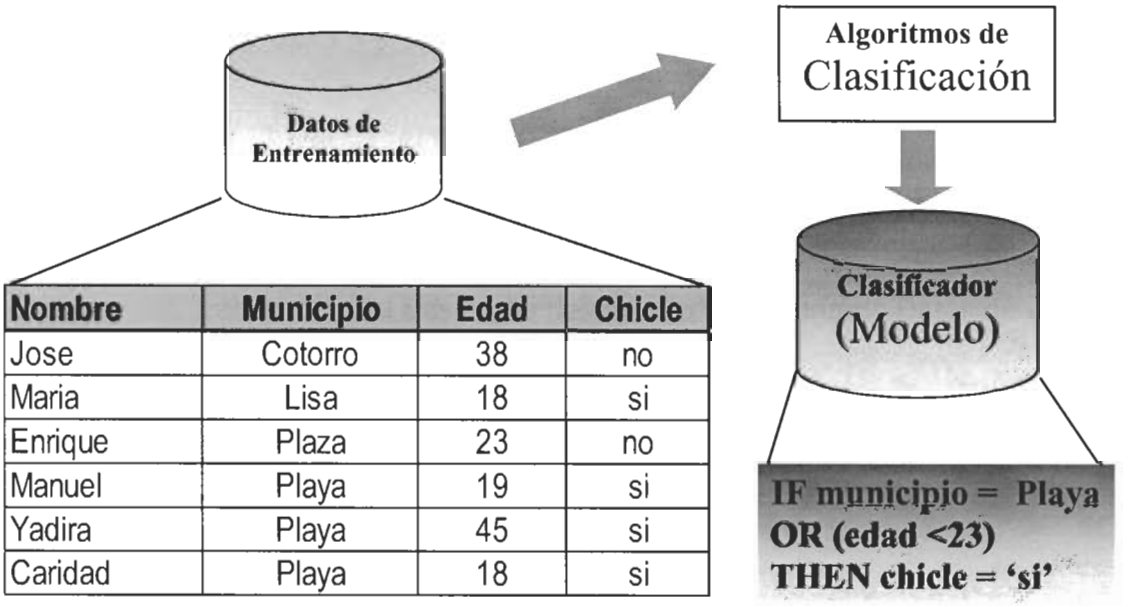
Construcción del modelo

- Describir un conjunto de datos en base a una característica.
- Cada tupla pertenece a una clase predefinida determinada por el atributo de condición.
- Se utiliza el conjunto de datos de entrenamiento.
- El modelo se representa a través de reglas de clasificación, árboles de decisión o mediante formulas matemáticas.

Ejemplo de la construcción de un modelo:

Condición: **Nombre, Municipio, Edad.**

Label: **Chicle.**



Nombre	Municipio	Edad	Chicle
Jose	Cotorro	38	no
Maria	Lisa	18	si
Enrique	Plaza	23	no
Manuel	Playa	19	si
Yadira	Playa	45	si
Caridad	Playa	18	si

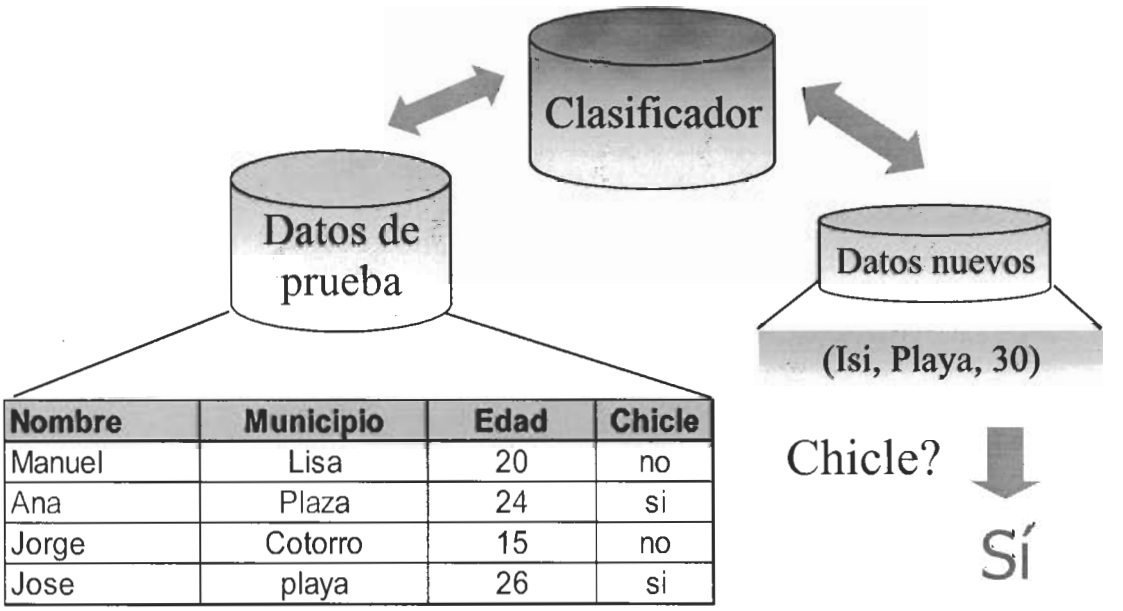
Figura 7.

Utilización del modelo

Para clasificar objetos nuevos de los que se desconoce su clase

- Determinación de la precisión del modelo.
- Se utiliza el modelo para clasificar el conjunto de datos de entrenamiento y se compara el resultado con la etiqueta original.
- La exactitud es el porcentaje de conjunto de datos de prueba que son clasificados correctamente por el modelo.
- El conjunto de datos entrenamiento y el conjunto de datos de prueba deben de ser disjuntos, para evitar el *over fitting* (sobre ajuste).

Aquí se puede observar como se puede emplear un modelo ya entrenado para clasificar nuevos datos.



Nombre	Municipio	Edad	Chicle
Manuel	Lisa	20	no
Ana	Plaza	24	si
Jorge	Cotorro	15	no
Jose	playa	26	si

Figura 8.

Representación del error

La matriz de confusión permite el análisis de los resultados obtenidos de la clasificación. Para la construcción de esta se utiliza una matriz la cual representa en forma de tabla del número de instancias clasificadas correctamente, haciendo posible la percepción del porcentaje de eficiencia de algoritmo empleado y el rango de error del mismo.

b <-- classified as

7 2 | a = yes

3 2 | b = no

Como se puede ver la matriz clasifica 7 prototipos de la clase yes cuya clase cierta es yes, y 2 prototipos de la clase no, cuya clase cierta es no.

Lo que indica que este algoritmo tuvo un nivel muy bajo de error, y un alto nivel de certeza.

Técnicas

Técnicas simbólicas (Árboles de inducción, ID3 y C4.5):

Tanto el ID3 como el C4.5 generan árboles y reglas de decisión a partir de datos preclasificados. Para construir los árboles se utiliza el método de aprendizaje "divide y reinarás", que particiona el conjunto de ejemplos en subconjuntos a medida que avanza; trabajar sobre cada subconjunto es más sencillo que trabajar sobre el total de los datos. Los nodos representan la verificación de una condición sobre un atributo, las ramas representan el valor de la condición comprobada en el nodo del cual derivan, Los nodos hoja representan las etiquetas de clase.

- Muy eficientes en tiempo de proceso.
- Resultados intuitivos.
- Particiones lineales.
- Algunos presentan problemas con variables continuas.

ID3: EL ID3 o Induction Decision Trees, desarrollado en los años ochenta por Quinlan, es un sistema de aprendizaje supervisado que construye árboles de decisión a partir de un conjunto de ejemplos. Estos ejemplos o tuplas están constituidos por un conjunto de atributos y un clasificador o clase. Los dominios de los atributos y de las clases deben ser discretos. Además, las clases deben ser disjuntas. Las primeras versiones del ID3 generaban descripciones únicamente para dos clases, como ser positiva y negativa. En las versiones posteriores, se eliminó esta restricción, pero se mantuvo la restricción de clases disjuntas. El ID3 genera descripciones que clasifican a cada uno de los ejemplos del conjunto de entrenamiento. Este sistema tiene una buena performance en un amplio rango de aplicaciones de diversos dominios, como el dominio médico, el artificial y el análisis de juegos de ajedrez. El nivel de precisión en la clasificación generalmente es alto. Sin embargo, el sistema tiene algunas desventajas. Recordemos que los atributos y las clases deben ser discretos y no pueden ser continuos. Además, aún cuando se cuente con conocimientos de dominio o conocimientos previos, el sistema no hace uso de ellos. A veces, los árboles son demasiado frondosos, lo cual conlleva una difícil interpretación. En esos casos pueden ser transformados en reglas de decisión para hacerlos más comprensibles.

C4.5: El C4.5 es una extensión del ID3 que acaba con muchas de sus limitaciones. Por ejemplo, permite trabajar con valores continuos para los atributos, separando los posibles resultados en dos ramas: una para aquellos **Ai<=N** y otra para **Ai>N**. Además, los árboles son menos frondosos porque cada hoja no cubre una clase en particular sino una distribución de clases, lo cual los hace menos profundos y menos frondosos. Este algoritmo fue propuesto por Quinlan en 1993. El C4.5 genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente, según la estrategia de profundidad-primero *(depth-first)*. Antes de cada partición de datos, el algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información o en la mayor proporción de ganancia de información. Para cada atributo discreto, se considera una prueba con n resultados, siendo n el número de valores posibles que puede tomar el atributo. Para cada atributo continuo, se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en los datos.

Ganancia de información: Seleccionar el atributo con mayor ganancia de información.

Si hay dos clases, **P** y **N.**

- Sea el conjunto de ejemplo S que contiene p elementos de la clase P y n elementos de la clase N.
- La cantidad de información, que se necesita para decidir si una muestra cualquiera de S pertenece a P o a N se define como:

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Si se utiliza un atributo A, un conjunto S se dividirá en conjuntos {S1, S2,..., Sv}

(Si) contiene (**pi**) ejemplos de **P** y (**ni**) ejemplos de **N**, la entropía, o la información necesaria para clasificar objetos en todos los subárboles **(Si)** es:

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

La ganancia de información de la rama **A** es:

$$Gain(A) = I(p,n) - E(A)$$

Evitar el "over fitting"

Una vez construido el árbol es de suma importancia identificar y eliminar las ramas que presentan ruidos (Poda del Árbol). El árbol generado es posible que sea muy exacto para el conjunto de entrenamiento. Para evitar esta anomalía se le da dos vías de solución al problema:

- Evitar el crecimiento (Prepruning): no se divide un nodo si ello supone que la medida de bondad caiga por debajo de un umbral
  - Dificultad de elegir el umbral
- Podar después de formar el árbol (Postpruning): Eliminar las ramas de un árbol una vez que se ha generado por completo.
  - Utilizar un conjunto de datos diferente al de entrenamiento para decidir que ramas hay que podar.

Redes neuronales (Redes de retropropagación)

Esta tecnología puede ser desarrollada en software o en hardware y es capaz de aprender y predecir. Es adecuada para problemas que hasta ahora eran resueltos sólo por el cerebro humano y difícil para las máquinas lógicas secuenciales. Pueden ser combinadas con otras herramientas de la Inteligencia Artificial (IA), tal como la lógica difusa (lógica fuzzy), los algoritmos genéticos y los sistemas expertos. Al hablar de redes de retropropagación o redes de propagación hacia atrás hacemos referencia a un algoritmo de aprendizaje más que a una arquitectura determinada. La retropropagación consiste en comparar la salida real con la salida deseada. La diferencia entre ambas constituye un error que se propaga hacia atrás desde la capa de salida hasta la de entrada permitiendo así la adaptación de los pesos de las neuronas intermedias mediante una regla de aprendizaje Delta. Sin embargo, también tiene sus limitaciones.

- Ventajas
  - La exactitud es generalmente alta.
  - Robusto, trabaja bien incluso cuando los datos contienen errores.
  - La salida puede ser discreta, valor real, un vector de valores reales.

Evaluación rápida de la función aprendida.

- Crítica

Largo tiempo de entrenamiento.

Dificultad de entender la función aprendida.

El método de aprendizaje de la red empleado es supervisado.

Aprendizaje supervisado: consiste en introducir una serie de patrones de entrada a la red y a su vez mostrar la salida que se quiere tener. La red es capaz de ajustar los pesos de las neuronas de forma que a la presentación posterior de esos patrones de entrada la red responde con salida memorizada.

La entrada *(x)* se transforma en *(y)* por medio de una función no lineal

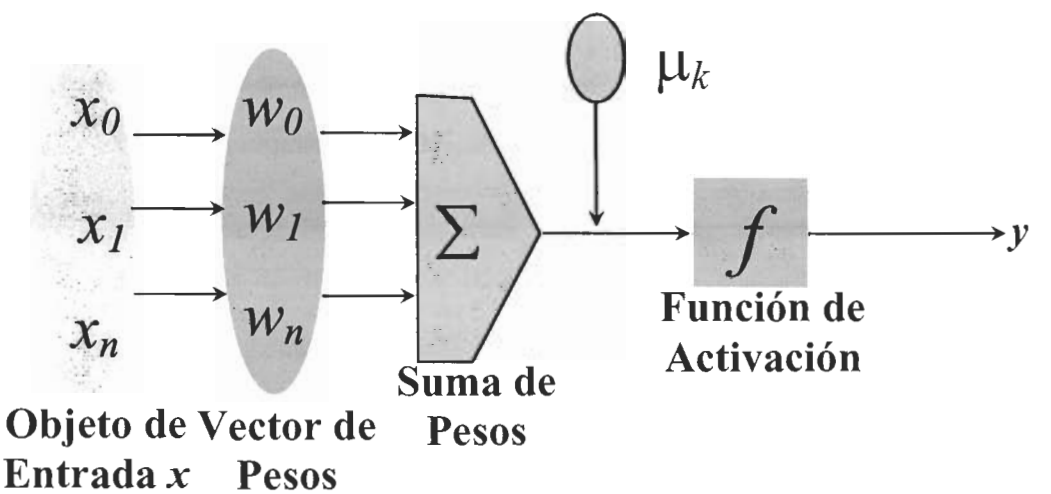


Figura 9.

Valoración

El algoritmo de retropropagación presenta ciertos problemas, algunos referentes a su dudosa plausibilidad neurofisiológica, y otros referentes a ciertos aspectos computacionales, que son los que vamos a comentar aquí [57].

- Los resultados dependen de los valores iniciales, aleatorios, de las conexiones. Esto hace que sea conveniente entrenar varias redes con distintas valores iniciales y elegir la que mejor funcione.
- A veces se requiere mucho tiempo para obtener soluciones sencillas. Este problema se reduce gracias al aumento de potencia de los procesadores y al uso de nuevas tecnologías, sin embargo, el tiempo de cómputo aumenta mucho al aumentar el tamaño de la red. Si bien el volumen de cálculo es proporcional al número total de conexiones. En la práctica, al aumentar el tamaño de la red, hacen falta más ejemplos de aprendizaje, y eso provoca un aumento mucho mayor del tiempo de aprendizaje. Para incrementar la velocidad de convergencia se han desarrollado diferentes modificaciones del algoritmo.
- La "interferencia catastrófica" o empeoramiento en el rendimiento del sistema, como consecuencia de la incorporación de nuevos ejemplos de aprendizaje.
- La parálisis: esto sucede cuando los pesos quedan ajustados a valores muy grandes, esto hace operar a las unidades de proceso con una activación muy próxima a 1, y por lo tanto, el gradiente del error, tiende a 0, en consecuencia no se producen modificaciones en los pesos, el aprendizaje queda detenido. Por eso es conveniente aleatorizar los pesos de las conexiones con valores pequeños y usar la tasa de aprendizaje, también pequeña, a pesar de que se haga más lento el aprendizaje.
- Inestabilidad temporal. Si usamos un coeficiente de aprendizaje elevado, se van a producir incrementos grandes en los pesos, de manera que es fácil pasarse de incremento y tener que tratar de compensarlo en el siguiente ciclo, de manera que se producirían oscilaciones continuas. Esto se soluciona usando un coeficiente pequeño, o, para no tener un aprendizaje muy lento, modificar dicho coeficiente adaptativamente (aumentarlo si el error global disminuye, y disminuirlo en caso contrario).
- El problema de los mínimos locales. El algoritmo de retropropagación usa una técnica por gradiente descendiente, esto significa que sigue la "superficie del error" siempre hacia abajo, hasta alcanzar un mínimo local, pero no garantiza que se alcance una solución globalmente óptima. Sin embargo, se ha comprobado que el hecho de alcanzar mínimos locales no impide

que se consigan resultados satisfactorios. Por otro lado, se han desarrollado métodos para solventar este problema, como el modo de operación asíncrona o probabilística y el uso de métodos estadísticos, como el equilibrio termodinámico simulado (ver siguiente apartado).

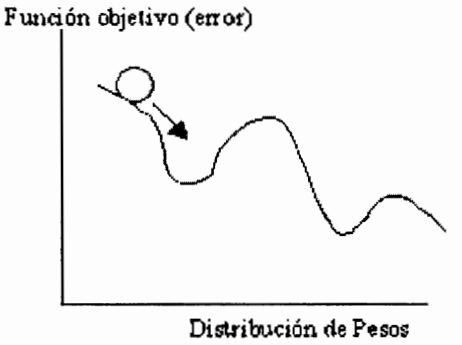


Figura 10. Problema de los mínimos locales.

- Podemos considerar el error como una superficie llena de desniveles, si soltamos una pelota caerá en algún valle, pero no necesariamente en el más hondo, sino en el más cercano (un mínimo local). Una idea intuitiva para solucionar esto, sería aplicarle cierta energía a esa superficie agitándola o haciéndola vibrar, esto haría saltar a la pelota de valle en valle, como de los valles más profundos es más difícil salir, tendería a estar en valles cada vez más profundos. Si dejamos de agitar esa superficie poco a poco, al final tendremos la pelota en el valle más profundo de la superficie.
- Otras técnicas que pueden ayudar a no caer en mínimos locales consisten en añadir cierto nivel de ruido a las modificaciones de los pesos de las conexiones. Otra medida propuesta es añadir ruido a las conexiones, pero esto es más útil para darle robustez y aumentar la capacidad de generalización de la red. Estas medidas, por contra, aumentan el tiempo de aprendizaje.

2.6.2.2 Predicción de valores

En múltiples ocasiones en la práctica nos encontramos con situaciones en las que se requiere analizar la relación entre variables cuantitativas y cualitativas. Los dos objetivos fundamentales de

este análisis serán, por un lado, determinar si dichas variables están asociadas y en qué sentido se da dicha asociación (es decir, si los valores de una de las variables tienden a aumentar –o disminuir- al aumentar los valores de las demás); y por otro, estudiar si los valores de algunas variable pueden ser utilizados para predecir el valor de otra.

La predicción es similar a la clasificación a diferencia que en vez de predecir clases de valores categóricos en esta los modelos de predicción son funciones continuas, primeramente construye un modelo y luego este predice el valor desconocido. Para la construcción del modelo se utilizan las siguientes herramientas:

Predicción **en regresión lineal simple** [58].

$Y = \alpha + \beta X$, los parámetros (α y β) especifican la línea y se estiman utilizando los datos. Se quiere predecir el valor de la variable aleatoria ($Y/X = x_t$) teniendo en cuenta que se ha ajustado una recta de regresión. Ahora se quiere predecir el resultado de una variable aleatoria. El predictor que se utiliza ($\hat{y}_t$) se obtiene como aquel que minimiza el Error Cuadrático Medio de Predicción. Esto es, ($\hat{y}_t$) se obtiene como el valor que minimiza la siguiente función:

$$\Psi(z) = \min_{z} E\left((z - Y_t)^2\right).$$

Al resolver este problema de minimización se obtiene como preeditor el resultado de sustituir el valor de (x_t) en la recta de regresión calculada:

$$\hat{y}_t = \hat{\alpha}_0 + \hat{\alpha}_1 x_t = \bar{y} + \hat{\alpha}_1 (x_t - \bar{x}).$$

Por tanto, en la predicción de ($Y/X = x_t$) aumenta la varianza ya que la variabilidad debida a la muestra $(Var(\hat{m}_t))$ se incrementa con la variabilidad propia de la variable aleatoria que se quiere predecir $(Var(y_t))$. Ahora la varianza de la predicción es

$$\text{Var}(\hat{y}_t - y_t) = \sigma^2 + \frac{\sigma^2}{n_t} = \sigma^2\left(1 + \frac{1}{n_t}\right)$$

Por la hipótesis de normalidad y razonando como en el apartado anterior se obtiene

$$\frac{\hat{y}_t - y_t}{\sigma\sqrt{1 + h_{tt}}} \sim N(0,1) \Rightarrow \frac{\hat{y}_t - y_t}{\hat{s}_R\sqrt{1 + h_{tt}}} \sim t_{n-2}.$$

Utilizando esta distribución se puede calcular un "intervalo de predicción" para (y_t,) con un nivel de confianza α, de la siguiente forma

$$y_t \in \hat{y}_t \mp \hat{s}_R \sqrt{1 + h_{tt}}\, t_{n-2}\left(1 - \frac{\alpha}{2}\right)$$

Predicción **en regresión lineal** múltiple [58]

Análogamente al caso de regresión lineal simple la predicción en la múltiple se ajusta un modelo de regresión lineal de la variable Y respecto al vector de variables regresoras $\vec{\mathbf{X}}$. Con esta herramienta se pueden responder preguntas como: conociendo que un determinado árbol tiene un diámetro 10 u. y una altura de 80 u. ¿qué volumen se predice para este árbol?"

El predictor $(\hat{y}_h)$ que minimiza el Error Cuadrático Medio de Predicción viene dado por, E $\left((\hat{y}_t - y_t)^2\right)$

$$\hat{y}_t = \vec{\mathbf{x}}_t^t \hat{\alpha}$$

El predictor $\hat{y}_t$ verifica las siguientes propiedades:

1. La predicción es centrada, ya que, $E(\hat{y}_t) = E(Y_t)$
2. La varianza de la predicción es,

$$\begin{aligned} Var(\hat{y}_t - y_t) &= E\left((\hat{y}_t - y_t)^2\right) = Var(y_t) + Var(\hat{m}_t) \Rightarrow \\ Var(\hat{y}_t - y_t) &= \sigma^2 + \sigma^2 h_{tt} = \sigma^2 (1 + h_{tt.}) \end{aligned}$$

3. Para calcular intervalos de predicción de y_t se utilizará el siguiente estadístico pivote

$$\frac{\hat{y}_t - y_t}{\hat{s}_R \sqrt{1 + h_{tt}}} \sim t_{n-(k+1)}$$

4. Un intervalo de predicción de y_t con nivel de confianza α viene dado por

$$y_t \in \hat{y}_t \mp \hat{s}_R \sqrt{1 + h_{tt}}\, t_{n-(k+1)}\left(1 - \frac{\alpha}{2}\right)$$

Modelos Log-lineal [59]

La Regresión Loglineal, es un método estadístico cuyo objetivo consiste en estudiar la "Clasificación" de las Variables Cualitativas. Es esencialmente un Modelo de Regresión Lineal Múltiple entre las Variables Cualitativas y el Logaritmo Neperiano de la Frecuencia de los datos (referenciales), de la forma:

$$Ln(frecuencia) = \mu + \lambda^{A} + \lambda^{B} + \lambda^{C} + \lambda^{AxB} + \lambda^{AxC} + \lambda^{BxC} + \lambda^{AxBxC}$$

Donde A, B y C; son Variables Cualitativas.

Se define como "Tabla de Contingencia" (Crosstabulation Tables), a una combinación de dos o más tablas de distribución de frecuencia, arregladas de manera que cada celda o casilla de la Tabla resultante represente una única combinación de las "variables cruzadas (crosstabuled)". De tal manera que la "Tabla de Contingencia" nos permita examinar las frecuencias observadas que pertenecen a cada una de las combinaciones específicas de dos o más variables.

La Bondad de Ajuste:

La bondad de ajuste de una Regresión Loglineal, se basa en cuan significativo es la desviación (residuo) entre la Frecuencia Observada de los datos y la Frecuencia Esperada que genera el modelo loglineal.

Es decir, el modelo será mejor en función de la minimización de la diferencia entre la Frecuencia Observada y la Esperada.

$$FREC_{OBSERV} - FREC_{ESPERADA} \Rightarrow 0$$

Se evaluará las nivel de significado de (p) o "Bondad de Ajuste" de un Modelo Loglineal particular, mediante: El Test del Chi Cuadrado (X^2) Tradicional y Estadístico de Máxima Verosimilitud de Pearson (L^2) o (Pearson Likelihood Ratio Chi-square, como es su denominación en inglés).

De tal manera, que se cumplan los siguientes parámetros:

1. Chi Cuadrado (X^2): Máximo
2. Estadístico de Máxima Verosimilitud de Pearson (L^2): Máximo

3. Nivel de significado de (Sig.): Mínima ($P < 0.0001$)

El Modelo Loglineal Saturado

El Análisis o Regresión Loglineal, analiza el Logaritmo Neperiano (Ln) de la Frecuencia de cada celda o casilla de una Tabla de Contingencia, por medio de un modelo lineal. Por lo tanto, el Ln de la frecuencia de cada celda o casilla se puede expresar como la suma de las contribuciones de las diferentes variables que intervienen en la formación del Modelo Loglineal.

Se define como Modelo Saturado (o Completo) a aquel que contiene TODOS los posibles efectos principales y TODAS las posibles combinaciones (efectos de 2do., 3er. o enésimo orden) de las Variables seleccionadas que lo componen.

Debido a que el Modelo Loglineal Saturado, puede reproducir perfectamente la data estudiada, debido a que contiene todas las posibles combinaciones de las variables seleccionadas; se supone en un modelo pesado y complejo, y usualmente no es el modelo más deseable. Por un principio elemental de parsimonia, se debe encontrar uno o más modelos más simples, que generen un resultado con un grado aceptable de precisión y los definimos como "Modelos Jerárquicos".

Por lo tanto, es necesaria la búsqueda de uno o varios modelos mas simples que den cuenta de dichas frecuencias con un grado de precisión aceptable para un nivel dado de confianza. Cuando se analizan Tablas de Contingencias de Cuarto Orden o mayor, la determinación del mejor modelo de Regresión Loglineal puede resultar altamente dificultosa. Aquí entraría la búsqueda de una Modelo de Correlación más simple.

Los Modelos Loglineal Jerárquicos

Se define como Modelos Loglineal Jerárquicos, a los diferentes modelos, todos sub-juegos (ecuaciones de menor orden que el Modelo Saturado) provenientes del Modelo Loglineal Saturado, que cumplan las condiciones siguientes:

a) Si un parámetro es nulo, también los serán aquellos términos de orden inferior.

b) Que exista completa independencia entre las variables seleccionadas

Si estas condiciones se cumplen, se genera un Modelo Loglineal más sencillo, más elegante y con un grado aceptable de precisión

El Método de búsqueda del Mejor Modelo Jerárquico más utilizado por los paquetes estadísticos dedicados en el conocido como "Retro-eliminación" (Backward Elimination). Esta metodología combina el uso de los **k** ésimos ordenes y el test Chi - cuadrado para encontrar un Modelo Jerárquico o varios Modelos Jerárquicos significativos.

La lógica del proceso es la siguiente:

a) Se comienza calculando el Modelo Saturado.

b) Se analiza el Modelo Jerárquico o los Modelos Jerárquicos de más alto orden

c) Se elimina el Modelo o Modelos de ese orden que no sean significativos ($\chi^2_i \leq \chi^2_0 \quad y \quad p > 0.01$)

d) Se eliminan los Modelos Jerárquicos de Orden Inferior en las mismas variables

e) Se analizan los Modelos Jerárquicos restantes

f) El proceso se continúa hasta el punto en que no puedan seguir eliminándose más efectos sin sacrificar el poder predictivo del modelo ($X^{2\,(y/0)}\ P$) permanezcan constantes o tiendan a disminuir χ^2 o aumentar p.

2.7 ANÁLISIS DEL DESCUBRIMIENTO INDIRECTO.

El resto de los objetivos que quedan fuera de descubrimiento directo pertenece al indirecto. Este tipo de descubrimiento responde a obtener una mejor comprensión de lo que ocurre en los datos y como consecuencia del mundo que reflejan.

2.7.1 Descubrimiento indirecto (Descriptivo):

No hay campo objetivo. Simplemente se le pide a los ordenadores que identifiquen patrones en los datos que sean significativos.

- ¿Qué productos se compran juntos?

- ¿Qué mezclas de colores deben ir juntas en un almacén?
- Segmentar la cartera de clientes

Este tipo de descubrimiento se divide en dos ramas:

- Segmentación de bases de datos
  - Clustering demográfico.
  - Algoritmo de las K-medias.
  - Mapas de Kohonen.
- Análisis de asociaciones y /o Patrones secuenciales
  - Matrices de concurrencias.
  - Algoritmo A priori.

2.7.2 El proceso del descubrimiento indirecto.

- Identificar las fuentes de datos disponibles
- Preparar los datos para el análisis
- Seleccionar la técnica dependiendo de la meta perseguida y de los datos de estudio:
  - Seleccionar el algoritmo que se usará
  - Establecer los parámetros del algoritmo
  - Definir las variables que se usan (no hay distinción entre entrada y salida)
- Utilizar la técnica seleccionada (algoritmo) para descubrir la estructura oculta en los datos.
- Analizar los resultados:
  - Utilizar variables que fueron ignoradas para ver su distribución en los resultados
  - Ejecutar técnicas de descubrimiento directo para explicar los resultados
- Generar nuevas hipótesis.

2.7.2.1 Segmentación de bases de datos

En esta rama se persigue como objetivo dividir en segmentos de registros similares. Se pretende que dado un conjunto de datos el algoritmo dé como resultado una división de la población de manera que se minimiza la distancia de los elementos de cada grupo o "cluster" y se maximiza la distancia entre clases.

Requisitos de la Segmentación:

- Número máximo de clusters.
- Número de iteraciones.
- Número mínimo de elementos en cada cluster.

Clustering

Cluster: Colección de objetos.

Un buen método de clustering producirá clusters de calidad:

- Alta similitud dentro de cada clase.
- Baja similitud de elementos de distintas clases.

La calidad de los resultados depende de la medida de similitud que se utilice y en su implementación otro indicador de calidad es la capacidad de descubrir patrones ocultos.

Medidas de calidad del cluster

- Similitud: Se expresa en término de una función de distancia: d(i, j).
- También se tiene una medida de calidad que mide la bondad del cluster.
- Las definiciones de funciones de distancia difieren dependiendo del tipo de variables.

Tipos de datos posibles

- Variables de intervalo.
- Variables binarias.
- Nominales, ordinales.

Variables de intervalo

Estandarizar los datos:

- Calcular la desviación respecto de la media

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$$

- Donde $$m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf})$$

- Calcular la medida estándar $$z_{if} = \frac{x_{if} - m_f}{s_f}$$

Usar la desviación con respecto a la media es más robusto que con respecto a la desviación estándar.

Similitud entre objetos:

- Las distancias se utilizan para medir la similitud de dos objetos:

- La distancia Minkowski se define:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + ... + |x_{i_p} - x_{j_p}|^q)}$$

Donde i = (xi1, xi2,..., xip) y j = (xj1, xj2,..., xjp) son dos objetos con p atributos y q es un entero positivo

- Si q = 1, d es la distancia de Maniatan:

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + ... + |x_{i_p} - x_{j_p}|$$

- Si q = 2, d es la distancia Euclídea:

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2)}$$

Variables Binarias

Para los datos binarios se emplea una tabla de contingencia

		Objeto *j*		
		1	0	*sum*
Objeto *i*	1	*a*	*b*	*a+b*
	0	*c*	*d*	*c+d*
	sum	*a+c*	*b+d*	*p*

Coeficiente invariante si la variable es simétrica:

$$d(i,j)=\frac{b+c}{a+b+c+d}$$

Variables Nominales

Una generalización de las variables binarias donde se toman más de 2 estados.

Método 1: Aplicación Simple

m: numero de valores iguales, **p:** total de números de variables.

$$d(i,j)=\frac{p-m}{p}$$

Método 2: Utilizar un número más grande de variables binarias creando una variable binaria para cada uno de los estados nominales.

Pueden ser discretas o continuas. El orden es importante $r_{if} \in \{1,...,M_f\}$

Llevar el rango de la variable a [0, 1] reemplazando el objeto i de la variable f mediante la fórmula:

$$z_{if}=\frac{r_{if}-1}{M_f-1}$$

Calcular la similitud mediante los métodos para variables de intervalos.

Tipos de algoritmos para la segmentación de bases de datos:

- Algoritmos divisivos: Construyen varias particiones y luego las evalúan.
- Jerárquicos: Crean una descomposición jerárquica.
- Basados en Densidad: se utilizan funciones de densidad.

Algoritmos divisivos

Estos algoritmos se caracterizan por segmentar la base de datos en *K* clusters, de esta familia contamos con dos de estos.

- k-means (MacQueen'67): Cada cluster se representa por el centro del cluster.
- k-medoids (Kaufman & Rousseeuw'87): Cada cluster se representa por uno de los objetos del cluster.

El método de las *K-Means:*

Partiendo de conocer el atributo *K* el algoritmo se implementa en cuatro pasos.

- Dividir los objetos en K subconjuntos no vacíos
- Calcular la semilla como el centroide (punto medio) del cluster.
- Asignar cada objeto al cluster más cercano.
- Ir al paso 2, hasta que no se puedan hacer más asignaciones.

Ventajas

- Relativamente eficiente
- Generalmente termina con un óptimo local.

Deficiencias

- Solo es aplicable cuando la media está definida. ¿datos categóricos?
- Se necesita especificar K de antemano.
- No es capaz de tratar con ruido.

- No es apropiado para descubrir cluster que no tengan formas no convexas

Variantes

Las variantes se diferencian por:

- Selección de las k medias iniciales.
- Cálculo de similitudes.
- Estrategias para calcular las medias.

Tratamiento de datos categóricos

- Reemplazar las medias por las modas.
- Usar las medidas de similitud para los objetos categóricos.

El método de los K-medoides:

Este método en esencia cuenta con dos pasos, en primer lugar se buscan los elementos más representativos hasta formar los clusters, posteriormente iterativamente se reemplaza uno de estos elementos por uno que no lo es (representativo) si ello mejora la calidad del clusters obtenido.

Este método es eficiente para conjuntos de datos pequeños.

Algoritmos jerárquicos

Utilizan la matriz de distancias como criterio. No requiere el número de cluster como entrada pero necesita una condición de terminación.

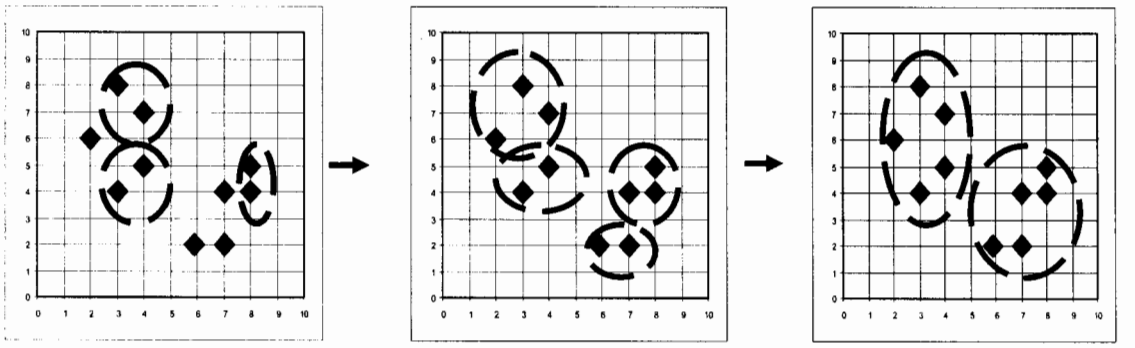


*Figura 11.**AGNES** (Agglomerative Nesting): Este algoritmo unifica los nodos que tengan una menor disimilaridad*

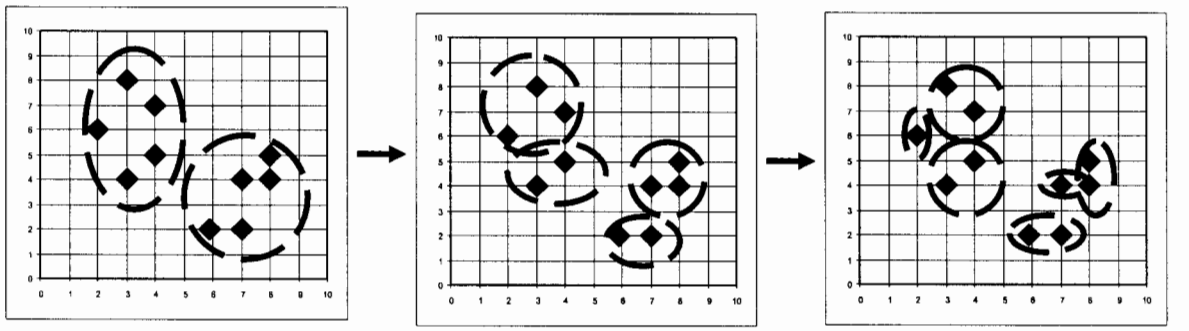


*Figura 12. **DIANA** (Divisive Analysis): Orden inverso de AGNES.*

Estos algoritmos pueden ser encontrados fácilmente en cualquier paquete estadístico.

De los métodos jerárquicos podemos concluir que:

- Su mayor desventaja es que no escalan bien (aglomerativos).
- Se pueden integrar con lo métodos basados en distancias (**BIRCH** (1996), **CURE** (1998)).

Algoritmos basados en la densidad

Esta familia de algoritmo se caracteriza por descubrir clusters de formas arbitrarias, tratan el ruido, obtienen los clusters en una sola pasada aunque dependen de parámetros como condición de parada.

- DBSCAN: Ester, et al. (KDD'96)
- OPTICS: Ankerst, et al (SIGMOD'99).
- DENCLUE: Hinneburg & D. Keim (KDD'98)

- CLIQUE: Agrawal, et al. (SIGMOD'98)

Parámetros de los métodos basados en densidad

- Eps: radio máximo de la vecindad.
- MinPts: número mínimo de puntos en el vecindario de ese punto
- NEps(p): {q pertenecientes a D | dist(p,q) <= Eps}.
- Un punto p es alcanzable directamente desde q con Eps, MinPts si
  - p pertenece NEps(q)
  - |NEps (q)| >= MinPts
- Un punto p es Densidad-alcanzable desde un punto q si hay una cadena de puntos p1,..., pn, p1 = q, pn = p tales que pi+1 es directamente alcanzable desde pi .
- Un punto p es Densidad-conectado con q si hay un punto o tal que, p y q son densidad alcanzables desde o.

Redes de Kohonen (Kohonen '95)

Pertenece a la categoría de las redes competitivas o mapas de autoorganización, es decir, con aprendizaje no supervisado de tipo competitivo. Poseen una arquitectura de dos capas (entrada-salida) (una sola capa de conexiones), funciones de activación lineales y flujo de información unidireccional (son redes en cascada) [60].

Este algoritmo dibuja un mapa bidimensional (Figura 13), sobre el cual localiza las instancias agrupadas por conjuntos. Como ventajas podemos resaltar que la representación grafica de los resultados es intuitiva, funciona de forma robusta con cualquier tipo de atributos. Como deficiencia se le puede señalar que los dos ejes descritos por el gráfico representan funciones complejas. La red mapea el espacio de entrada hacia un espacio de salida con cierto orden topológico, Kohonen propone un método para que este orden se conserve al entrenar la red, la clave está en reducir el tamaño del vecindario de la unidad ganadora en cada iteración.

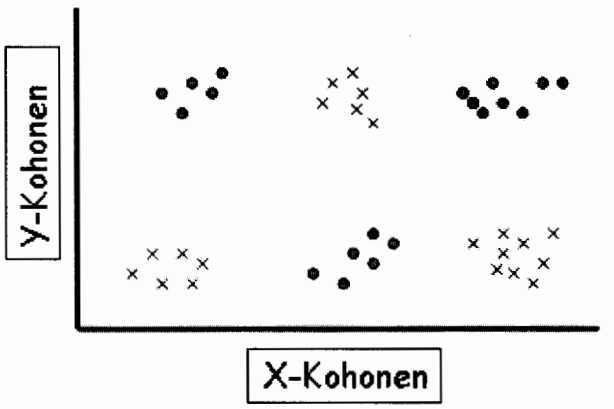


Figura 13.

Resumen

Luego de haber descrito una de las ramas del descubrimiento indirecto (Segmentación de bases de datos) y las técnicas que permiten su solvencia podemos centrarnos en la utilización de estas herramientas para darle cumplimiento a los objetivos de este trabajo. Dadas de las bondades de la segmentación de bases de datos es fácil percatarse que más a los objetivos, esta contribuye al preprocesado de otros algoritmos, en este caso los de descubrimiento directo. Es preciso aclarar que la bondad de conocer más los datos y escudriñar información oculta no es de mucha utilidad en este proyecto dado por el carácter empresarial del proyecto no investigativo.

2.7.2.2 Análisis de Asociaciones.

En esta rama del descubrimiento indirecto las acciones se centran en dar cumplimiento a establecer vínculos entre los registros, asociaciones (servicios que se solicitan juntos), patrones secuenciales (si se compra algo en una fecha en x meses se adquiere otro producto) y secuencias similares (detectar fenómenos comportamientos similares). Luego de haber definido someramente los objetivos de las Asociaciones es fácil percatarse que esta herramienta brinda una potencial vía de dar alcance a los objetivos que queda fuera del descubrimiento directo.

Análisis de la cesta de compra

Aquí se utiliza la información de las solicitudes de los usuarios para intentar descubrir ¿Quién? y ¿Por qué? Se solicitan esos productos o servicios. Esto tiene gran aplicación en lugares donde de

algún modo se monitoreen las acciones de los clientes. No es difícil percatarse que el CC nos brinda esta potencialidad, además que los resultados obtenidos son de gran compatibilidad con los objetivos de trabajo puesto que satisfacen las requerimientos del CRM.

Los resultados son muy claros y generalmente muy útiles no obstante hay que distinguir entre las reglas que se ofrecen:

- Útiles.
- Triviales: Conocidos por todos o porque podrían estar explicando resultados de campañas anteriores.
- Inexplicables.

Requisitos

- No necesita decir los atributos de los lados derecho e izquierdo de las reglas pues se generan de manera automática.
- Existen variedades para tratar todo tipo de datos.
- Especificar mínimo soporte.
- Especificar máximo número de reglas.

La obtención de reglas de asociación está sujeta a las siguientes restricciones:

- En problemas reales el número de instancias a tratar es muy elevado (centenares de miles o más).
- También en casos reales el número de atributos es muy elevado (cientos o miles de atributos).
- Los algoritmos de este tipo consideran que los datos de entrada están recogidos en una tabla transaccional. Para otro tipo de datos es necesario binarizar ciertos atributos.

Para solucionar el problema de encontrar las reglas de asociación se opta por dos vías:

1. Encontrar el conjunto de productos que tienen el soporte mínimo requerido.
2. Usar los conjuntos frecuentes para generar las reglas.

Técnicas

Se pueden generar tablas de concurrencias que van calculando las ocurrencias de pares de valores, tripletas,...

Esto generaría inconvenientes como explosión combinatoria, se necesita algún método (soportes mínimos) para evitar el crecimiento de las tablas.

Algoritmos de Asociación

A priori (Agrawal '93) este algoritmo pretende obtener ítems (conjunto de valores que se repiten) de un determinado tamaño, para combinarlos en reglas.

Ventajas:

- A priori y sus variantes son los más usados dentro de este tipo de análisis.
- Eficiencia para grandes volúmenes de datos muy elevada.
- Ciertos SGBD (sistema gestor de bases de datos) son capaces de ejecutar este algoritmo dentro del núcleo del gestor.

Deficiencia: Para ciertos datos de entrada, los resultados intermedios consumen gran cantidad de recursos (memoria).

Algoritmo de Asociación: A priori

Ii= {**a1**, **a2**, **a3**,...} conjunto de todos los atributos. **i=1**.

I) Se recorre la tabla de entrada y se actualiza el conjunto

Li= {l1, l2, l3,...} Donde cada **li** es el par formado por un elemento de **Ii** y por el número de veces que dicho elemento ocurre en los datos de entrada.

II) Se eliminan **Li** los elementos cuyo contador no supere el umbral mínimo.

III) Se genera un nuevo conjunto **Ii**+1 como el de grupos de atributos de tamaño **i**+1 a partir de los conjuntos de atributos de **Li**

IV) Si **Li** no esta vacío e **i**<MAX regresar a **I).**

Uso de taxonomías y valores virtuales

A veces es muy útil disponer de generalizaciones o agrupaciones de los productos que se están considerando. Eligiendo de manera inteligente el nivel al que generalizar se puede mejorar los resultados. Para establecer hábitos de los clientes (productos bajos en calorías,...) se pueden insertar productos virtuales y fijarnos en las asociaciones que los contengan.

A veces las combinaciones obtenidas pueden ser de utilidad, no obstante, se pueden obtener reglas que pueden ser interesantes. Las reglas de asociación toman la forma: si condición entonces resultado

Donde tanto la condición como el resultado son combinaciones disjuntas de productos. Hay que establecer parámetros para medir la bondad de las reglas.

Soporte: es el porcentaje de cestas (transacciones) que contienen tanto la condición como el resultado. Esto es, porcentaje de transacciones donde la regla es cierta.

Confianza: es el porcentaje de transacciones que conteniendo la condición también contienen el resultado. Es decir es la probabilidad de que se encuentre el resultado una vez que se tiene la condición.

Lift: (mejora): Es una medida que se utiliza para comparar la probabilidad de encontrar los productos del resultado en una transacción que sabemos que posee la condición con la probabilidad de encontrar esos mismos productos en una cesta cualquiera.

Patrones secuenciales mediante asociaciones

Mediante las asociaciones descubrimos sucesos que ocurren juntos. Los patrones secuenciales requieren la identidad del cliente. Para poder calcularlos necesitamos unir (ordenar) a cada suceso una fecha. Generalmente no se van a calcular las reglas, el interés está en las asociaciones. Para el análisis de series temporales existen métodos más específicos (redes neuronales y regresión).

Resumen

Respecto a esta parte de Asociaciones podemos resumir que es de suma importancia para alcanzar algunos de los objetivos trazados en la comprensión del problema. Como ya se ha mencionado en

otras ocasiones el CC sobre el cual se desarrollará el módulo de minería tiene la peculiaridad de brindar servicios a distintos proveedores lo cual imposibilita concretar cualquiera de estas técnicas, siendo necesaria el criterio de un experto por cada proveedor el cual daría criterios indispensables para aplicación de estas técnicas de manera que respondiesen a sus intereses.

2.8 EVALUACIÓN (FASE V CRISP-MD)

En este momento se dispone de al menos un modelo que parece tener buena calidad desde la perspectiva del análisis de datos. Antes de la implantación es importante revisar el proceso para cerciorarse de que también ha logrado los objetivos de negocio. Es importante en este punto determinar si algún aspecto de negocio no ha sido tenido suficientemente en consideración. Al final de la fase se tendrá la decisión sobre el uso de los resultados de minería.

Es evidente que en esta fase de la metodología es necesario tocar resultados, de los cuales se carecen por cuestiones antes mencionadas. No obstante es preciso aclarar que de la forma que se trazaron los aspectos de negocio y de minería de datos y a su vez la relación con la **CRM** es muy poco probable que quede sin tocarse algún aspecto del negocio.

2.8.1 Fases y Salidas

Evaluar los resultados

- Contrastar los resultados de minería con los criterios de éxito del negocio.
- Modelos aprobados.

Proceso de revisión

- Revisión del proceso.

Determinar los pasos siguientes

- Lista de posibles acciones futuras.
- Decisión sobre la implantación.

Aunque en esta fase y la próxima no se logren los objetivos a los cuales estas responden no se logran, se tiene que tener en cuenta que el objetivo de este trabajo es planear una estrategia para la elaboración del modulo de minería de datos del CC.

2.9 IMPLANTACIÓN (FASE VI CRISP-MD)

La creación del modelo no es el final del proyecto. Incluso cuando se trata de incrementar el conocimiento, este se tiene que poner en orden y presentarlo de manera que se pueda hacer uso del mismo. Esta fase por tanto, puede ser tan simple como la generación de un informe o tan compleja como la implantación de un proceso de minería en toda la empresa. Es importante que al usuario se le deje claro las acciones necesarias para hacer uso efectivo de los modelos obtenidos.

En nuestro caso de estudio particular, la estrategia trazada consiste en organizar el trabajo. Esto para que una vez creadas las condiciones para la aplicación, este todo previamente analizado y se halla sobrepasado un periodo importante de tomas de decisiones a lo largo de todo el estudio y aplicación de la metodología. Como consecuencia, esta fase se convertiría en la generación de un informe. Informe que no será presentado pues para ello sería necesario contar con la suficiente información en las bases de datos del servidor central del CC como para poner a prueba la aplicación de la metodología. Y recordamos que dicho CC aún se encuentra en fase de implementación.

2.9.1 Fases y salidas Implantación.

- Desarrollo del plan de implantación
  - Plan de Implantación
- Desarrollo del plan de monitorización y mantenimiento
  - Plan de seguimiento
- Realización del informe final
  - Informe final
- Revisión del proyecto
  - Experiencia
  - Documentación

CONCLUSIONES

Se culmina el presente trabajo, donde se han logrado los objetivos trazados con técnicas y metodologías específicas para cada uno de ellos, en vista de lograr la mejor aproximación a estos, podemos concluir los siguientes resultados:

- Se analizó las metodologías usadas para resolver problemas de Minería de Datos, haciendo una selección conveniente.
- Se logró desarrollar cada una de las fases de la metodología CRISP-DM para el problema en cuestión del CC, aunque algunas de estas no a plenitud por razones de encontrarse el CC en fase de implementación, quedando indicadas para un completamiento. Permitiéndose una mejor comprensión del problema y dándole un enfoque desde la perspectiva de negocio.
- Se analizó el proceso de Minería de Datos, por cada una de las ramas que lo componen (Descubrimiento Directo e Indirecto) y a su vez las técnicas más usadas en cada uno de estos casos.

Con los resultados expuestos anteriormente no cabe duda de haber logrado una consistente estrategia de trabajo para la elaboración del Módulo de Minería de Datos del CC y así queda vencido el objetivo trazado.

RECOMENDACIONES

El simple y la vez importante hecho de que este trabajo sea solo el primer y gran paso para el real diseño e implementación del Módulo de Minería de Datos del CC, hace que una vez llegado a su finalización, queden una gran gama de líneas futuras a seguir. Se recomienda:

- Una vez que se cuente con los datos necesarios, llevar a profundidad la aplicación de la metodología aquí estudiada.
- Llevar a cabo la implementación del Módulo de Minería de Datos apoyándose en la estrategia aquí descrita.
- Búsqueda y estudio de estándares para el desarrollo de proyectos de CRM.
- Del estudio sobre la Gestión de Relación con Clientes (CRM) ha de existir información suficiente como para apoyar los objetivos del negocio en esta base. Tenerle muy en cuenta.

BIBLIOGRAFÍA

[1]. Agrawal. R. Imeilinski, T. Swami. A. "Mining association rules sets of items in large databases". Proceedings of ACM SIGMOD conference on management of data SIGMOD'93.(1993).

[2]. Agrawal, R.; Srikant, R. "Fast algorithms for mining association rules in large databases". Proc International Conference on Very Large Databases, pp. 478-499. Santiage, Chile: Morgan Kaufmann, Los Altos, CA. (1994).

[3]. Andrássyová, E. Paralic, J. "Intelligent Knowledge Discovery" Dept. of Cybernetics and Artificial Intelligence (1998).

[4]. Andrássyová, E. Paralic, J. "Knowledge Discovery in Databases: A Comparison of Different Views" Dept. of Cybernetics and Artificial Intelligence (1999).

[5]. Berry, Michael J.A.; Gordon Linoff. "Data Mining Techniques For Marketing, Sales and Customer Support". Editorial Wiley, (1997).

[6]. Bigus Joseph P. "Data Mining with Neural Networks". Ed. McGraw Hill, (1996).

[7]. Braun H., Riewdmiller M., "Rprop: a fast adaptative learning algorithm", En Proc.of the Inst. Symposium on Computer and Information Science VII. (1992) human centered approach, Advances in Knowledge Discovery and Data Mining".

[8]. Cabena, Peter; Hadjinian, Pablo; Stadler, Rolf; Verhees, Jaap; Zanasi, Alessandro; "Discovering Data Mining. From Concept to Implementation". IBM. Prentice AAAIIMIT Press. (1996)

[9]. Chang, G.; Healey, M.; McHugh, Jason; Wang, J. "Mining the World Wide Web. An information Search Approach". Kluwer academic Publishers. London, (2001).

[10]. COTEC, Documentos sobre oportunidades tecnológicas., "Redes Neuronales", Diciembre 1998, 1ª Ed.(1998)

[11]. Chapman, P. Clinto, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. "CRISP-DM 1.0. Step-by-step data mining guide". Dirección Web: http://www.crisp-dm.org. (2002).

[12]. Cuadrado Vega, Abel Alberto. "Supervisión de Procesos Complejos mediante Técnicas de Data Mining con Incorporación de Conocimiento Previo". Tesis Doctoral. Universidad de Oviedo. Noviembre, (2.002).

[13]. Daedalus. "Minería de Datos: Conceptos y Objetivos". Dirección Web: http://www.daedalus.es. (2002).

[14]. Davis, L. "Handbook of Genetic Algorithms" Van Nostrand Reinhold, (1991).

[15]. Dirección Web: http://www.gtic.ssr.upm.es/encuestas/delphi.htm (2002).

[16]. Kloesgen, W; Zytkowhttp, J. Dirección Web: http://orgwis.gmd.de/projects/explora/terms.html."Machine Discovery Terminology". (2002).

[17]. Dixon, J.K. "Pattern recognition with partly missing data". Rev: IEEE Transactions on Systems, Man ans Cybernetics, 9:617-621, (1979).

[18]. Famili, A.; Shen , W.-M. "Data Preprocessing and Intelligent Data Analysis". Rev: Intelligent Data Analysis. Vol. 1, nº 1, pp 3-23 (1997).

[19]. Fayyad, U.; Haussler, D.; Stolors, P. "Mining Science Data". Comunications of ACM. Vol 39, Nº 11. (1996). [FAY96a] Fayyad, UM; Simoudis, E. "Data Mining and Knowledge Discovery". Tutorial Notes at PADD'97 - First International Conf. Prac. App.KDD & Data Mining, London, (1997).

[20]. Fayyad, UM; Piatetsky Shapiro, G; Smyth, P. "From Data Mining to KnowledgeDiscovery: an overview". Advances Discovery and Data Mining. AAAI Press/MIT Press, (1996), 1-36.

[21]. Gaines, B. R.; Compton, P. "Induction of ripple-down rules applied to modeling large data bases". Journal of Intelligent Information Systems. 5(3):211-228. (1995).

[22]. Garcke,J; Griebel, M. "Data mining with sparse grids using simplicial basis functions". Knowledge Discovery and Data Mining (2001)

[23]. Goebel, Michael; Gruenwald, Le. "A survey of data mining and knowledge discovery software tools". ACM SIGKDD Explorations. (1999).

[24]. Hand, David; Mannila, Heikki; Smyth, Padhraic. "Principles of Data Mining". A Bradford Book. The MIT Press. London, (2001).

[25]. Hansen, James V.; Nelson, Ray D.; "Data mining of time series using stacked generalizers". Neurocomputing. Noviembre, (2002).

[26]. Hecht-Nielsen, R. "Neurocomputing". Addison Wesley, (1990).

[27]. Hulten, Geoff. Spencer, Laurice. Domingos, Pedro. "Mining Time-Changing Data Streams". ACM. 2000.

[28]. IT Innovation Centre. "CRITIKAL. European Project for Large Scale Data Mining". Dirección Web: http://www.attar.com/pages/critikal.htm. (1999).

[29]. Kargupta, H.; Chan, Philip. "Advances in Distributed and Parallel Knowledge Discovery". AAAI Press. California, (2000).

[30]. Dirección Web: http://www.kdnuggets.com. KDNuggets. Portal de Minería de Datos, (2002).

[31]. Kimball, Ralph; Reeves, Laura; Ross, Margy; Thornthwaite, Warren; "The Data Warehouse Lifecycle Toolkit". Wiley Computer Publishing.USA (1998).Artificial. Universidad Politécnica de Madrid. (2002).

[32]. Maojo, Victor. "Adquisición de Conocimientos". Departamento de Inteligencia Artificial. Universidad Politécnica de Madrid. (2002).

[33]. Martín del Brío, Bonifacio. Sanz Molina, Alfredo. "Redes Neuronales y Sistemas Borrosos. 2ª Edición ampliada y actualizada". Ra-Ma. Madrid, (2001).

[34]. Martínez de Pisón, F.J.; Pernía Espinoza, A.; Castejón Limas, M.; González Marcos, A. "Minería de Datos: Herramientas, Técnicas y Metodologías". Proceeding VI International Congress on Project Engineering. Barcelona, (2002)

[35]. Martínez de Pisón, F.J. "Optimización, Mediante Técnicas de Data Mining, del Ciclo de Recocido de una Línea de Galvanizado". Universidad de La Rioja. Servicio de Publicaciones. 2003.

[36]. Martin, B. "Instance-Based learning: Nearest neighbour with generalisation". MSc Thesis, Department of Computer Science. University of Waikato, New Zealand (1995).

[37]. Mehta, M.; Agrawal, R.; Rissanen, J. SLIQ: A fast scalbel classifier for data mining. Proceedings of the Fifth International Conference on Extending Database Technology, 1996.

[38]. Menasalvas, Ernestina Facultad de Informática "Data Mining: Ténicas y Herramienmtas.ppt" Universidad Politécnica de Madrid. 2001

[39]. Michalski, R; Bratko, I.; Kubat, M. "Machine Learning and Data Mining. Methods and Applications". John Wiley & Sons LTD. England, (1998).

[40]. Morales, E. "Descubrimiento de Conocimiento en Bases de Datos". Curso KDD On- Line. Dirección Web: http://dns1.mor.itesm.mx/~emorales/Cursos/KDD01/. (2000).

[41]. OPEAL. "Tutorial sobre algoritmos genéticos". Direccion web:http://opeal.sourceforge.net/tindex.html. Universidad de Granada. (2001).

[42]. Pernía, A.; Mtnez de Pisón F.J.; Ordieres J.B.; Castejón M.; De Cos F.J. "Gestión del Conocimiento y Minería de Datos". Actas del XVII Congreso Nacional de Ingeniería de Proyectos. Murcia, (2001).

[43]. Pernía, A.; González, A.; Alba, F. "Medida de calidad en modelos de procesos industriales". Proceeding VI International Congress on Project Engineering. Barcelona, (2002)

[44]. Piatetski-Shapiro G.; Frawley W.J. "Knowledge Discovery in Databases". Ed. AAAI/MIT Press. (1991).

[45]. Dirección Web: http://www.r-project.org. "The R Project for Statistical Computing" (2002).

[46]. SAS. "SEMMA. A Proven Data Mining Process". Dirección Web: http://www.sas.com/products/miner/semma.html. (2001).

[47]. Scott, C J and Al-Attar, A and Schneider, W and Nisbet, D and Barth, T and Schwarz, H. "CRITIKAL Final Report". Departament of ECS. University of Southampton, (1999).

[48]. Sebzalli, Y.M.; Wang, X.Z. "Knowledge discovery from process operational data using PCA and fuzzy clustering". Engineering Applications of Artificial Intelligence. Num 14, (2001).

[49]. Siebes, A. "Data Mining and Statistics". Cism Courses and Lectures, nº 408. International Centre for Mechanical Sciences. CISM. Pag. 1 a 38 (2000).

[50]. SPSS "Clementine 6.0: Users Guide" SPSS (2001).

[51]. [SPS02] Dirección Web: http://www.spss.com. "SPSS Home Page" (2002).

[52]. Dirección Web: http//www.statsoft.com/textbook/stathome.html. "Libro electrónico sobre algoritmos de data mining" (2001).

[53]. Thuraisingham, B. "Data Mining. Technologies, Techniques, Tools and Trends". Ed. CRC Press LLC, (1999).

[54]. Wang, Xue Z. "Data Mining and Knowledge Discovery For Process Monitoring and Control". Advances in Industrial Control. Ed. Springer. London, (1999).

[55]. Dirección Web: http://www.cs.waikato.ac.nz/~ml/weka/. "Weka 3 – Machina Learning Software in Java" (2002).

[56]. Witten, Ian H.; Frank, Eibe. "Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations". Morgan Kaufmann Publishers. San Francisco, California (2000).

[57]. http://www.iiia.csic.es/~mario/rna/tutorial/RNA_backprop.html

[58]. http://www.udc.es/dep/mate/estadistica2/

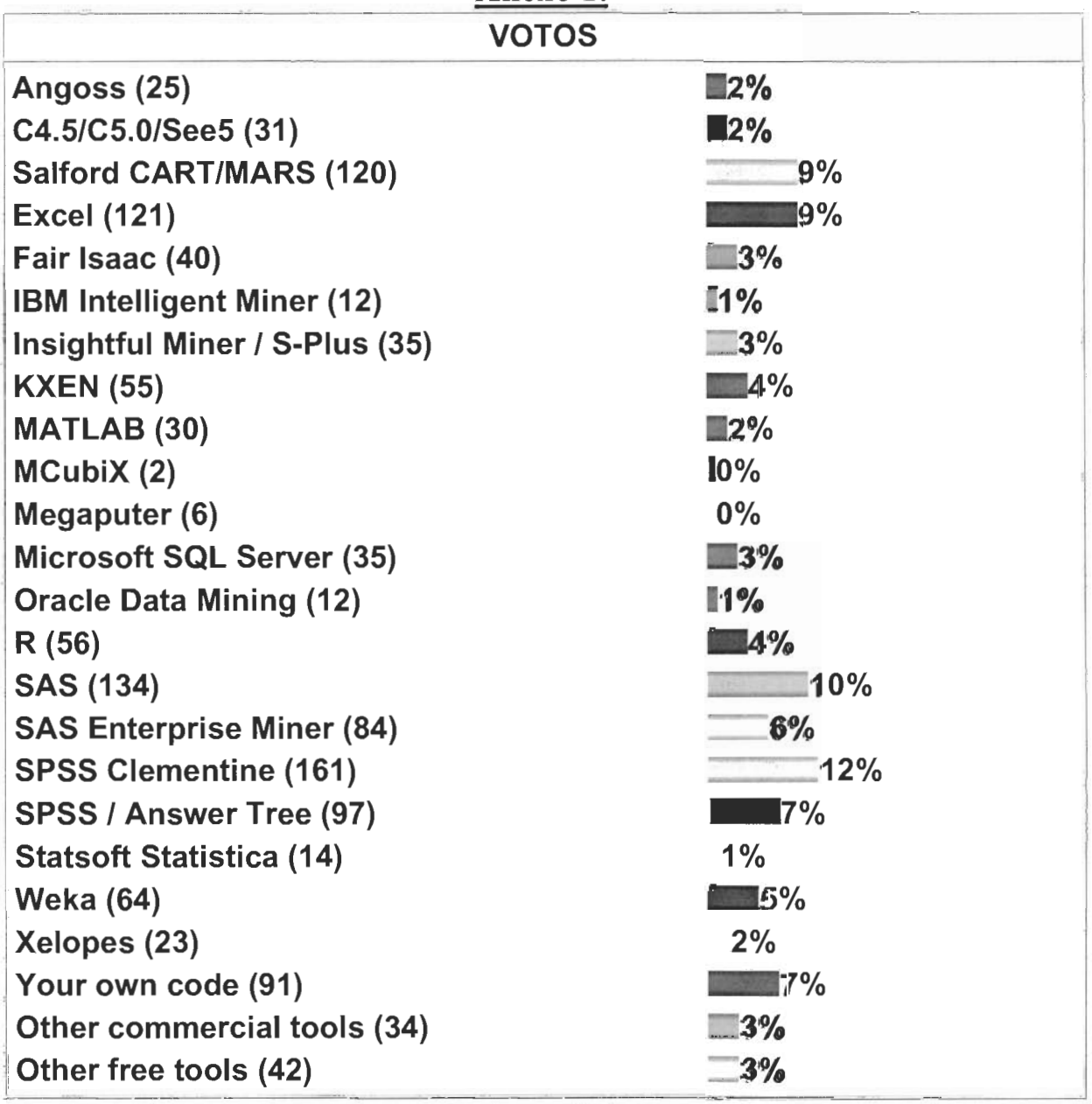
[59]. http://www.monografias.com/trabajos15/loglineal/loglineal.shtml

[60]. http://www.iiia.csic.es/~mario/rna/tutorial/RNA_kohonen.html

[61]. http://www.gestiopolis.com/recursos/documentos/archivodocs/demarketing

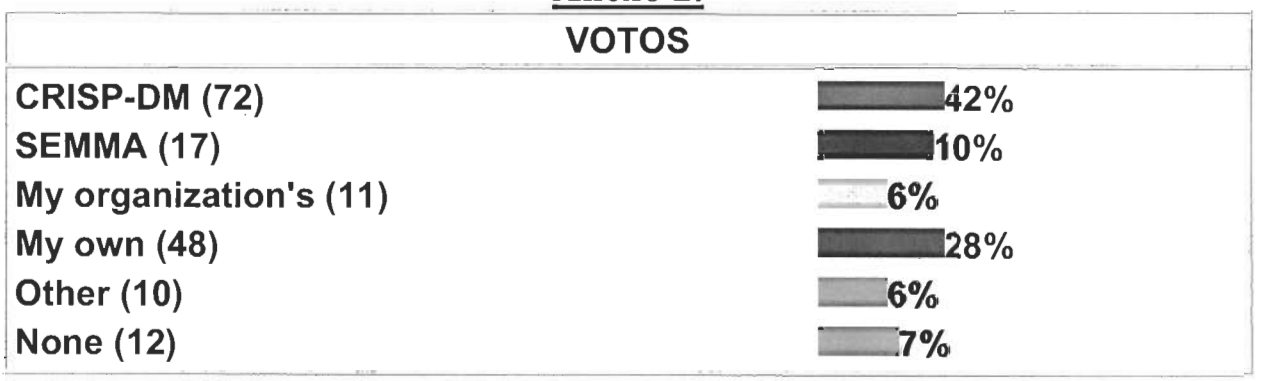
ANEXOS

Anexo 1.



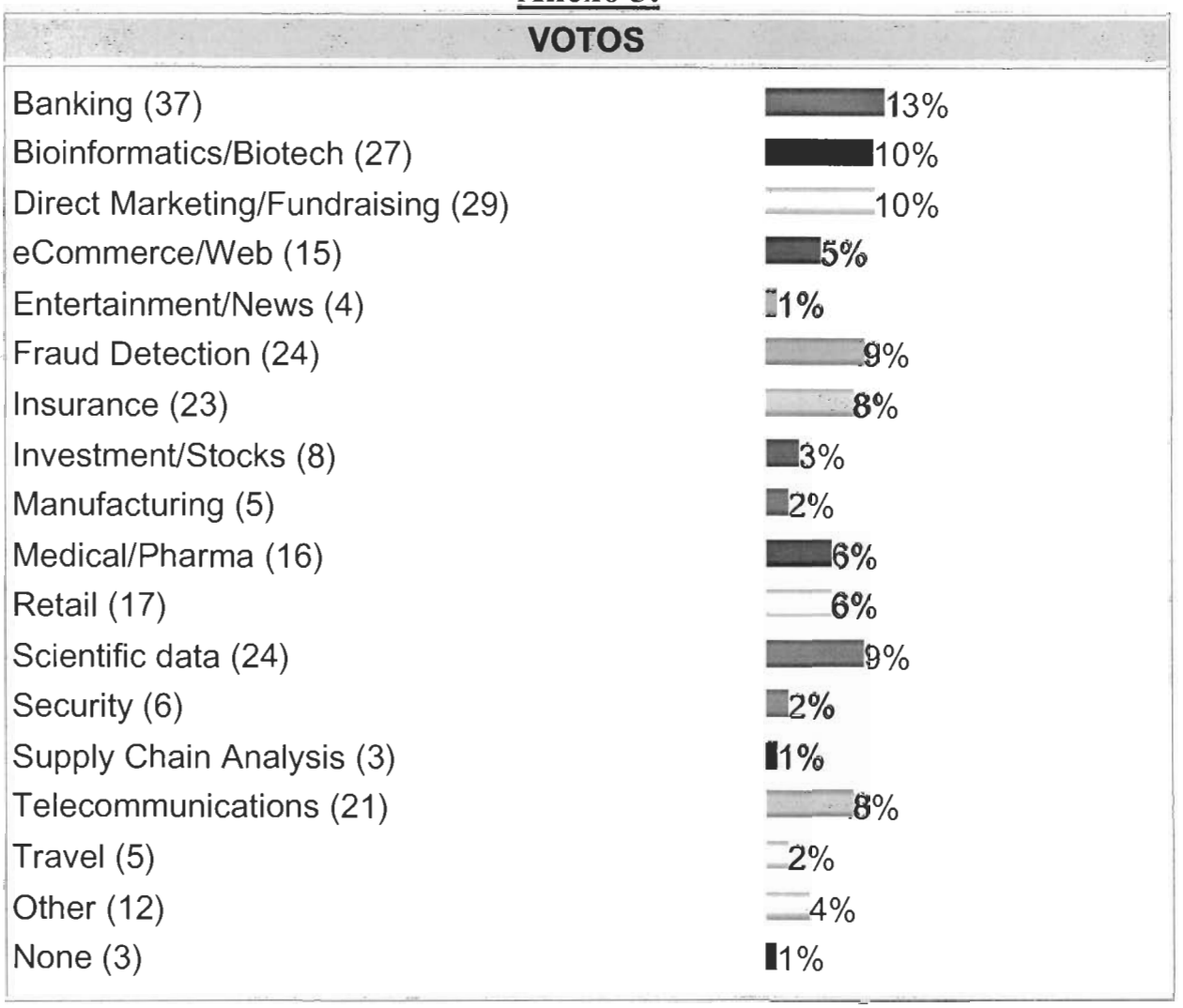
Encuesta sobre herramientas de Minería de Datos usadas regularmente. Mayo 2004[1324 votos]. http://www.kdnuggets.com/polls/2004/data_mining_software.htm

Anexo 2.



Encuesta sobre metodologías de Minería de Datos usadas regularmente. Abril 2004 [170 votos]. http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm

Anexo 3.



Campos donde actualmente de aplica la MD. Agosto 2003 [279 votes total]. http://www.kdnuggets.com/polls/2003/data_mining_applications_industries.htm

GLOSARIO

- Call Center (CC) - Centro Informatizado de Atención de Llamadas.
- ACD (Automatic Call Distribution) - Distribución automática de Llamadas. Un sistema ACD tiene por objetivo el manejo de la correcta distribución de las llamadas entrantes a las distintas posiciones de operadores telefónicos, en forma rápida y permitiendo una carga de trabajo uniforme por operador. La distribución de la llamada se puede basar en algoritmos y sistemas de asignación parametrizables por el cliente.
- IVR (Interactive Voice Response) - Sistema de Respuesta Interactiva de Voz. Su principal función es la atención de la llamada entrante durante un cierto tiempo en base a "vocalización de datos que se encuentran previamente grabados, o se acceden desde una base de datos operacional, y que son manejados interactivamente durante la llamada, en función de una programación determinada por la aplicación específica del cliente".
- CTI (Computer and Telephony Integration) - Integración Telefonía e Informática. Interviene al transferir la llamada a otro operador, o de un **IVR** a un operador. Esta necesidad puede surgir del hecho de que, la llamada requiera una especialización distinta al operador que se determinó originariamente, o bien de la necesidad de intervención de un nivel de supervisión para la resolución definitiva del problema, o al hecho que el cliente quiera resolver varios problemas en la misma llamada, o que luego de navegar por el **IVR** el cliente requiera atención personalizada, etc.
- CRM (Customer Relationship Management) – Gestion de Relaciones con Clientes.