ESTUDIO SOBRE EL ESTADO DE LAS SOLUCIONES ICT Y DE LOS CASOS PRÁCTICOS DE APLICACIÓN DE LA MINERÍA DE DATOS A NIVEL MUNDIAL EN AL MENOS 5 CASOS REPRESENTATIVOS.

MELINA ALEJANDRA RENDON HERRERA
JUAN DAVID ACOSTA VASQUEZ

Proyecto de Grado para optar el título de Ingeniero de Sistemas

Asesor
SONIA CARDONA RIOS
INGENIERA DE SISTEMAS
MAGISTER EN ADMINISTRACION

DEPARTAMENTO DE INFORMÁTICA Y SISTEMAS
ESCUELA DE INGENIERÍA
UNIVERSIDAD EAFIT
MEDELLIN
2006

Nota de aceptación

_____

_____

Presidente del Jurado

_____

Jurado

_____

Jurado

_____

A mis padres, Francisco Javier Acosta S y Emperatriz Vásquez; a mi hermana Carolina.

Juan David Acosta V

A mi Familia, William Rendón H., Omaira Herrera A., Jorge Rendón. A mis Profesores y Amigos que hicieron de este reto una realidad.

Melina Alejandra Rendón H

AGRADECIMIENTOS

Expresamos nuestros agradecimientos a:

Sonia Cardona asesora de este proyecto, por su colaboración, guía y constante motivación haciendo posible el desarrollo del Proyecto de Grado.

A nuestros padres por su esfuerzo y motivación. Gracias a ellos hemos consolidado nuestros estudios profesionales.

A nuestros amigos y en general a todas aquellas personas que nos han brindado su apoyo incondicional.

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

INTRODUCCION

La tecnología actual ha facilitado el almacenamiento y procesamiento de grandes volúmenes de datos que surgen de una infinidad de procesos que se han sistematizado en las diferentes actividades que realiza el hombre a diario. Sea el manejo de cuentas bancarias, de compras en grandes supermercados o la simple información que maneja un gobierno o una institución pública o privada.

El problema de tener estos grandes volúmenes de datos[1] es que resulta imposible para un humano realizar análisis sobre estos, para que aporten información y conocimiento que facilite un proceso de toma de decisiones. Aquí es donde entra a jugar lo que se conoce como minería de datos con la cual se trata de identificar patrones, asociaciones, reglas y nuevo conocimiento; que permita a las compañías que están en un ambiente bastante competitivo poder guiar sus estrategias de mercadeo, investigación y administración [DEA02].

La idea es generar conocimiento que sea útil, valido y relevante para esa toma de decisiones, mediante algoritmos eficientes. También se quiere que la información sea presentada con un formato adecuado y claro de modo que la interpretación sea mas clara [DEA06].

El objetivo del proyecto de grado es Identificar, clasificar y comparar casos reales en los que se han aplicado por completo un proceso de minería de datos en algunas empresas del mundo, reconociendo, describiendo y analizando las herramientas o soluciones ICT de minería de datos mas utilizadas con el fin de

[1] Según estudios de la Universidad de California en Berkeley, la actual producción mundial de nueva información es del orden de las decenas de exabits (1018 bits) por año.

proporcionar un documento que sirva de apoyo y guía a las empresas colombianas que quieren empezar a adoptar soluciones completas de minería de datos.

El primer capitulo de este proyecto es fundamental para el empresario que quiera implementar una solución de minería de datos, pues en este, se definen los tipos de tareas mas comunes de la minería, en las cuales se pueden traducir los problemas de negocio a tratar; y además se nombran los métodos mas utilizados para resolver este tipo de problemas. También se exponen las metodologías SEMMA y CRISP-DM, que se consideran los modelos más importantes para el desarrollo de proyectos de extracción de conocimiento [Gon04], haciendo mas énfasis en la segunda, pues esta es la que se considera estándar en el mundo, y ha sido utilizada por gran cantidad de empresas para desarrollar proyectos de esta clase. Además esta metodología va a servir como referencia para el desarrollo de los casos mas representativos que se presentaran en el capitulo 2.

El siguiente capitulo expone casos reales en los cuales se ha aplicado un proceso completo de minería de datos teniendo como referencia la metodología CRISP. Se muestra de una forma muy general como se tradujo el problema del negocio en un problema de minería, como se escogieron los datos y como se prepararon estos para que el algoritmo a utilizar trabajara correctamente, luego se generaron los modelos de acuerdo a uno o varios algoritmos, se evalúan estos resultados y por ultimo se publican cuales son las posibles formas en que se emplean los resultados obtenidos.

El ultimo capitulo se enfoca en las principales herramientas y proveedores de minería de datos que hay en la actualidad. Inicialmente se le presenta a las empresas u organizaciones que desean adquirir una herramienta de minería de datos los criterios y características que a consideración se deberían de tener en cuenta. Después de contextualizar al lector y entregarle un marco de referencia en

cuanto a la selección de una herramienta, se muestran cuales son las más importantes a nivel mundial, de acuerdo a dos criterios. El primero se basa en una evaluación de posicionamiento y funcionalidad de las herramientas de minería de datos, presentada por la empresa GARTNER [Met04], la cual se dedica al análisis de tecnología. El segundo basado en los resultados obtenidos en una encuesta realizada por KDNUGGETS a los usuarios de estas herramientas, de una forma un poco informal y con un rango de error un poco más grande. Después de identificadas cuales son las herramientas y los proveedores mas trascendentales en el mundo, se ubican cuales son las que tienen presencia en Colombia y se añade otro grupo de proveedores que aunque son reconocidas en el ámbito nacional, no fueron identificadas en las evaluaciones anteriores.

Por ultimo se presentan un conjunto de conclusiones del trabajo, acorde a lo visto capitulo por capitulo, se hacen recomendaciones de trabajos futuros y se ven las complicaciones respecto al alcance.

1. METODOLOGIAS, TAREAS Y TECNICAS DE MINERIA DE DATOS

Antes de comenzar a ver los distintos casos documentados y las diferentes herramientas que hay en el mercado para el desarrollo de soluciones de minería de datos; es necesario tener en cuenta ciertos aspectos teóricos que son de vital importancia para entender los capítulos posteriores.

Al momento de las investigaciones pertinentes para la realización de la construcción de los casos, se encontró que no se tenían bases muy sólidas en cuanto a la estructura y el porque de los casos.

Es por esto que, primero se realizo la tarea de investigar cuales eran los tipos de problemas o tareas mas comunes de la minería de datos y cuales eran los algoritmos mas utilizados según la tarea. Utilizando varias fuentes [Her04], [Par01], [CCK+00] se llego a un consenso donde se estableció 4 tareas principales de las que se podían desprender otras con sus respectivas particularidades. Estas tareas son: clasificación, pronóstico, segmentación, asociación.

El segundo paso fue identificar y analizar las principales metodologías para el desarrollo de proyectos de minería de datos en el medio. Se encontró que existen muchas formas de desarrollar procesos de minado, pero las más conocidas en el medio son: la metodología CRISP-DM y la metodología SEMMA. Ambas metodologías fueron desarrolladas por proveedores de estas herramientas. Después de haber realizado el estudio de cada una de estas dos metodologías, se encontró que la metodología CRISP-DM es considerada la metodología estándar en el medio, según comparaciones y análisis encontrados en Internet y otras

fuentes, y fue concebida como una metodología abierta, lo cual significa que no esta basada en una herramienta especifica, contrario de lo que ocurre con SEMMA, pues esta fue definida para la herramienta SAS Intelligent Miner [Gon04].

Por esta razón, en este primer capitulo se hace un hincapié en la metodología CRISP-DM, identificando y analizado cada una de sus etapas, las tareas que se encuentran en cada una de estas y los resultados que se obtienen después de realizar las tareas [CCK+00]. Junto con la descripción y el análisis de esta metodología, se presentan a su vez un conjunto de mejores prácticas o tips que servirán de guía o referencia a las organizaciones que deseen seguir esta metodología en sus proyectos de minería de datos.

1.1. TIPOS DE TAREAS EN LA MINERÍA DE DATOS

Una buena forma de iniciar un proceso de Minería de datos, después de identificar el problema del negocio que se quiere tratar, es traducir el problema del negocio en un problema de minería de datos. Cuando se habla de un problema de este tipo, normalmente se hace referencia a las tareas que se pueden realizar dentro de los proyectos de minería [Ber05].

Para propósitos de traducción de un problema de negocio a un problema de minería de datos. Inicialmente se debe identificar el contexto del problema, es decir, si se busca predecir un comportamiento o si se busca describir un comportamiento [Lar06]. Una vez se haya identificado el contexto del problema, se deberá seleccionar alguna de las tareas o problemas de minería de datos que mas se ajuste al problema del negocio que desea solucionar. Estas tareas se presentan en la siguiente tabla, con una pequeña descripción que se espera sea útil en el momento de determinar cual será la tarea a realizar en un proyecto de minería de datos según el problema del negocio que se quiere resolver. Para más

información sobre el proceso de traducción del problema de negocio, remitirse a la descripción y definición de la primera etapa de la metodología CRISP-DM que se presenta a continuación.

Tabla 1. Tareas y Métodos en la Minería de Datos

	Tipo de tarea	Técnicas	Ejemplos
Predictiva	**Clasificación:** La tarea o problema de clasificación consiste básicamente en examinar cada uno de los atributos o características que se tengan de una elemento dado, basado en estas características o atributos, realizar un estudio para identificar la categoría o clase a la cual podrá ser asignado. Cada una de estas clases o categorías deben estar previamente identificadas y definidas. Después de que se tengan las clases definidas, se podrá catalogar cada nuevo elemento que se tenga dentro de alguna de las clases dadas.	1. Árboles de decisión 2. Vecino más cercano (K Nearest Neighbor). 3. Redes neuronales. 4. Análisis de vínculos (Link Análisis) 5. Regresión logística y polinomial. 6. Naive bayes 7. Algoritmos genéticos y evolutivos. 8. Maquinas de vectores soporte. 9. Reglas CN2 (cobertura) 10. Análisis discriminante multivariante 11. Reglas de método inductivo 12. Razonamiento caso-base	1. los bancos generalmente tienen la información del comportamiento de pago de los clientes. Combinando la información financiera, con otra como sexo, edad, ingresos. Es posible desarrollar un sistema de clasificación de nuevos clientes, como buenos o malos clientes. 2. Escoger el contenido para desplegar en una pagina Web. 3. Encontrar reclamos de seguros fraudulentos. 4. Determinar que números de teléfono corresponden a que maquinas de fax.
	Pronostico (Estimación y predicción): Cuando se dice que se tiene un problema predictivo de pronostico, se quiere decir que se tiene un problema en donde se debe hacer una estimación de unos valores o atributos de un elemento dado y del cual no se tiene mucha información, es por esto que se utiliza la información de los atributos de otros elementos para así poder identificar un valor “estimado” de este primer valor del cual no se tiene mucho conocimiento. También se puede hablar de un problema de	1. Árboles de decisión 2. Árboles de decisión CART. 3. Redes neuronales. 4. Vecino más cercano (K Nearest Neighbor). 5. Análisis de vínculos (Link Análisis) 6. funciones de base radial (RBF)	1. Predecir que clientes dejaran de serlo en los próximos 6 meses. 2. Predecir que suscriptores al servicio telefónico ordenaran un servicio adicional. 3. Predecir que productos venderles a que clientes. 4. Estimar el ingreso total de un grupo familiar. 5. Estimar la expectativa de vida de un cliente.

	pronostico, cuando se realiza una estimación inicial de datos, como se menciono anteriormente, pero se tiene como objetivo de esta estimación, predecir un evento futuro, basado en datos estimados.	7. Reglas de asociación. 8. regresión logística y polinomial 9. Árboles de Regresión 10. Método Box-Jenkins 11. Algoritmos Genéticos 12. Modelos de regresión. 13. Series de Tiempo	
Descriptiva	**Asociación:** Consiste en identificar que atributos o elementos tienen algún nivel de asociación a otros atributos o elementos en un ambiente determinado.	1. Reglas de asociación. 2. Regresión logística. 3. Pruebas de independencia 4. Pruebas de bondad de ajuste 5. Algoritmos genéticos y evolutivos 6. Reglas CN2 (cobertura) 7. Análisis de Correlación 8. Redes Bayesianas 9. programación lógica inductiva	**1**. Determinar que artículos van juntos en las compras de una persona, en el supermercado. **2**. Identificar oportunidades de cross-selling. **3**. Aplicando algoritmos de reglas de asociación a datos de accesorios de carros, una compañía de automóviles encontró que si un radio es ordenado, el cliente automáticamente también ordena un gearbox el 95% de las veces. Basados en esta dependencia, la compañía decidió ofrecer estos accesorios como una combinación de los mismo para así reducir costos

	Agrupamiento (Clustering): Esta tarea consiste en segmentar una población heterogénea en un número de subgrupos o clusters homogéneos. La diferencia entre clustering y clasificación es que el primero no se apoya en clases predefinidas. El agrupamiento divide la población en clases, donde cada uno de los miembros de cada clase tiene similitudes con otro miembro de la misma clase.	1. Redes neuronales. 2. Redes de kahonen. 3. Kmeans 4. Vecinos más próximos. 5. Twosteb, cobweb 6. Algoritmos genéticos y evolutivos 7. Maquinas de vectores soporte.	1. realizar una segmentación de clientes, para así desarrollar estrategias de mercado especificas a cada grupo

Fuente: [Men00], [BeL04], [CCK+00].

1.2. METODOLOGÍAS PARA EL PROCESO DE MINERÍA DE DATOS

1.2.1. Propósito de las Metodologías en la Minería de Datos

El principal objetivo de las Metodologías que se encuentran en estos momentos en el campo de la Minería de Datos, surge básicamente de la necesidad que se tenía en el medio de tener un proceso o pasos estándares para la resolución de problemas con las herramientas y técnicas presentadas por el concepto de minería de datos.

En Este campo, cuando se desea solucionar uno o varios problemas, no es suficiente con tener las herramientas que me ayuden a resolver preguntas y predecir comportamientos, con un conjunto de métodos y algoritmos formulados para dar respuesta a estos, sino que además se requiere de un conjunto de pasos sistematizados que guíen el proceso que se debe seguir desde que se estudia los problemas que se desean tratar hasta que se tienen las respuestas a estos problemas formulados [CCK+00].

Debido a la creciente necesidad en el medio, distintos proveedores de herramientas de Minería de Datos se pusieron en la tarea de esquematizar el proceso que se recomendaba seguir para el desarrollo de un procesos completo de minería de datos [Gon04], surgiendo así dos de las principales metodologías que hoy existen en el medio: SEMMA y CRISP-DM.

SEMMA (Simple, Explore, Modify, Model, Assess), como metodología para procesos de Minería de Datos fue propuesta por uno de los proveedores de herramientas, llamado SAS. Mientras que la metodología CRISP-DM, por sus siglas en ingles (Cross- Industry Standard Process for Data Mining), fue propuesta por un consorcio de varias empresas europeas de herramientas de minería de datos como NCR de Dinamarca, AG de Alemania, SPSS de Inglaterra y OHRA de Holanda [Gon04].

Es por esto, que se plantea como propósito principal de las metodologías de Minería de Datos crear un proceso sistematizado de desarrollo de proyectos de minado. Dado que es la forma mas efectiva de llegar a resultados óptimos además de ser lo mas recomendado por los expertos.

A continuación se presentaran una descripción mas profunda de estas dos metodologías.

1.2.2. Metodología de minería de datos SEMMA

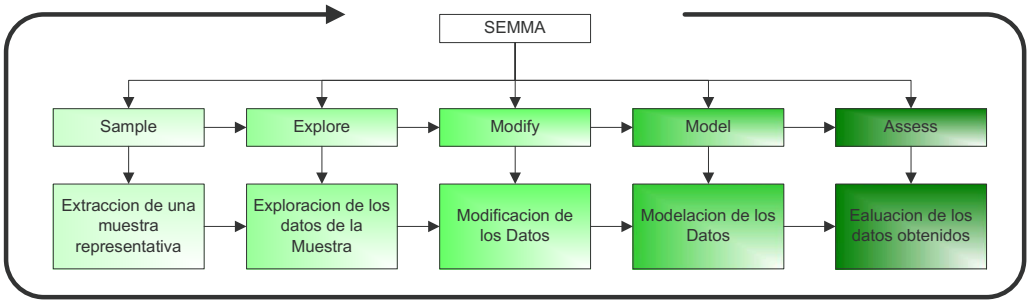
Como lo mencionamos anteriormente, SEMMA con sus siglas en ingles (Sample, Explore, Modify, Model, Assess), fue desarrollada por SAS institute y la define como una herramienta que ayuda a los usuarios en los procesos de selección, exploración y modelación de cantidades significativas de datos almacenados, para así poder responder a preguntas o predecir eventos que pueden pasar.

Según SAS, mas que una metodología de procesos de minería de datos, SEMMA se puede identificar como un conjunto de herramientas funcionales, enfocándose mas en los aspectos del desarrollo del modelo de minería de datos [Sas05].

A continuación se presentan las fases de esta metodología con sus características esenciales[2].

[2] http://www.sas.com/technologies/analytics/datamining/miner/semma.html

Figura 1. Fases de la Metodología SEMMA[3]



Sample → Extracción de una muestra representativa

En esta primera fase de la metodología, se realiza la extracción de un conjunto de datos que sean una buena representación de la población a analizar, esto se hace con el objetivo de facilitar los procesos de minado sobre los datos, reduciendo los tiempos que se necesita para determinar la información valiosa para el negocio.

Explore → Exploración de los datos en la muestra.

En esta fase, se hace un recorrido a través de los datos extraídos en la muestra para detectar, identificar y eliminar datos anómalos, ayudando a refinar los procesos de descubrimiento de información en fases siguientes del proceso. En este punto del proceso, la exploración se puede realizar a través de medios visuales, aunque muchas veces no es suficiente este método, es por eso, que además de la visualización se pueden manejar diferentes técnicas estadísticas como análisis de factores, análisis de correspondencias, entre otros.

Modify → Modificación de los datos.

Esta modificación de los datos se puede realizar creando, seleccionando y transformando las variables en las cuales se va a enfocar el proceso de selección del modelo. Muchas veces se tendrá la necesidad de

[3] Construcción propia basado en información encontrada en [Sas05]

realizar modificaciones cuando los datos que se están analizando cambien. Esto se debe a que el entorno en el que se trabaja la minería de datos es dinámico e iterativo.

Model→ Modelación de los datos,

En esta fase, las herramientas de software se encargan de realizar una búsqueda completa de combinaciones de datos que juntos predecirán de una manera confiable los resultados buscados. Es en esta parte donde las técnicas y métodos de minería de datos entran a jugar un papel importante para la solución de los problemas que fueron identificados al iniciar el proyecto de minería de datos.

Assess → Evaluación de los datos obtenidos

Después de que la fase de modelación presente los resultados obtenidos de la aplicación de los métodos de minería de datos al conjunto de datos. Se deberá realizar un análisis de los resultados para ver si estos fueron exitosos de acuerdo a las entradas que se tuvieron para analizar el problema. Una buena practica para identificar si los resultados con el modelo creado son los esperados, es aplicar este modelo a una porción de datos diferente. Si el modelo funciona correctamente para esta muestra y para la muestra utilizada para el proceso de creación del modelo, se tiene una buena probabilidad de tener un modelo valido.

Después de realizar este proceso de manera iterativa, se llegara a un punto donde se obtendrá el modelo más efectivo, el cual se pasara a producción para iniciar el proceso de predicción de nuevos elementos que ingresen al sistema. Y es con esta etapa, con la cual se finaliza el proceso de minería de datos.

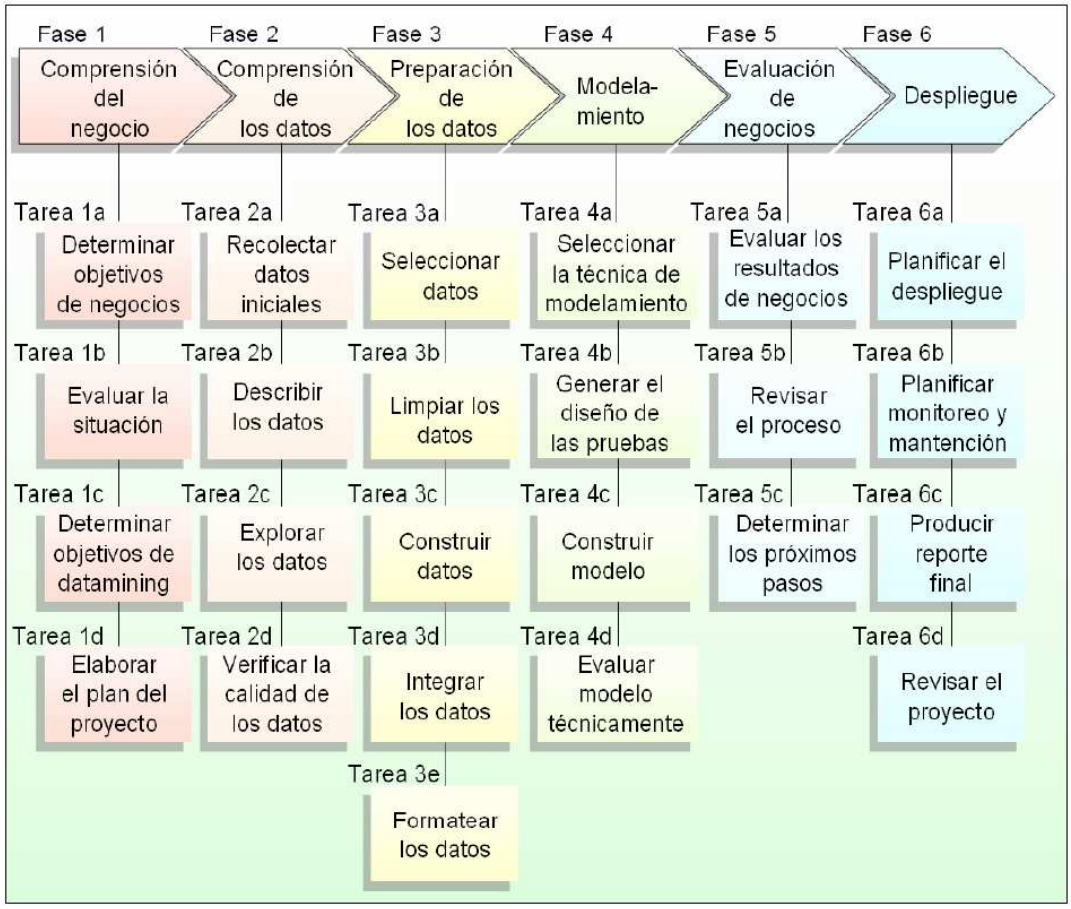
1.2.3. Metodología de minería de datos CRISP-DM

CRISP-DM es una metodología de minería de datos concebida en el año 96. Surge de la necesidad de tres grandes empresas (DaimlerChrysler, SPSS y NCR) interesadas en realizar proyectos de minería de datos, pero a las cuales les faltaba mucha experiencia en el tema. El resultado de la unión y los estudios realizados conjuntamente por estas organizaciones fue esta metodología [CCK+00], la cual proporciona un vistazo general del ciclo de vida de un proyecto de minería de datos. Cada fase se divide en un conjunto de tareas, las cuales se relacionan entre si, esas tareas a su vez se componen de actividades especificas y de un conjunto de resultados concretos.

Esta metodología, muestra el camino que se debe seguir en el proceso de extracción de conocimiento y permite determinar que actividades desarrollar en que etapa, de forma que se puedan lograr los objetivos propuestos. La mejor forma de lograr los objetivos es seguir al máximo este estándar, pues ya ha sido ampliamente utilizado por una gran cantidad de organizaciones que han obtenido resultados muy valiosos.
Los secretos del éxito de proyectos de minería de datos radican en que tan bien se han seguido o desarrollado estas etapas. Lo ideal es evitar el rompimiento del ciclo de la minería de datos, entendiendo las distintas formas en que se puede fallar y que acciones preventivas tomar para evitar estas posibles fallas [BeL04].

A continuación se presenta el ciclo de vida de un proyecto de minería de datos definido por la metodología CRISP–DM, especificando las tareas de cada etapa

Figura 2. Etapas y tareas de un proyecto de Minería de Datos en la Metodología CRISP-MD



Fuente: [Luc06] [CCK+00]

Aunque esta figura muestra las etapas de forma lineal, este proceso es en realidad un ciclo iterativo, que puede ser aplicado varias veces hasta conseguir los resultados deseados [CCK+00], el recorrido no necesariamente tiene que hacerse de una forma rígida, es decir, no limita a que una etapa tenga que estar precedida por una inmediatamente anterior, se puede mover dentro de la estructura, hacia adelante y hacia atrás, según el proyecto lo requiera.

Fases de la metodología CRISP-DM

1. Comprensión del Negocio:

Esta fase se enfoca en entender los objetivos del proyecto y requerimientos desde una perspectiva del negocio, luego convertir este conocimiento en la definición de un problema de minería de datos y el diseño de un plan preliminar para conseguir el objetivo deseado. [CCK00]

Para esta etapa de comprensión del negocio la única forma de obtener buenas respuestas a las preguntas formuladas es involucrando a los dueños de los problemas del negocio y entender como los resultados de la minería de datos serán usados; y a el personal de TI y los administradores de bases de datos para entender como serán presentados los datos. Es recomendable entrevistar a las personas de los distintos departamentos de la empresa en conjunto que hacerlo por separado. De esa forma las personas de las diferentes áreas de conocimiento y con la experiencia adquirida tienen la posibilidad de reaccionar ante otras ideas y aportar otras nuevas. El papel de la persona encarga del proceso de minería de datos en las discusiones o entrevistas debe ser de asegurarse que la declaración final del problema del negocio pueda ser traducido en un problema de minería [BeL04].

Las tareas que se realizan dentro de esta etapa son las siguientes:

A. Determinar los objetivos del negocio

Esta tarea trata de entender a fondo, desde una perspectiva del negocio, que es lo que realmente desea el cliente lograr o cumplir. Cometer errores en esta parte puede conducir a encontrar las respuestas correctas a preguntas erróneas [CCK00].

El destino correcto de un proyecto de minería de datos es el resultado de un problema de negocio muy bien definido. Los objetivos particulares de un proyecto de minería no deben enunciarse en términos muy generales o muy amplios, porque seria más difícil medir. A los proyectos difíciles de medir, son más difíciles de sacarles algún valor. Los objetivos generales deben romperse en objetivos mas específicos, para hacer mas fácil el monitoreo del progreso del cumplimiento de estos [BeL04].

Los resultados que se deben obtener como frutos de esta tarea son el desarrollo de un documento donde se registre el conocimiento que se tenga sobre la situación de negocios actual de la organización al principio del proyecto [Luc06]. Después se debe describir el objetivo primario de los clientes, desde una perspectiva del negocio, se pueden adjuntar otros objetivos relacionados que el cliente quiera tratar de responder. [BeL04] Por ultimo se deben describir los criterios de éxito del proyecto, desde un punto de vista del negocio. Estos criterios deben ser muy específicos y deben ser medidos objetivamente.

B. Evaluar la situación

Esta tarea involucra una exploración mas detallada sobre los recursos, restricciones y otros factores que deben ser considerados al momento de elaborar el plan de proyecto de minería de datos. [CCK00]

En esta actividad es importante contar con una lista de los recursos disponibles para el proyecto incluyendo personal, datos, recursos computacionales y software. También se deben listar todos los requisitos del proyecto, incluyendo fechas de terminación, calidad de los resultados, seguridad, así como cuestiones legales; una lista con las asunciones hechas para el proyecto y una lista con las restricciones tanto técnicas como de recursos [BeL04].

Es significativo saber cuales son los riesgos o eventos que pueden retardar o atrasar el proyecto o aquellos que lo podrían hacer fallar, y tener para cada uno de estos riegos un plan de contingencia, para saber como actuar en caso de que se presente estos riegos.

Se debe hacer un glosario con la terminología relevante para el proyecto, sea terminología del negocio o terminología de la parte de minería.

Por ultimo y no menos importante se debe construir un análisis de costo beneficio para el proyecto, la idea es saber si el negocio es exitoso en cuanto a esta parte [BeL04].

C. Determinar los objetivos de minería de datos

Una meta del negocio indica los objetivos en términos del negocio. Una meta de minería de datos indica los objetivos del proyecto en términos técnicos. [CCK00]

Se debe traducir el problema del negocio en un problema de minería de datos, por lo cual debe ser reformulado en términos de alguna de las tareas de la minería. Como por ejemplo: clasificación, pronóstico, segmentación o asociación [BeL04]. Estas tareas de minería de datos fueron presentados en la tabla 1 de este capitulo, donde se enfatiza en las tareas de minería de datos y las técnicas mas utilizadas para resolverlas.

Los resultados de esta tarea son dos. Primero se debe describir las salidas previstas del proyecto que permiten el logro de los objetivos de negocio y segundo se deben definir los criterios para un resultado exitoso del proyecto en términos técnicos [CCK00].

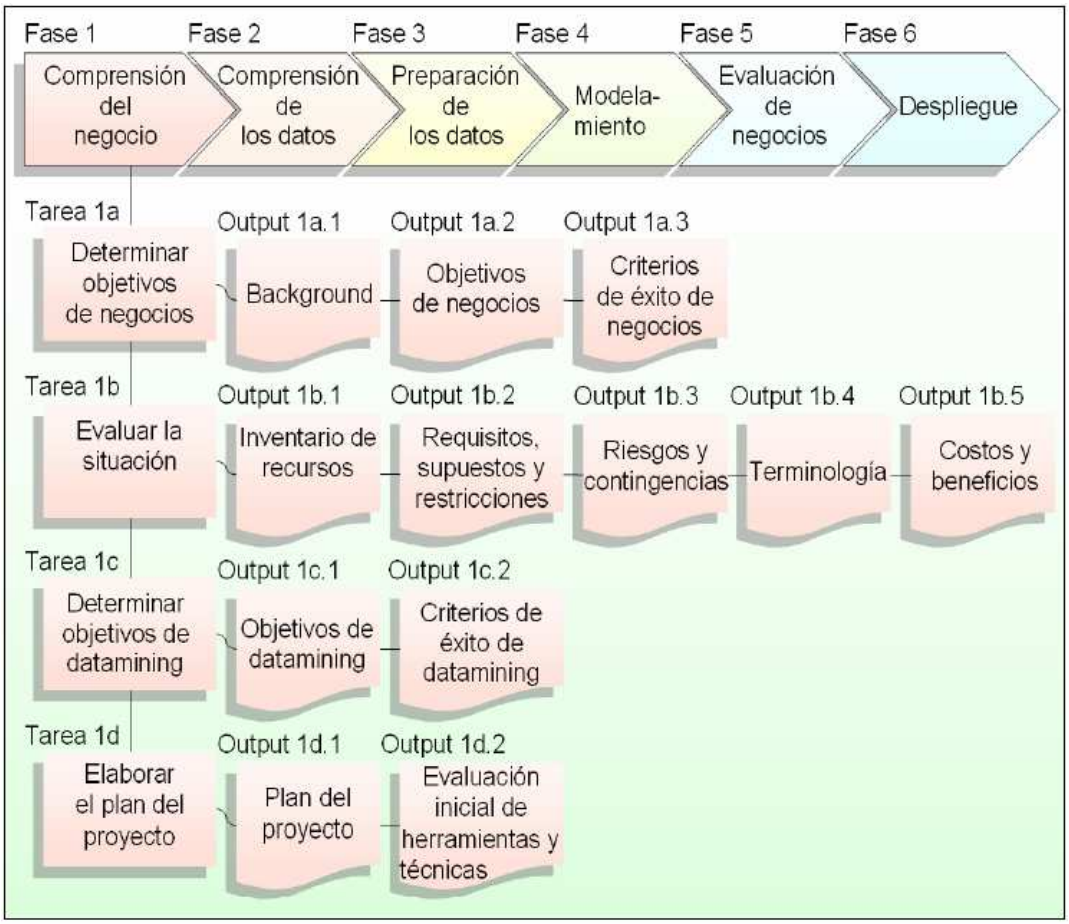
D. Generar el plan del proyecto

Esta actividad indica el plan previsto para alcanzar las metas de minería de datos y por consiguiente logar los objetivos del negocio. Este plan debe de estar especificado en pasos a ser desarrollados durante el resto del proyecto, incluyendo la selección inicial de técnicas y herramientas. [CCK00]

Esta tarea debe arrojar un plan del proyecto que describa las etapas a ser ejecutadas durante el proyecto, donde se especifique la duración, los recursos requeridos, entradas, salidas y dependencias; en lo posible tratar de planificar repeticiones de ciertas fases como las de modelado y evaluación.

Por otra parte, se deben hacer una valoración inicial de las herramientas y técnicas a utilizar.

Figura 3. Fase de entendimiento del Negocio con tareas y resultados



Fuente: [CCK+00], [Luc06]

2. Comprensión de los datos:

Antes de comenzar un proceso de minería de datos es necesario contar con datos apropiados para hacerlo. Por lo que en esta etapa se hace una primera colección de datos y se procede a tener actividades que permitan tener una mayor familiaridad con los datos [CCK+00], identificar posibles problemas de calidad, detectar subconjuntos que puedan ser interesantes para formar hipótesis relativa a la información oculta.

Dentro de la comprensión de los datos, se tienen las siguientes actividades:

A. Recolectar datos iniciales

En esta tarea lo que se hace es adquirir o acceder los datos enumerados en los recursos disponibles para el proyecto. [CCK00]

Como resultado, se debe hacer un reporte que liste el conjunto de datos adquiridos, conjuntamente con su ubicación, los métodos utilizados para adquirirlos, los problemas encontrados y como se trataron de solucionar [BeL04].

El primer lugar para buscar los datos es en la bodega de datos corporativa. Los datos de esta bodega han sido limpiados, verificados y juntados de múltiples fuentes. El problema que puede presentarse es que muchas organizaciones no tienen una bodega de datos o por el contrario existen varias bodegas de datos. En el primer caso los encargados del proceso de minería de datos deben de buscar los datos en las bases de datos de los departamentos y en los mismos sistemas operacionales, por lo que se hace más difícil acceder a estos datos.

La cantidad de datos a utilizar no se puede establecer a ciencia cierta, esto depende del algoritmo que se va a usar, la complejidad de los datos y la frecuencia relativa de los resultados obtenidos. La minería de datos es más efectiva cuando el total del volumen de datos no permite sacar patrones como se haría fácilmente en una base de datos pequeña. La idea es comenzar con decena de miles o hasta millones de registros preclasificados para el conjunto de datos de entrenamiento, validación y pruebas, cada uno con miles de registros [BeL04].

La minería de datos usa también datos del pasado para hacer predicciones sobre el futuro. Para muchas aplicaciones que se enfocan en los clientes es recomendable contar con datos de 2 o 3 años de historia. Sin embargo esto depende de la aplicación [Her05].

Escoger las variables también es una actividad de mucho cuidado, los encargados de esto no se pueden apresurar a eliminar variables, sin haber comprobado su poder de predicción, o si juntándolas con otras aportan considerablemente para la construcción del modelo.

B. Describir los datos

En esta tarea se examina superficialmente de los datos adquiridos y se debe reportar los resultados. [CCK00]

Se debe hacer un reporte en donde se describan los datos en términos de tipo, distribución, tablas de frecuencia, valores máximo y mínimo, y estadígrafos tales como el promedio, la varianza, la asimetría y la curtosis, entre otros. Para realizar esto, se pueden utilizar los histogramas de las variables para realizar observaciones e identificar posibles casos excepcionales y poco comunes. Se debe tener presente el rango de las variables, cuando esta toma valores negativos, si toma valores muy bajos y si esto es razonable; si la media es muy diferente de la mediana, cuantos valores ausentes hay. Para hacer esto se hace necesaria la ayuda de herramientas estadísticas [Luc06].

Se debe tener diccionarios de datos, entrevistas con los usuarios y administradores de bases de datos para determinar cual es la información disponible [Bel04].

C. Explorar los datos

Esta tarea dirige las preguntas de minería de datos, usando queries, técnicas de visualización y de reporte. Corresponde a un conjunto de análisis de los datos realizados mediante el empleo de gráficos y tablas. [CCK00]

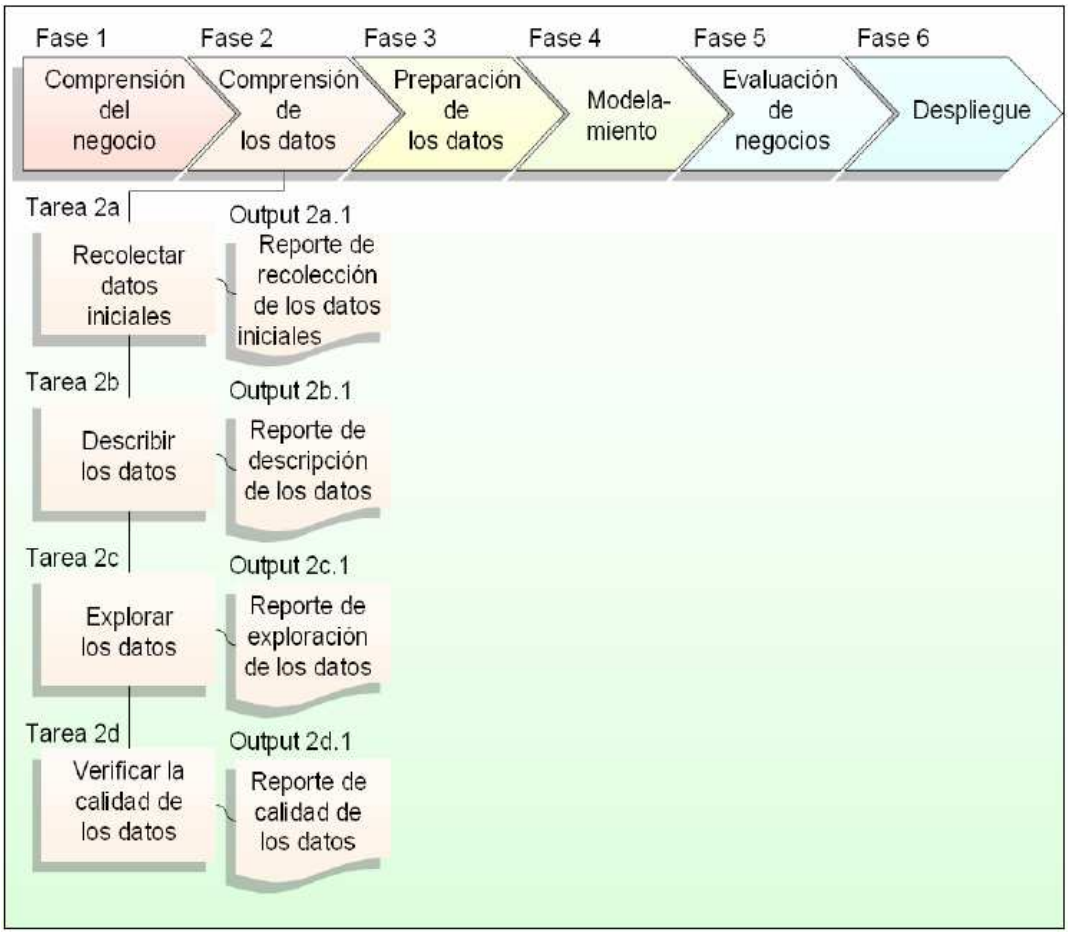
El resultado de esta tarea es la descripción de los resultados de la tarea de exploración de datos, incluyendo los primeros descubrimientos o las hipótesis iniciales y su impacto sobre el resto del proyecto. Este reporte también puede incluir graficas que indiquen las características de los datos o que conduzcan a la obtención de subconjuntos de datos interesantes para exámenes posteriores. [Luc06]

D. Verificar la calidad de los datos

Esta tarea consiste en examinar la calidad de los datos, chequeando que estos estén completos, que no tenga valores ausentes, y que cubran todos los casos requeridos. Determinar si los datos contienen errores y con que frecuencia aparecen.

La salida de esta actividad es un listado con los resultados de la verificación de la calidad de los datos, si hay problemas de calidad y listar las posibles soluciones, las cuales pueden depender fuertemente de los mismos datos y del conocimiento del negocio [CCK00].

Figura 4. Comprensión de los Datos con actividades y resultados



Fuente: [CCK00], [Lac06]

3. Preparación de los datos:

Esta fase cubre todas las actividades necesarias para la construcción del conjunto de datos que alimentara las herramientas de modelado. Estas tareas deben ser ejecutadas varias veces y no tienen un orden preestablecido. Consisten en la selección de tablas, registros y atributos; así como la transformación y limpieza de los datos [CCK00].

Esta fase contiene dos resultados específicos que no están vinculados a ninguna tarea específica y estos son:

- Un conjunto de datos que son usados para el modelamiento y para la mayor parte del trabajo de análisis del proyecto de minería [Luc06].

- Una descripción del conjunto de datos y contiene toda la información que a priori se considera relevante para elaborar el modelo [Luc06].

La creación del modelo requiere ensamblar datos de múltiples fuentes y luego preparar los datos para el análisis.

El modelo de datos es una tabla o conjunto de tablas con una fila por ítem a ser estudiado y campos, por todo lo conocido sobre ese ítem que pueda ser usado para el modelado.

Una buena practica para esta fase, es que hay que tratar de escoger datos de diferentes épocas del año, ya que lo que se quiere es crear un modelo estable, que trabaje bien en cualquier época y sobre todo en el futuro; esto es más probable cuando los datos escogidos no pertenecen a un espacio de tiempo único. Si un modelo esta basado en solamente 3 meses de historia, los diferentes registros del modelo deben corresponder a este intervalo [BeL04].

También se deben coger las variables categóricas que tienen muchos valores [BeL04] y agruparlos en clases de acuerdo a su similitud o también podrían ser reemplazadas por otras variables que cumplan y se ajusten a lo que se necesita.

Las distribuciones sesgadas y los valores extremos también causan problemas a cualquiera de las técnicas de minería que usan valores aritméticos. En algunos casos es recomendable descartar aquellos registros con valores extremos. Sin embargo la mejor estrategia es transformar las variables para reducir el rango de valores, reemplazando cada valor con su logaritmo, por ejemplo [Bel04].

La preparación de los datos tiene las siguientes actividades:

A. Seleccionar los datos

En esta tarea se seleccionan los datos que serán usados para el análisis. Los criterios de selección incluyen tanto la relevancia para los objetivos de minería de datos, así como calidad, y restricciones técnicas como los limites en el volumen de datos o los tipos de datos [CCK00].

El resultado de esta tarea es una lista con los datos a incluir y excluir y las razones para estas decisiones [Luc06].

B. Limpiar los datos

Esta tarea pretende aumentar la calidad de los datos al nivel requerido por las técnicas de minería de datos. Puede involucrar la selección de subconjuntos de datos que no presentan errores, como así también la inserción de valores por omisión en le caso de existir datos faltantes. [Luc06]

Algunos algoritmos son capaces de tratar con valores ausentes o faltantes y lo incorporan en las reglas. Otros no pueden manejar estos valores faltantes [Lar06]. Lo más recomendable para este tipo de situaciones con los valores faltantes es reemplazar ese espacio en blanco con un algún valor como la media o el valor más común [Luc06].

El resultado de esta tarea es un reporte donde se describe que acciones y que decisiones fueron tomadas para enfrentar los problemas de calidad de los datos los cuales fueron reportados durante la ejecución de la tarea de verificación de calidad de los datos de la fase anterior.

C. Construir los datos

Esta tarea incluye la generación de atributos derivados a partir de otros campos, incluye también el cambio de los formatos de los campos existentes [CCK00].

Los resultados de estas tareas son el conjunto de atributos derivados construidos de uno o más atributos existentes. El segundo resultado es la descripción de los registros que incorporan nueva información.

D. Integrar los datos

En esta tarea se utilizan métodos por medio de los cuales la información es combinada de múltiples tablas o registros para crear nuevos registros o valores. El resultado de esta actividad es la fusión de 2 o más tablas que tienen distinta información sobre los mismos objetos; la fusión de tablas también cubre las agregaciones[4] [CCK00].

E. Formatear los datos

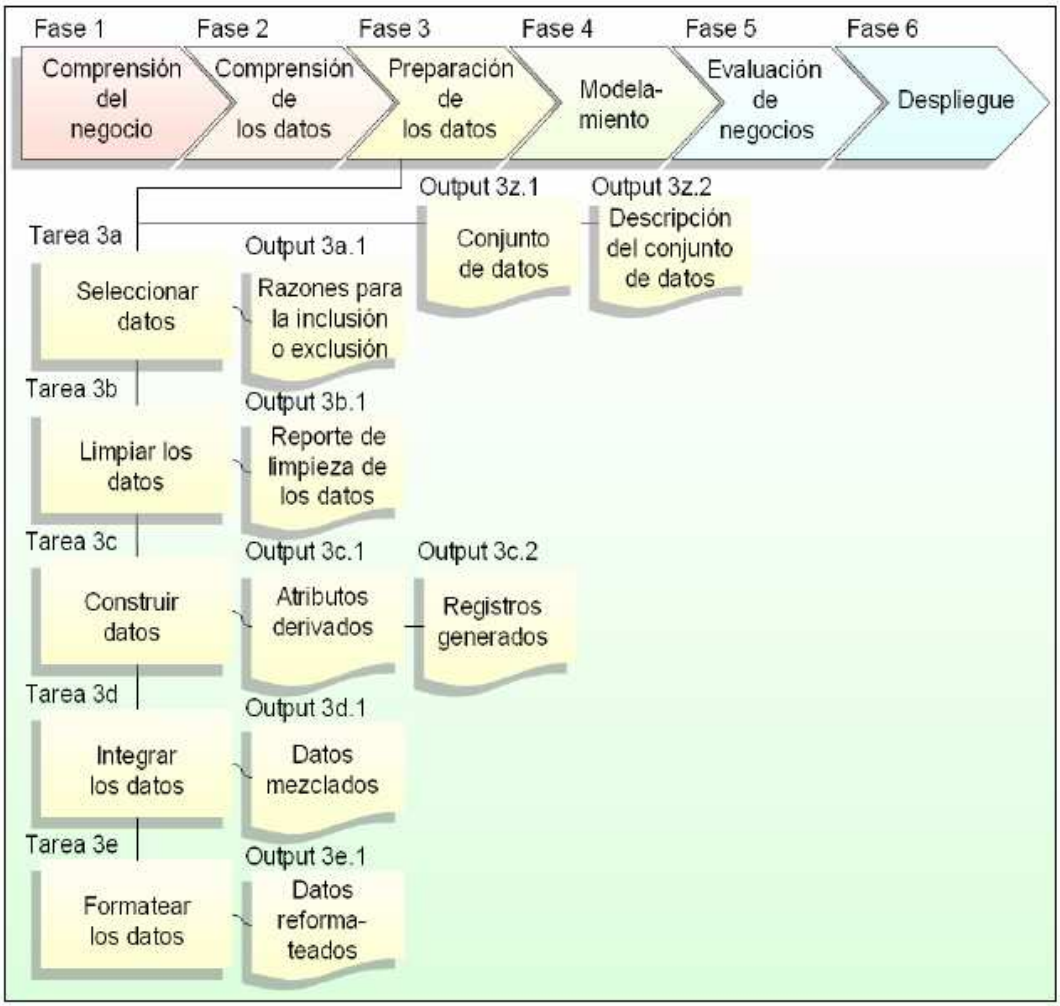
En esta tarea se realizan modificaciones sobre los datos que no alteran su significado, pero que puedan ser necesarias para trabajar con la herramienta de minería [Luc6].

Cuando la información del mismo tema es recolectada de múltiples fuentes, y cada fuente tiene su forma de representar un mismo dato, y no se toma en cuenta esto, se pueden obtener conclusiones erróneas, por lo cual hay que tener en cuenta el formato en el que se esta codificando

[4] Las agregaciones se refieren a las operaciones en donde nuevos valores son calculados por la extracción de información de múltiples registros y/o tablas.

La salida de esta tarea son los datos reformateados. Algunas herramientas tienen requerimientos en el orden de los atributos. También puede ser importante cambiar el orden de los registros en el conjunto de datos, etc.

Figura 5. Preparación de los Datos con actividades y resultados



Fuente: [CCK00], [Luc06]

4. Modelamiento:

Se seleccionan varias técnicas de modelación y se aplican; y los parámetros son calibrados para obtener los valores más óptimos. Hay varias técnicas de minería que se aplican al mismo tipo de problema, por lo cual se trata de escoger aquellas en las cuales pueda obtener el mejor modelo [CCK00].

El proceso de modelado tiene las siguientes actividades:

A. Seleccionar la técnica de modelamiento

El primer paso del modelamiento, es seleccionar la técnica de modelado a utilizar, aunque en etapas anteriores ya se había hablado de herramientas para modelado, en esta tarea se escoge la o las técnicas específicas de modelado [CCK00].

Los resultados de esta tarea son un documento donde se describa la técnica de modelamiento a utilizar y las asunciones específicas de la técnica sobre los datos.

B. Generar diseño de las pruebas

Antes de construir el modelo, se hace necesario generar un procedimiento o mecanismo para probar la calidad y validez del modelo [Luc06].

En esta tarea se describe el plan deseado para el entrenamiento, prueba y evaluación de los modelos. El componente primario del plan es determinar como dividir el conjunto de datos disponibles en los conjuntos de datos de entrenamiento, de pruebas y de validación [CCK00].

Una vez teniendo el conjunto de datos total, la metodología indica que se debe dividir en 3 partes. La primera parte, corresponde al conjunto de entrenamiento, el cual es usado para construir el modelo inicial; la segunda parte, que corresponde al conjunto de validación, es usado para ajustar el modelo inicial para hacerlo mas general y menos sujeto a las idiosincrasias del conjunto de entrenamiento. Por ultimo se tiene el conjunto de pruebas, es usado para calibrar la probable efectividad del modelo cuando se aplica a datos que nunca se han manejado [CCK00].

El conjunto de datos debe ser balanceado de forma tal que haya un porcentaje de registros similares en cada uno de los conjuntos de datos

C. Construir modelo

En este paso se corre la herramienta de modelación sobre el conjunto de datos que se ha preparado para crear uno o más modelos.

Los resultados de esta tarea son: el conjunto de parámetros que necesita la herramienta, los cuales deben ser ajustados, con sus valores seleccionados, junto con la razón de porque se escogieron esos parámetros. El segundo resultado es el o los modelos obtenidos; y por ultimo una descripción de los modelos resultantes, en esta parte se hace una interpretación de los modelos y se documentan las dificultades encontradas y su significado [Luc06].

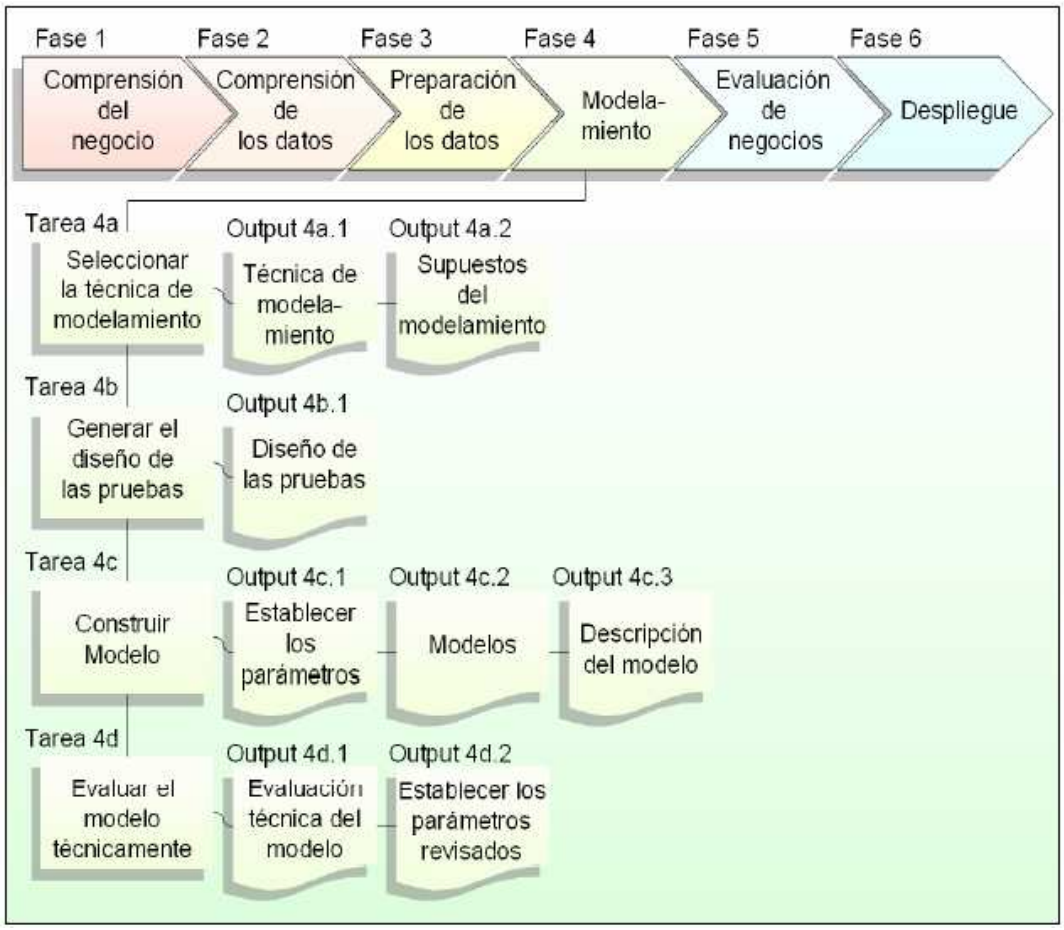
D. Evaluar el modelo técnicamente

En esta parte, los ingenieros de minería de datos interpretan los resultados de acuerdo a su domino del tema, a los criterios de éxito de minería de datos, y el diseño deseado de la prueba. Se juzga el éxito de la aplicación del modelado de una forma técnica. Luego se discuten los resultados con analistas del negocio y expertos del tema en el contexto del negocio.
También se clasifican los modelos. Se evalúan los modelos de acuerdo a los criterios de evaluación, se tiene en cuentan los objetivos del negocio y los criterios de éxito del negocio [Luc06].

El primer resultado de esta tarea es una evaluación del modelo, donde se resumen los resultados de esta tarea, se lista la calidad de los modelos generados.

El segundo resultado, es una revisión de los parámetros y se afinan para la próxima corrida, en la tarea de construcción del modelo, la idea de esto es tratar de encontrar el mejor modelo. Se deben documentar todas las revisiones y evaluaciones [Luc06].

Figura 6. Modelamiento con tareas y resultados.



Fuente: [CCK00], [Luc06]

5. Evaluación:

En esta etapa cuando ya se ha construido un modelo o conjunto de modelos que aparentemente tienen un alto grado de calidad desde la perspectiva del análisis de los datos; es importante hacer una evaluación a fondo del modelo y

revisar los pasos ejecutados para la construcción del modelo y estar seguros de que se lograran correctamente los objetivos del negocio [CCk00].

Una de las formas de evaluar los resultados obtenidos de un proyecto de minería, es evaluando con respecto a costos y beneficios obtenidos con determinado modelo [Luc06]. Este es uno de los mecanismos más utilizados por lo general para evaluar resultados obtenidos con varios modelos.

En el proceso de evaluación se realizan las siguientes actividades:

A. Evaluar los resultados de negocios

Las evaluaciones hechas en tareas anteriores se realizaban en términos de la precisión y generalidad del modelo. En esta tarea se evalúa el grado en el cual el modelo satisface los objetivos del negocio y busca determinar si hay alguna razón del negocio del porque el modelo seria deficiente [CCk00].

El primer resultado de esta tarea es un resumen con los resultados de la evaluación en términos de los criterios de éxito del negocio [Luc06], incluyendo una declaración final diciendo si el proyecto satisface los objetivos iniciales del negocio.

El segundo resultado son los modelos aprobados. Después de la evaluación de los modelos respecto a los criterios de éxito, los modelos generados que cumplen con estos criterios se convierten en aprobados [Luc06].

B. Revisar los procesos

En este punto es apropiado hacer una revisión más minuciosa del proceso de minería de datos para determinar si una tarea o factor importante ha sido pasado por alto [Lar06].

El resultado de esta tarea es un resumen de la revisión del proceso de minería [Luc06], donde se especifiquen las actividades en las que se ha errado y aquellas que deberían ser repetidas.

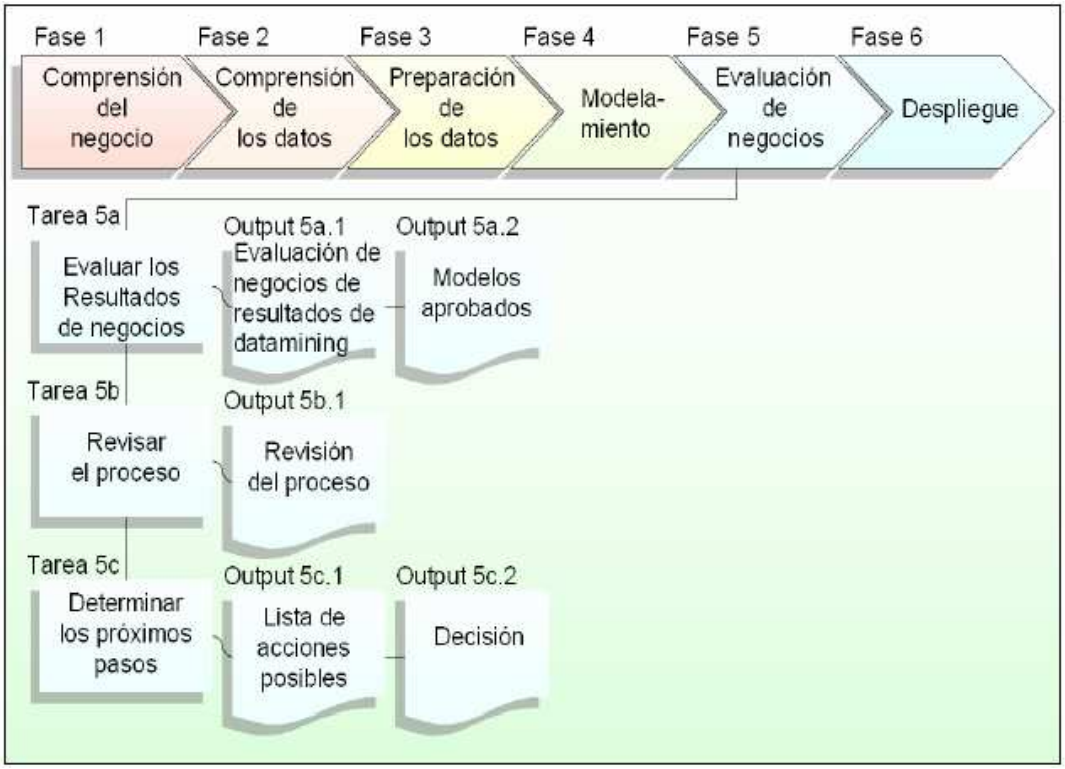
C. Determinar los próximos pasos

De acuerdo a los resultados de las revisiones y evaluaciones hechas, el equipo de proyecto debe decidir como proceder. Deben decidir si terminar el proyecto y pasar a la etapa de despliegue, iniciar más iteraciones o levantar nuevos proyectos de minería de datos. Esta tarea incluye analizar las sobras de recursos y de presupuesto, ya que esto puede influenciar las decisiones a tomar [CCk00].

El primer resultado de esta tarea es una lista de las posibles acciones a tomar, junto con las razones de pro y en contra de cada una de esas acciones.

El segundo resultado es la decisión en cuanto ha como proceder, junto con el análisis razonado.

Figura 7. Evaluación con sus tareas y resultados



Fuente: [CCk00], [Luc06]

6. Ejecución:

La creación del modelo no es el último paso del proyecto; se deben presentar los resultados organizados y de una forma en que el usuario final pueda utilizarlos. Depende de los requerimientos, la fase de despliegue puede ser tan simple como generar un reporte o tan compleja como implementar otro proceso de minería de datos en la empresa [CCK00].

La importancia y el éxito de un proyecto de minería de datos se mide en términos de cómo pueden ser aplicados los resultados obtenidos al negocio como tal. La idea no es conseguir un resultado teórico el cual no vaya a ser aplicado en el ambiente laboral y que solo haya servido como una experiencia más [Luc06].

Uno de los conceptos más manejado para medir el éxito de un proyecto de minería de datos es el del retorno sobre la inversión (ROI), pues la parte

directiva de una empresa esta siempre interesada en saber cuanto se ganara en términos económicos con el proyecto. Además es muy difícil hacerle entender a un directivo que las ganancias con un proyecto de este tipo no solamente se miden en dinero. Aunque esta es una buena medida para calificar el éxito de un proyecto de minería de datos desde ciertos puntos de vista, existen otros criterios para medir el éxito [Luc06].

En esta etapa se mide el impacto que tuvieron los resultados obtenidos en las diferentes áreas de la empresa, los cuales presentaron una relación directa o indirecta con el desarrollo del proyecto. La idea es que se logre un beneficio autentico aplicando estos resultados, que apoyen los procesos del negocio y los mejoren, tratando de darle a la empresa una ventaja competitiva.

Las actividades que se realizan en la fase de ejecución son:

A. Planificar el despliegue

Esta tarea toma los resultados de la evaluación y determina una estrategia para el despliegue en el negocio [CCK00].

Se deben resumir los resultados obtenidos, desarrollar y evaluar planes alternativos de despliegue, determinar como se va hacer para propagar el conocimiento o la información obtenida hacia los usuarios, como se van a medir los beneficios, establecer como el modelo o software resultante será desplegado dentro de los sistemas de la organización y por ultimo tratar de identificar posibles problemas durante el despliegue.

El resultado de esta tarea es un resumen donde se describa la estrategia de despliegue, que incluya los pasos necesarios y como ejecutarlos [Luc06].

B. Planificar el monitoreo y la manutención

En esta tarea debe determinarse como se monitoreara la precisión del modelo y sus resultados, y cuando deben dejar de utilizarse estos. Un buen plan de manutención ayuda a evitar el peligro del uso incorrecto de los resultados de la minería de datos [CCK00].

La salida es un resumen de la estrategia de monitoreo y manutención, incluyendo los paso necesarios para hacerlo y como ejecutarlos.

C. Producir el reporte final

Al final del proyecto, el equipo hace un reporte final. Dependiendo del plan de despliegue este reporte puede ser solo un resumen del proyecto y las experiencias o puede ser una presentación final de los resultados de minería [CCK00].

Esta tarea tiene dos resultados. El primero es un reporte final del proceso de minería de datos e incluye todos los entregables previos, a la vez que resume y organiza sus resultados. El segundo es una reunión al final del proyecto en la cual los resultados del proyecto de data mining son presentados a su patrocinador.

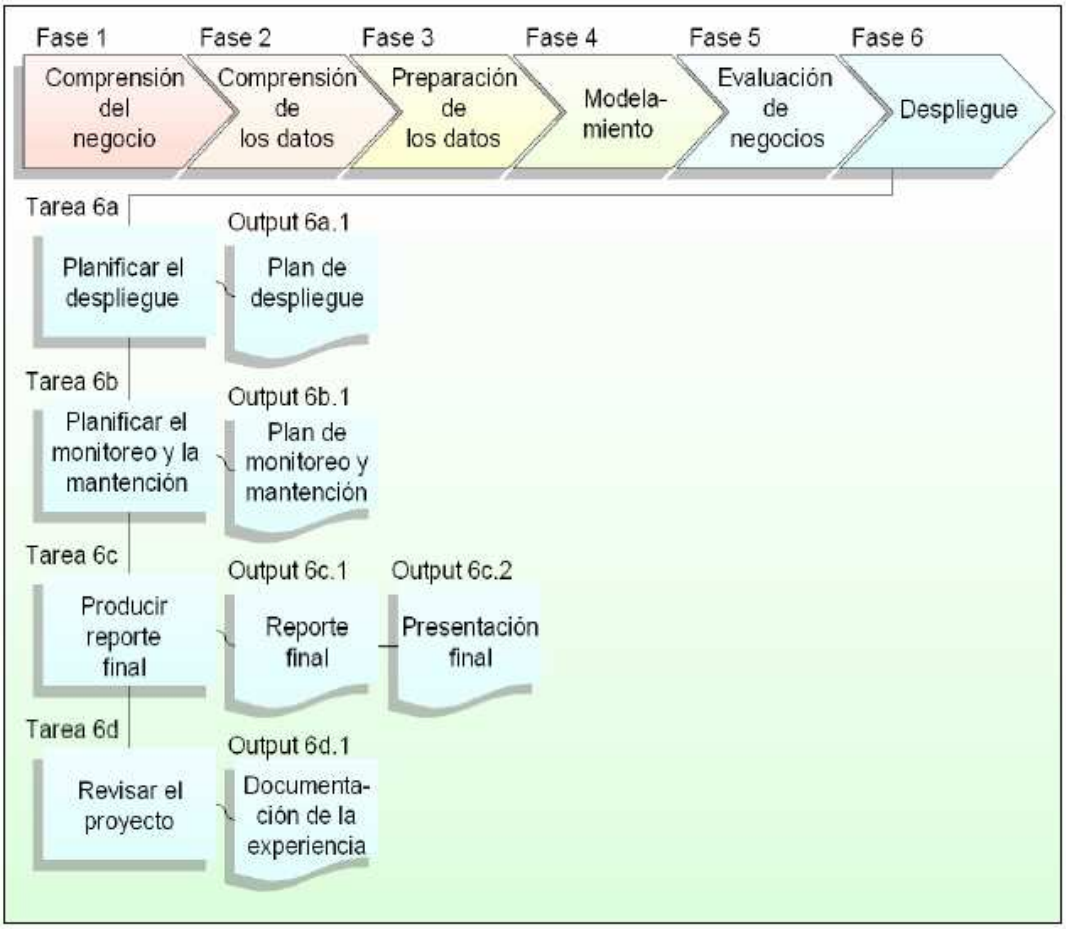
D. Revisar el proyecto

Esta tarea corresponde a evaluar lo que ocurrió directamente y lo que ocurrió mal en el proyecto [CCK00].

Es demasiado importante para proyectos futuros documentar la experiencia obtenida. Por ejemplo peligros o trampas, acercamientos engañosos, o tips para escoger las mejores técnicas de minería y los reportes hechos.

El resultado de esta tarea corresponde a un resumen con las experiencias más importantes adquiridas durante el proyecto.

Figura 8. Ejecución con sus tareas y resultados



Fuente: [CCK00], [Lac06]

La importancia de este capitulo radica en que la compañía u organización que pretenda implementar una solución de minería de datos, para solucionar uno o varios problemas del negocio; tendrá una imagen mas clara de cuales son los métodos o problemas que se presentan en el campo de la minería, los pasos a seguir en un proceso convencional de minería de datos de acuerdo a las metodologías vistas y cuales son las técnicas o algoritmos que usualmente se utilizan para resolver estos problemas.

Las mejores practicas presentados en este capitulo se realizaron con base en diferentes autores y fuentes recopiladas durante la investigación, pues se considera que una de las mejores formas de reunir estos tips es basado en experiencias con definiciones e implementaciones realizadas y aportes de diferentes expertos.

Se puede decir que la mejor práctica a seguir, es utilizar una metodología de minería de datos y hacer cada una de las tareas o actividades establecidas en cada una de sus etapas, pues es la mejor forma de llegar a unos buenos resultados. Además estas metodologías no han sido fruto de poco trabajo, se han ido refinando a medida de que han sido utilizados en los diferentes campos de la industria, por lo que el mejor camino para lograr los objetivos con un proyecto de minería de datos, es seguir los consejos o tips de autoridades en este campo.

2. CASOS REALES DE APLICACIÓN DE MINERÍA DE DATOS A NIVEL MUNDIAL

Este capitulo, trata algunos casos de aplicación de minería de datos en ciertos sectores de la industria, utilizando algunas de las técnicas posibles según la tarea que se tratando y algunas de las herramientas mencionadas en el capitulo anterior.

Para la construcción de los casos, como fue realmente complicado encontrar los casos reales bien documentados, acerca de todo el proceso de minería de datos, se trataron de construir utilizando diferentes fuentes bibliográficas sobre el tema especifico que se estaba tratando.

El desarrollo de cada caso se hizo utilizando la metodología CRISP-DM aunque no con todo el detenimiento y nivel de detalle que exige cada fase del ciclo, con sus respectivas tareas y salidas. Por tal razón algunas de las etapas fueron presentadas en un mismo paso en el proceso, según el grado de similitud de las etapas. Esto se dio sobre todo con las etapas de entendimiento y preparación de los datos y las etapas de modelado y evaluación.

La forma como se desarrollo cada caso fue similar a los demás. En ciertas fuentes se encontraba el método que se podía seguir para realizar un proyecto de minería de datos en una determinada industria y para cierta tarea especifica. Lo que se hizo fue ajustar ese método a algunos de los casos encontrados, que cumplían en cierta forma con los requerimientos y las características del negocio; pero cuya documentación era muy insuficiente. Por lo que se puede decir que aunque las bases son de casos reales de proyectos de minería de datos de algunas empresas

del mundo, no se puede verificar que de esta forma como se presentan los casos y como se construyeron, si fue la forma que la empresas emplearon para levantarlos.

Lo que se muestra es un camino posible a seguir para construir proyectos de este tipo, siguiendo una metodología, la cual se considera un estándar para este tipo de tareas y utilizando una de las herramientas especificas de minería de datos.

Los nombres utilizados en los distintos casos son ficticios. Pues como se mencionaba anteriormente, la documentación que se consiguió de los datos reales fue muy poca como para poder ilustrar el proceso de minería que estas empresas siguieron, por lo cual se prefirió usar otros nombres, pero utilizando cierta información de las fuentes reales.

2.1. PERFILES DE PACIENTES CON TROMBOSIS PARA DETERMINAR SIMILITUDES DE COMPORTAMIENTO

Este caso muestra el proceso que se siguió para sacar perfiles de pacientes que tienen trombosis y que tienen pronósticos o comportamientos similares.

La tarea de minería de datos que se utilizo en este caso fue la de segmentación y se utilizo la herramienta DB2 Intelligent Miner for Data para el desarrollo del proyecto de IBM.

2.1.1. Comprensión del negocio o del tema a tratar:

Calit health services[5], con mas de 8000 médicos, 14 hospitales, 1000 clínicas y mas de 400 farmacias; es la organización de manejo de la salud mas grande de Israel. Esta organización provee servicios de salud a mas de 3.7 millones de miembros asegurados, lo cual representa un 60% de la población israelí.

La organización siempre esta buscando caminos para incrementar el bienestar de sus miembros. Para esto almacenan la información de sus pacientes en formato electrónico; y buscan la forma de usar esos datos de una forma más eficiente, que le permitan mejorar la calidad de salud de sus beneficiarios.

Una alternativa fue realizar estudios de los pacientes que fueron diagnosticados con trombosis de vena profunda (DVT, por sus siglas en ingles), para describir un método que permita hacer perfiles de este tipo de pacientes.

Trombosis de vena profunda[6]: es una condición donde un coagulo de sangre se forma en una vena del sistema mas interno de venas con el que cuenta el ser humano. Este sistema de venas abarca el conjunto de venas que están dentro de los músculos del cuerpo.

Según lo visto en el libro Mining Your Own Business in Health Care Using DB2 Intelligent Miner for Data, la DVT puede ocurrir en cualquier parte del cuerpo, pero es más común encontrarlas en las venas de las piernas y de los muslos. Muchos factores pueden causar DVT incluso una lesión en la vena, la reducción del flujo de sangre y las condiciones que incrementan la tendencia de que la sangre se

[5] Para ver la información real de la organización base, se puede dirigir al documento: Microsoft SQL Server 2005 data mining helps Clalit preserve health and save lives. [SQL05]
[6] Ver referencia bibliografica [BAB+01]

coagule. La causa más común de herir una vena es un trauma en la pierna que ocurre cuando hay fracturas de huesos, una herida severa del músculo. La inmovilización es la causa más común para que haya una reducción del flujo de sangre en la vena, porque el movimiento de los músculos del pie ayuda a mantener constante el flujo de sangre en las venas. Las personas que pueden tener una o más de las siguientes condiciones están en alto riesgo de desarrollar DVT:

- Parálisis causada por una apoplejía o por una lesión de la medula espinal.
- El enyesado en una pierna para curar un hueso fracturado.
- El confinamiento a una cama debido a sugerencia medica.
- Sentarse prolongadamente con las piernas cruzadas.

Según las investigaciones realizadas por Baragoin, Andersen, Bayer y compañia los síntomas más comunes de la DVT en una pierna son hinchamiento y dolor de la pierna afectada. Estos síntomas son causados por la acumulación de sangre que es incapaz de pasar el coagulo en la vena y el resultado es una fuga de fluido de la sangre en el músculo. Una DVT presenta una cierta dificultad al momento de diagnosticar sin pruebas específicas en donde el sistema de venas internas sea examinado, esto porque muchas otras condiciones exhiben síntomas similares a la DVT. Además muchos pacientes no presentan síntomas, a menos de que el coagulo se desplace hasta el pulmón y cause una embolia pulmonar. En este caso el paciente puede desarrollar taquicardia, fallas de respiración, dolor agudo en el pecho que empeora con las fallas de respiración y tos con sangrado. Si la embolia pulmonar es grande y bloquea una o ambas de las principales arterias enviando sangre a los pulmones, el paciente puede desarrollar una muy baja presión sanguínea, desmayos, y posiblemente la muerte por una falla del corazón o los pulmones.

Estos autores también mostraron que existen diversas pruebas cada una de las cuales tiene ciertas ventajas y limitaciones, para diagnosticar la presencia de DVT. La más antigua de estas pruebas es la venografia. Esta prueba es hecha inyectando un fluido opaco a la radiación en una vena que por lo general esta localizada en la parte superior de pie. El tinte fluye con la sangre y llena las venas de la pierna así como las del muslo. Un coagulo obstructor en una de estas venas puede ser visto mediante rayos X, como un área libre de color dentro de la vena. La venografia es el examen mas preciso para identificar DVT, pero es doloroso, caro y ocasionalmente causa una inflamación dolorosa en las venas. Además requiere de un alto grado de habilidad para ejecutar e interpretarlo correctamente.

El reto fue realizar un proyecto de minería de datos para entender como tratar de mejorar el cuidado de los pacientes.

El primer desafió fue encontrar grupos de pacientes que fueron diagnosticados con DVT tratando de encontrar similitudes en su comportamiento. Y también se espera detectar indicadores nuevos y útiles que se puedan derivar del análisis que se va a realizar.

La pregunta[7] que se formulo y que se quiso responder con este estudio fue: ¿que grupo de pacientes comparten una conducta similar durante un diagnostico de DVT?

2.1.2. Comprensión y preparación de los datos a utilizarse:

Los datos fundamentales fueron recolectados por la red de hospitales. Cada paciente era enviado al centro medico como recomendación por su doctor o por un terapeuta del hospital. Luego la persona se examina para comprobar si tiene

[7] Ver capitulo 3 del libro de referencia. [BAB+01]

trombosis usando las pruebas estándar. Como resultado el paciente obtiene información de si tiene trombosis o no. [BAB+01]

Según Baragoin, Andersen, Bayer y compañia los siguientes fueron los datos utilizados:

- Datos demográficos
- Datos de pruebas médicas especiales que fueron hechas por el hospital.
- Datos históricos de pruebas de sondeo.

Mientras que los datos demográficos son más fáciles de entender**,** los datos de las pruebas medicas y los datos históricos necesitan una descripción un poco mas detallada. A continuación se veran las variables utilizadas en el proyecto. [BAB+01]

Datos demográficos: los datos demográficos contienen información general del paciente. Los datos incluyen todos los pacientes y contiene aproximadamente 1000 registros. [BAB+01]

Tabla 2. Lista de variables demográficas

Variable	Tipo de Variable	Descripción
Id	Categórica	Identificación del paciente
Sexo	Categórica	Genero del paciente
cumpleaños	Categórica	Cumpleaños del paciente
Date_of_rec	Categórica	El primer día que el paciente fue registrado
Date_of_del	Categórica	La fecha cuando el paciente fue al hospital.
Admisión	Categórica	Indica si el paciente fue admitido en el hospital o fue remitido desde otro.
Dia_2	Categórica	Diagnostico

Fuente: [BAB+01]

Datos de las pruebas medicas: los datos son exámenes de laboratorio que fueron medidos por el hospital. Estos datos solo cubren un subgrupo de los pacientes. A continuación se muestra la lista de variables que fueron registradas cuando una prueba de trombosis se realizaba en el hospital. [BAB+01]

Tabla 3. Lista de las variables de pruebas medicas.

Variable	Tipo de Variable	Descripción
Id	Categórica	Identificación del paciente.
EXAMINATION_ DATE	Categórica	Fecha cuando el examen fue hecho
ACL_IGL	Numérica	Anti-cardiolipin concentración de inmunoglobina tipo G.
ACL_IGM	Numérica	Anti-cardiolipin concentración de inmunoglobina tipo M.
ACL_IGA	Numérica	Anti-cardiolipin concentración de inmunoglobina tipo A.
ANA_PATTERN	Categórica	Concentración de anticuerpos antinucleares.
Diagnostico	Categórica	Diagnostico.
KCT	Binaria	Medida del grado de coagulación para KCT
RVVT	Binaria	Medida del grado de coagulación para RVVT.
LAC	Binaria	Medida del grado de coagulación para LAC
síntomas	Categórica	Otros síntomas observados
trombosis	Categórica	Trombosis

Fuente: [BAB+01]

Los siguientes, son algunos de los términos que los autores definieron para tener un mejor conocimiento de las variables de pruebas médicas:

Inmunoglobina (lg): es una proteína producida por los linfocitos B en su forma unida a la membrana celular. Inmunoglobulina es una parte esencial del sistema inmunológico que ataca sustancias extrañas o agentes como bacterias. Algunas clases de inmunoglobulinas son por ejemplo la A, M y G.

ANA: es la abreviatura de anticuerpos antinucleares que están dirigidos contra la estructuras dentro del núcleo de las células. El núcleo es el centro interno dentro de cada célula del cuerpo, este contiene el material genético.

Los ANA son encontrados en pacientes que cuyo sistema inmunológico puede estar predispuesto para causar inflamación contra su propio tejido. La predisposición para que el sistema inmune trabaje contra su propio cuerpo hace referencia a la auto inmunidad.

KCT, RVVT y LAC: son variables binarias que indican cualquier valor de análisis sea alto o bajo.

Síntomas: se refiere a otros diagnósticos, contiene por ejemplo si el paciente presento infarto cerebral, epilepsia, o CVA. CVA es la muerte súbita de algunas células del cerebro, debido a la falta de oxigeno, cuando el flujo de sangre al cerebro se ve deteriorado por un bloqueo o ruptura de una arteria al cerebro.

Pruebas medicas históricas: los datos históricos están basados en resultados de análisis que fueron hechos en el pasado. A continuación se muestran las variables de estas pruebas. [BAB+01]

Tabla 4. Lista de las variables de pruebas medicas históricas.

Variable	Tipo de Variable	Descripción
ID	categórica	Identificación del paciente
Fecha	categórica	Fecha cuando el examen fue realizado
GOT	continua	trans aminasa glutamin oxaloacetica
GPT	continua	trans aminasa glutamin pylvic
LDH	continua	lactato deshidrogenada
ALP	continua	Fosfato alcalino
TP	continua	Numero total de proteínas.
ALB	continua	Albumina
UA	continua	Acido urico
UN	continua	Nitrogeno ureico
CRE	continua	Creatinina
T-BIL	continua	Bilirrubina
T-CHO	continua	Colesterol
TG	continua	Triglicéridos
CPK	continua	Creatinina fosfoquinasa
C4	continua	Complemento 4
WBC	continua	Numero de glóbulos blancos en un volumen de sangre

RBC	continua	Numero de glóbulos rojos.
HGB	continua	Hemoglobina
HCT	continua	Hematocritos
PLT	continua	Plaquetas
IGG	continua	Inmunoglobina G.

Fuente: [BAB+01]

Los siguientes, son algunos de los términos que los autores definieron para tener un mejor conocimiento de las variables de pruebas médicas históricas:

LDH (lactato deshidrogenasa): es una enzima que cataliza la conversión de lactato a piruvato. Este es un importante paso en la producción de energía en las células. Muchos tipos diferentes de células en el cuerpo contienen esta enzima. Algunos de los órganos relativamente ricos en LDH son el corazón, los riñones, el hígado y los músculos.

T-ALB (albúmina): es la principal proteína de la sangre humana y la llave para la regulación de la presión osmótica de la sangre.

Creatinina: es una molécula de desperdicio químico que es generada por el metabolismo del músculo. Es producida por la creatina, una molécula de mayor importancia para la producción de energía en los músculos. Creatinina es transportada a través del flujo de sangre a los riñones. Los riñones filtran la creatinina y se deshace de esta en la orina.

T-BIL (bilirrubina): es un componente de color amarillo-naranja producido por la falla de hemoglobina desde RBC.
RBC (glóbulos rojos): son las células que cargan el oxigeno y el dióxido de carbono a través de la sangre.

HGB (hemoglobina): es un pigmento en los glóbulos rojos. Forman un lazo irreversible con el oxigeno. En este estado oxigenado son llamados

oxihemoglobina y tienen un color rojo brillante. En estado reducido son llamados de oxihemoglobina y son de color azul-violeta.

HCT (hematocritos): es la proporción, por volumen de los glóbulos rojos en las sangre.

PLT (plaquetas): son las estructuras celulares más pequeñas presentes en la sangre y son importantes para la coagulación de la sangre y para el taponamiento de daños en los vasos sanguíneos.

Evaluación de los datos:

Para la evaluación de los datos Baragoin, Andersen, Bayer y compañia utilizaron la estadística descriptiva observaron el comportamiento de las diferentes variables y los posibles valores que podian tener. Por ejemplo lograron ver que los datos contienen registros con aproximadamente el 80% de tipo femenino y el 20% restante masculino. Para la variable Admisión, cerca del 60% de los pacientes eran pacientes desconocidos o externos al hospital y el otro 40% fueron admitidos. Las variables Date_of_rec y Date_of_del tenian muchos valores asociados; por consiguiente seria adecuado tener en cuenta tener otras variables alternativas que fueran mas fáciles de utilizar.

De acuerdo a la inspección de los datos que ellos realizaron de las pruebas médicas concluyeron lo siguiente:

- KCT, LAC, y RVVT pueden asumir los valores alto o bajo.
- ANA es expresada por una D, S, P o una combinación de estos valores.
- Trombosis tiene varios valores:
    - Un valor de 0 indica que el paciente no tiene trombosis.
    - Un valor de 1 indica que el paciente tiene trombosis en un alto grado.

  - Un valor de 2 indica que el paciente tiene trombosis en un mediano grado.
  - Un valor de 3 indica que el paciente tiene trombosis en un bajo grado.
- Examination_Date se refiere a la fecha cuando el paciente fue examinado y cuando la prueba fue hecha.

La distribución de los datos históricos también la realizaron con estadísticas descriptivas. La mayoría de estas variables contienen valores que no fueron tomados como resultados de las pruebas, pero fueron utilizados con el fin de superar el problema de los valores ausentes. [BAB+01]

Construcción del centro de datos:

Para el análisis fue necesario unir los datos de las diferentes fuentes, por lo que se obtuvieron en total un centro de datos con 39 variables y aproximadamente 20000 registros. [BAB+01]

Para los autores fue necesario considerar que variables iban a utilizar y como. Por lo que tuvieron que calcular nuevas variables. Por ejemplo la variable date_of_rec contiene 98 datos diferentes, y la variable date_of_del tiene 791 datos diferentes. Como se sospecho que estas variables podrían ser útiles, procedieron a definir otras variables las cuales fueron medidas en años:

- DeliveredSinceN: indica el año cuando fueron enviados al hospital.
- RecordedSinceN: indica el año cuando el primer registro fue hecho.

Existia un problema similar con las variables Examinacion_Date la cual tiene 454 valores diferentes y Diagnostico tiene 187 valores diferentes. Ambas son variables categóricas, y el uso de este tipo de variables con muchos valores es un factor

muy caro al momento de comenzar un proceso de minería de datos; por lo que introdujeron 2 nuevas variables para usar en vez de estas. [BAB+01]

- ExaminateSinceN indica el año cuando el análisis fue hecho.
- Dia2P: contiene términos médicos mas generalizados, como es especificado en el diagnostico.

"La ventaja de contar con variables numéricas, es que son más fáciles de Interpretar".

En el caso de las variables históricas, era mejor tener una buena claridad en las variables. Esto porque los valores normales y anormales eran conocidos, y era mas fácil discretizar las variables correspondientes. Esas discretizaciones fueron aplicadas con base a las fuentes médicas convenientes, encontradas por los autores.

Tabla 5.Rangos normales de las variables, mediante los cuales se hizo la discretizacion de algunas variables.

Variable	Rangos Normales
GOT_D	GOT < 60
GPT_D	GPT < 60
LDH_D	LDH < 500
ALP_D	ALP < 300
TP_D	TP_D 6.0 < TP < 8.5
ALB_D	3.5 <= ALB < 6.5
UA_D	UA > 6.5
UN_D	UN < 30
CRE_D	CRE < 1.5
TBIL_D	TBIL < 2.0
TCHO_D	TCHO < 250
TG_D	TG < 200
GOT_D	GOT < 60
CPK_D	CPK < 250
C4_D	C4 > 10
WBC_D	3.5 <= WBC < 9.0
RBC_D	3.5 <= RBC < 6.0
HGB_D	10 <= HGB < 17
HCT_D	29 <= HCT < 52
PLT_D	100 <= PLT < 400

IGG_D	900 <= IGG < 2000

Fuente: [BAB+01]

2.1.3. Modelado y Evaluación:

Como se quería hacer perfiles de cierto tipo de pacientes, se utilizaron algoritmos de segmentación; pero como la segmentación es un proceso tedioso, porque desde el punto de vista técnico las variables pueden estar correlacionadas o ser redundantes. Para poder aplicarla se hizo necesario desarrollar una estrategia que permitiera encontrar un buen modelo de datos. [BAB+01]

Lo que hicieron Baragoin y compañía fue revisar cuidadosamente las variables, y en caso de alguna incorrección, estas se transformaban.

Algunas veces se perdían algunas variables porque estaban corruptas o por la correlación. En el caso de la correlación se aplicaba una función estadística llamada análisis de componente principal (Principal Component of Analysis (PCA[8]) por sus siglas en ingles) para tratar de arreglar este tipo de problemas, el inconveniente de esto, era que esta función solo se aplicaba a variables numéricas; por lo que no servia para las variables categóricas. [BAB+01]

Con el fin de solucionar este inconveniente se utilizaron árboles de clasificación[9] en la búsqueda de variables categóricas apropiadas. La ventaja de los árboles de clasificación es que usa una variable que divide el centro de datos de la mejor forma. Las variables que tienen menor importancia en este sentido son muy pocas veces tenidas en cuenta y en ocasiones ni se utilizan. [BAB+01]

[8] PCA: es una técnica usada para simplificar un conjunto de datos.
[9] Los árboles de clasificación o también llamados de decisión son una forma de representar el conocimiento obtenido en el proceso de aprendizaje inductivo.

La estrategia que se aplico con los árboles de clasificación para la reducción de las variables fue la Bottom Up[10], que básicamente realiza lo siguiente:

- Divide las variables categóricas en conjuntos distintos y ejecuta una clasificación para cada conjunto de variables.
- Usa solo las mejores variables en el árbol de clasificación y remueve aquellas que no aparecen en el árbol.
- Abstrae las variables categóricas resultantes y chequea el modelo en conjunto.

En cada uno de los grupos de datos que se tenían se aplico esta técnica y los resultados que encontraron los autores fueron los siguientes:

Grupo de variables demográficas

Para este grupo se usaron las siguientes variables:

- Admisión
- RecordedSinceC (DATE_OF_REC)
- Sexo
- Cumpleaños

Dia2P (DIA_2) no pudo ser utilizada porque era obvio que estaba relacionada con la variable trombosis. Lo mismo ocurrió con DeliveredSinceC (DATE_OF_DEL), porque todos los pacientes que fueron diagnosticados con trombosis fueron enviados al hospital. Por consiguiente el grupo de variables demográficas solo tiene 4 variables.

[10] Ver el libro Mining Your Own Business in Health Care Using DB2 Intelligent Miner for Data para tener una mejor descripción de esta técnica.

Aplicando la técnica de árboles de clasificación se obtuvo que la variable más importante fue cumpleaños, seguida de Admisión, RecordedSinceC y sexo.

Sin embargo, cumpleaños es una variable muy predomínate y puede que esta variable tenga algún impacto en la variable trombosis. Por lo que por su predominancia se dejara en la lista de variables a utilizar, pero se transformara a la variable numérica Edad (Age), para tratar de minimizar su fuerte correlación con otras variables.

Grupo de variables de Pruebas

Para este grupo se utilizaron las siguientes variables:

- ExaminedSinceC (Examination_Date)
- ACL_IGL
- ACL_IGM
- ACL_IGA
- ANA
- KCT
- RVVT
- LAC

La variable síntomas no puede ser utilizada porque esta contiene valores que están correlacionados con algunas etiquetas de la variable trombosis.

El árbol de clasificación resultante presentado en [BAB+01], que se construyo con las variables seleccionadas del grupo de variables de pruebas, arrojo los siguientes resultados: la variable más importante fue la variable LAC, seguida por ACL_IGA y ANA. ExaminateSinceC y ACL_IGM aparecen en el tercer y cuarto nivel respectivamente. Sin embargo RVVT, KCT y ACL_IGG no aparecen en el

árbol. Estas variables fueron removidas de la lista de variables categóricas, aunque se dejaron como variables suplementarias.

Según el resultado[11] que arrojo una matriz de confusión[12], se veía que algunas correlaciones podían presentarse en el futuro y se sospechaba que la variable LAC era la razón de la correlación por lo cual se procedió a removerla.

Se volvió a construir otra matriz de confusión, pero de nuevo se volvió a observar que había una alta correlación entre las variables, por lo que se decidió remover todas las variables del grupo de pruebas y mantenerlas como variables suplementarias

Grupo de variables Históricas

Para este grupo se utilizaron todas las variables. Para este caso no se mostrara el árbol resultante el cual puede ser encontrado en la referencia [BAB+01].

La variable más importante fue TP_D, seguida por TCHO_D y RBC_D. TBIL_D, C4_D, GOT_D, CPK_D, CRE_D siguen en los siguientes niveles. Aunque había muchas variables categóricas, todas aparecen en el árbol. Por lo cual se usaran todas las variables para la construcción del modelo de segmentación.

Después de haber hecho esto para cada grupo de variables, se combinaron las variables resultantes de los 3 grupos y se creo un nuevo árbol de clasificación que puede ser encontrado en [BAB+01]. El resultado de este nuevo árbol genero un modelo donde la variable Admisión se encuentra en la cima del árbol, seguida por

[11] Resultado encontrado en la fuente bibliografica [BAB+01], pero el cual no fue sustentado.
[12] También es utilizada para encontrar variables que están altamente correlacionadas.

las variables TP_D y RecordedSinceC. La calidad de este modelo[13] fue muy aceptable, por lo que no fue necesario hacer más reducciones de variables.

Aplicación de las técnicas de minería de datos:

De acuerdo al proceso de clasificación que se realizo anteriormente, se utilizaran las variables resultantes de este, así como la variable numérica edad. El resto de las variables que fueron removidas, fueron puestas como variables suplementarias. [BAB+01]

Para realizar la segmentación y encontrar segmentos con pacientes que comparten un comportamiento similar para la trombosis se utilizo la técnica del clustering demográfico con el valor condorcet [BAB+01].

Clustering de los vecinos más cercanos con Condorcet

Este algoritmo que también es conocido como clustering demográfico[14], fue una implementación hecha por IBM, del algoritmo estándar de los vecinos más cercanos, trabajando además con un valor condorcet. Por lo cual se convierte en algoritmo propietario.

El cluster demográfico opera en primer lugar sobre registros de variables categóricas. Usan una técnica de medida de distancia la cual se basa en el principio del condorcet.

[13] El nivel de calidad fue de aproximadamente 81%.Pero no tenemos la evidencia bibliográfica para juzgar a partir de que porcentaje el modelo se puede considerar como de buena calidad. Solo nos basamos en lo que encontramos en la fuente [BAB+01].
[14] Ver [BrM05].

Valor condorcet[15]: el valor condorcet de un clustering demográfico es un valor de calidad que mide la bondad de los registros perteneciente a un cluster. Este valor es calculado de la siguiente forma:

- Normalizan todas las variables numéricas.
- Se toman los dos primeros registros y se comparan para cada variable.

    - Si el valor de la variable es el mismo, entonces se aumenta en uno un valor de similitud que se tiene, en caso contrario, no.
    - Si todos los valores de las variables comunes se comparan, entonces estos registros, consiguen un valor de similitud que demuestra la semejanza entre estos dos registros.

- Este procedimiento es hecho para todos los pares de registros. Luego se obtiene una matriz de similitud de tamaño N veces N, donde N es el numero de variables; donde todas las entradas de la diagonal contienen un numero N, esto porque cada registro es exactamente similar a el mismo. Las demás entradas tienen un valor menor a N.

- Luego se reestructura la matriz de similitud dependiendo de cual registro pertenece a que cluster.

El valor condorcet es luego calculado sumando todas las entradas, de cada cluster identificado, dividido por el valor máximo de similitud; por lo cual el valor condorcet obtenido se encuentra entre 0 y 1.

Para realizar la segmentación por medio de clustering demográfico estándar, fue necesario tener en cuenta los siguientes parámetros:

[15] Ver [BAB+01] para complementar la información sobre esta técnica.

- El numero máximo de clusters que son permitidos.
- Similitud: es el umbral que limita los registros aceptados como los más adecuados para el cluster. Si la similitud es alta, entonces los registros que satisfacen el umbral pueden ser adicionados al cluster.

A continuación se mostrara una tabla que resume los resultados de algunos modelos que se construyeron, en la búsqueda de ese mejor modelo.

Tabla 6. Ejecuciones realizadas para encontrar el mejor modelo de clustering.

No	No. clusters	Similitud	Condorcet	Modificación de variables.
1	9	0.5	0.669	NO
2	9	0.6	0.711	NO
3	9	0.7	0.744	NO
4	9	0.8	0.753	NO
5	9	0.85	0.749	NO
6	9	0.9	0.738	NO
7	7	0.6	0.709	NO
8	20	0.8	0.801	NO
9	30	0.8	0.817	NO
10	50	0.8	0.831	NO
11	100	0.8	0.841	NO
12	200	0.8	0.844	NO
13	500	0.8	0.846	NO
14	9	0.5	0.644	SI. Las variables suplementarias fueron utilizadas
15	20	0.8	0.758	SI. Las variables suplementarias fueron utilizadas
16	50	0.8	0.799	SI. Las variables suplementarias fueron utilizadas

Fuente: [BAB+01]

Para los autores fue interesante observar que los modelos 14, 15 y 16 tienen un peor condorcet que los demás modelos que no utilizaron las variables suplementarias. Lo que indica que le pre-procesamiento con los árboles de clasificación tuvo un impacto positivo en la generación de los modelos.

El mejor modelo resultante de acuerdo a los parámetros utilizados, es la ejecución numero 13. El cual da un valor condorcet de 0.846, que es el más alto y que aparentemente dio una segmentación casi perfecta.

El problema que se ve con este modelo, es que tiene un total de 500 clusters, lo que hace que sea muy específico, y donde probablemente cada cluster tiene un tamaño con menos del 1% de la población. Además la legibilidad e interpretación de cada cluster se hace cada vez más difícil. [BAB+01]

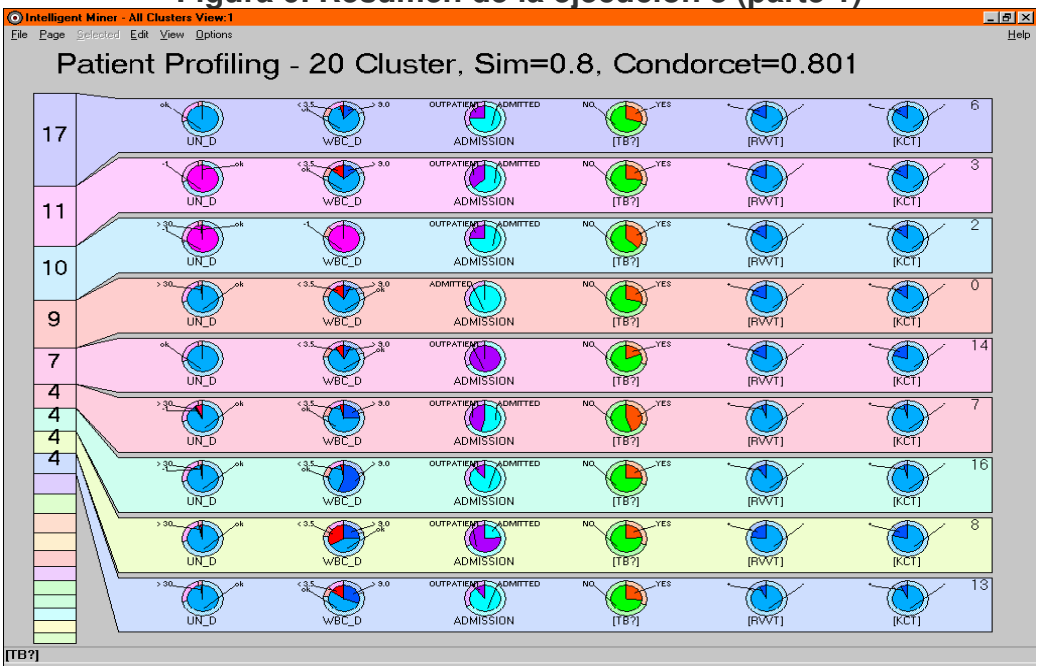
Por lo cual, a pesar de que el condorcet de esta corrida fue la mas alta, es obvio que se pueden obtener mejores resultados para analizar con la ejecución numero 8 que tiene 20 clusters, similitud de 0.8 y un condorcet un poco mas bajo que el anterior. [BAB+01]

Según Baragoin y compañía “En ocasiones es mejor tener un condorcet mas bajo con menos números de clusters y mas grandes; que un condorcet alto con una cantidad grande de pequeños clusters.”

Interpretación de los resultados:

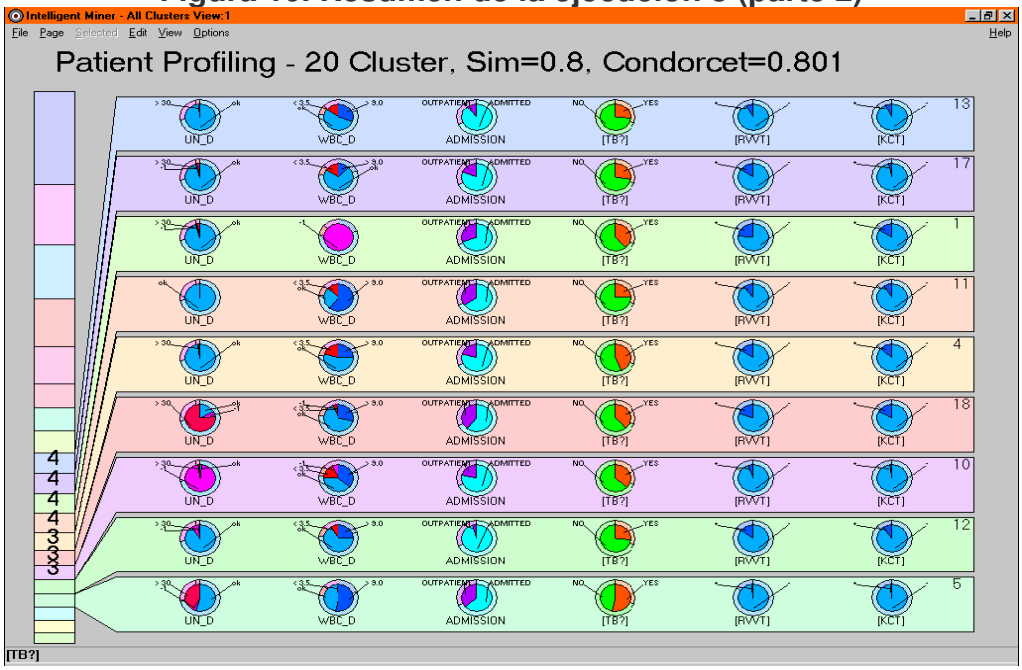
En las siguientes figuras se mostraran los resultados obtenidos en la ejecución número 8 que fue la que se escogió para el análisis.

Figura 9. Resumen de la ejecución 8 (parte 1)



Fuente: [BAB+01]

Figura 10. Resumen de la ejecución 8 (parte 2)



Fuente: [BAB+01]

En la parte superior de las graficas se ve el número de clusters resultantes, el valor de similitud y el valor condorcet que arrojo el modelo. Los clusters están organizados por tamaño, empezando por el mayor; cada cluster esta separado por franjas de distintos colores, dentro de las franjas se encuentra el comportamiento de las variables escogidas para la construcción del modelo, pero dentro del cluster especifico. Estas variables son: UN, WBC, Admisión, Trombosis (TB), RVVT y KCT.

Se observara que el cluster 6 contiene el 17% de los pacientes, seguido por el cluster 3 con el 11% y el cluster 2 con el 10%. Aunque estos son los clusters de mayor tamaño, cabe resaltar que los clusters mas pequeños (1, 4, 5, 10 y 18) son los más interesantes, pues es en estos donde se encuentra relativamente un alto porcentaje de pacientes con trombosis. [BAB+01]

Pero los cluster en los que se concentraron fueron el 4 y el 5 que eran los que tenían mayor cantidad de pacientes con esta enfermedad.

Cluster 4:

Este cluster contiene aproximadamente el 3.29% de los pacientes. En resumen, las variables con una alta relevancia para este cluster muestran las siguientes características:

- Albúmina (ALB_D) era principalmente ‘<3.5’
- Hematocritos (HCT_D) eran principalmente ‘<29’
- Hemoglobina (HGB_D) estuvo siempre ‘<10’
- Los glóbulos rojos (RGB_D) eran principalmente ‘<3.5’
- El numero total de proteínas (TP_D) eran principalmente ’<6.0’
- Acido úrico (UA_D) estuvo principalmente ‘>6.5’

Se hecho un vistazo a las variables suplementarias, se observo que muchas de las pacientes mujeres eran muy jóvenes, se encontraban especialmente entre los 17 y 20 (8% en este cluster, comparado con el 3% en toda la población), entre los 25 y 27 (15% en este cluster, comparado con el 10% en toda la población) y entre los 40 y 42 años (14% en este cluster comparado con el 6% en toda la población). [BAB+01]

El síntoma mas típico que se encontró en este cluster fue el “CNS lupus” (25% en este cluster, comparado con el 8% en toda la población) esta enfermedad es una condición inflamatoria crónica causada por una enfermedad auto inmune. Este tipo de enfermedades ocurre cuando los tejidos finos del cuerpo son atacados por el propio sistema inmunológico. Los pacientes con lupus tienen en su sangre anticuerpos inusuales que atacan el propio tejido del cuerpo. Aproximadamente el 58% (comparado con el 69%) de los síntomas estuvieron ausentes. [BAB+01]

Cluster 5

Este cluster contiene aproximadamente el 2.42% de los pacientes.

En resumen, las variables con una alta relevancia para este cluster muestran las siguientes características:

- Alkalifosfato (ALP_D) era principalmente ‘>300’
- Complemento 4 (C4_D) era principalmente ‘<10’
- Creatina (CRE_D) era principalmente ‘>1.5’
- El genero (SEX) era principalmente masculino.
- Nitrógeno ureico (UN_D) era principalmente ‘>30’

Se hecho un vistazo a las variables suplementarias, se observo que muchos de los pacientes hombres pertenecían a la tercera edad. Especialmente entre 45 y 47 años (33% comparado con el 4%) [BAB+01]

Algunos de los síntomas no fueron registrados (46% comparado con el 69%), el principal síntoma sin embargo fue identificado como el síndrome de Budd-Chiari (46% comparado con el 2%). [BAB+01]

Este síndrome incluye una variedad de condiciones con la obstrucción de la salida de la vena hepática, en el nivel de las venas hepáticas grandes o del segmento sobre hepático de la vena cava inferior. Dependiendo de la causa o de las manifestaciones pato fisiológicas se puede clasificar como primario o secundario. [BAB+01]

2.1.4. Despliegue y utilización de los resultados de minería de datos:

La idea específica de esta organización fue direccionar con los avances en la tecnología de minería de datos, los modelos de clustering a nuevos grupos de pacientes. [BAB+01]

Calit utilizo este modelo para sacar perfiles de pacientes, para tratar de identificar a personas que tuvieran el riesgo de un deterioro de su salud. Una vez que se identificaban estas personas, los doctores trataban de persuadir a los pacientes para establecer un plan de tratamiento pro-activo con el fin de prevenir el deterioro de la salud de estos. [BAB+01]

Como sacar perfiles de pacientes no es solo un método para explicar las características de un grupo de pacientes respecto a una enfermedad especifica,

se utilizo este mismo método para hacerlo con otro tipo de enfermedades y pacientes. [BAB+01]

Los siguientes fueron los casos donde la organización Calit healt services aplico este método para sacar perfiles de los pacientes

Caso 1: tuberculosis[16]

La tuberculosis es una enfermedad causada por una bacteria llamada mycobacrium tuberculosis o también tubérculo bacilos. Es propagada de persona a persona a través del aire por la inhalación de partículas que contienen la bacteria. Estas partículas son producidas cuando una persona infectada con la bacteria exhala, como cuando esta tosiendo, estornudando, riéndose, hablando o cantando.
Las partículas infecciosas pueden permanecer suspendidas en el aire y ser inhaladas por alguien que comparte el mismo aire. La tuberculosis es transmitida en áreas de cerradas donde la ventilación es poca. El riesgo de infección aumenta cuando se comparte el mismo aire con una persona que no esta siendo tratada.

Si el sistema inmunológico del cuerpo no puede contener la bacteria, esta continuara en la producción de más y más bacteria. La infección se da en la parte superior de los pulmones y pueden pasar varios meses para que los síntomas se empiecen a mostrar. Estos síntomas son principalmente cansancio, debilidad, perdida de peso, fiebre y sudor. Si la infección en los pulmones empeora mas síntomas se pueden presentar como tos, dolor en el pecho, y fallas de respiración.

Al respecto, hubo ciertos aspectos que reflejaron la necesidad de hacer perfiles:

[16] Ver [BAB+01] para complementar la información sobre este caso de aplicación.

- La descripción de los síntomas hace referencia a características que son conocidas por los médicos. Sin embargo puede ser el caso de que hayan síntomas aun desconocidos. Esto puede depender por ejemplo del país, las diferentes condiciones o las diferentes situaciones higiénicas.

- La bacteria de la tuberculosis puede mutar en el tiempo y en lugar, esto dificulta tener siempre la ultima bacteria justo a tiempo.

- Algunas veces no es fácil probar si un paciente tiene tuberculosis o no; aunque esta puede se diagnosticada en diferentes formas; por ejemplo mediante rayos X que revelen la evidencia de tuberculosis, neumonía o una cicatriz que muestre una tuberculosis inactiva. Es importante tener el diagnostico lo mas temprano posible. Entre mas avance el estado de la enfermedad se hace más difícil el tratamiento a seguir mediante las pruebas estándar.

Caso 2: Narcosis[17]

Otro escenario en que aplico este modelo fue con la narcosis que toma lugar antes de que un paciente sea operado. Normalmente los doctores deciden que clase de narcosis será usada para un paciente que se va a operar.

La narcosis no es tan inofensiva como parece. Si el paciente muestra alguna alergia a los medicamentos esto puede inducir a diferentes formas de dolor; por ejemplo con los pulmones, el corazón o el cerebro. En el peor de los casos el paciente puede caer en un coma.

Hacer perfiles de pacientes ayudo a identificar pacientes que tenían, por ejemplo, alergia a algún medicamento en especial. Basados en este modelo los doctores

[17] Ver [BAB+01] para complementar la información sobre este caso de aplicación.

pueden usar esta información como una fuente adicional para asegurar la salud del paciente.

Caso 3: sistema para el manejo de un hospital[18]

Para el manejo del hospital, es importante conocer cuantos pacientes son admitidos. Si la capacidad del hospital esta completa, entonces solo el servicio medico para casos de emergencia es factible, y en un caso extremo, el paciente no puede permanecer en el hospital para cuidados médicos, por lo cual lo tiene que irse, ya que no existen camas o habitaciones disponibles para una posible operación.

Hacer perfiles de pacientes ayudo a sobreponer problemas de capacidad, por ejemplo:

- Se supone, que se tienen los registros de los pacientes que tiene una enfermedad parecida o similar, pero que se quedan diferentes días en el hospital. Las razones para esto es por algo tan simple como los diferentes estados de salud de los pacientes, si fuman o no, etc. Estas razones son conocidas y frecuentemente dan una indicación de cómo pueden manejar estos casos respecto a los servicios médicos.

  Sin embargo, puede haber otras razones para que la permanencia de los pacientes del hospital sea diferente.

  Por ejemplo, pueden haber pacientes que prefieren habitaciones con ventanas con vistas hacia el sur o el oeste, y por consiguiente tener una mejoría en general. Mientras otros pacientes prefieren estar en ventanas

[18] Ver [BAB+01] para complementar la información sobre este caso de aplicación

con vistas hacia el norte, o el este. Claro que esto a simple vista no parece ser algo relevante.

- Como se tienen registros sobre los pacientes, que fueron enviados al hospital. Entonces sacar perfiles de los pacientes puede ayudar, como un instrumento adicional, para estimar que personas probablemente necesitan quedarse en el hospital y cuales no.

  Mediante las reglas generadas por los perfiles de pacientes que se tienen, el sistema puede ayudar a predecir que grupos de pacientes tienen que permanecer en el hospital.

Esta aplicación básicamente, le permitió a los hospitales y clínicas de esta institución contar siempre con habitaciones y quirófanos disponibles, para cuando fuera necesario, por lo cual no había necesidad de enviar a los pacientes a otros centros hospitalarios. [BAB+01]

2.2. CLASIFICACIÓN DE CLIENTES PARA UNA CAMPAÑA DE MERCADEO A TRAVÉS DE CORREO DIRECTO

Este caso trata sobre la necesidad de una cadena de almacenes de ropa de predecir la respuesta que tendría una campaña de mercadeo utilizando correo directo.

En este caso la tarea de minería de datos que se trato fue la de clasificación, y se utilizo la herramienta Clementine de SPSS.

2.2.1. Comprensión del negocio o del tema a tratar:

La cadena de almacenes de ropa RK Clothes, conduce anualmente una campaña de mercadeo usando correo directo, con el fin de obtener más ganancias y establecer una relación más directa con sus clientes.

Como ya se había mencionado anteriormente, la cadena esta interesada en incrementar sus ventas. Para ello quieren predecir que clientes son los que mas probablemente responderán a una campaña de promoción de los diferentes productos ofrecidos.

El desafió es clasificar que clientes responderán a una campaña de mercadeo mediante correo directo, basados en la información colectada sobre esos clientes, y utilizando técnicas de minería de datos

Construcción de la tabla de costo/beneficio

Según Larose los modelos de clasificación son muchas veces evaluados, de acuerdo a porcentajes de precisión, a tasas de error, porcentajes de falsos positivos y falsos negativos. Estas medidas pueden ser aplicadas a cualquier problema de clasificación, sin embargo para un problema en particular, es probable que estas medidas no seleccionen el modelo mas optimo. La razón es que cada problema de clasificación acarrea con un único conjunto de costos y beneficios, que son el resultado particular de un conjunto de circunstancias únicas de cada negocio o de cada problema de investigación.

Para este autor, en problemas de negocios, como este que se esta tratando, los gerentes pueden requerir que las comparaciones del modelo sean hechas en términos de análisis de costo beneficio, además es muy útil cuando se esta realizando una clasificación. Esto es hecho para suministrar una comparación del

modelo en términos del beneficio o la pérdida anticipada, asociando un costo o beneficio a cada una de las 4 posibles combinaciones de clasificaciones correctas e incorrectas.

Considérense cada uno de los 4 posibles resultados definidos por Larose en la construcción del caso y el costo razonable al que cada uno conlleva.

1. **verdadero negativo (TN[19]):** el modelo predice que estos clientes no responderían a la campaña de promoción de mercadeo mediante correo directo, así que por esto la tarjeta postal no será enviada por correo a estos clientes. En realidad estos clientes no responderán a la promoción. Por consiguiente se tomo la decisión correcta. No se incurre en costos pues no se manda ninguna tarjeta postal; no se hacen ventas y tampoco se pierden ventas. [Lar06]

2. **Verdadero positivo (TP):** el modelo predice que estos clientes responderían a la campaña de promoción de mercadeo mediante correo directo, así que la promoción fue enviada a estos clientes. En realidad estos clientes podrían responder a la promoción. Por consiguiente se tomo la decisión correcta. El costo del correo directo, con materiales, estampillas y manejo es de $2 por unidad de promoción enviada por correo. Ahora la pregunta que surge es cuanto dinero seria razonable esperado que gaste un cliente y cuanto de esta cantidad se puede considerar como ganancia. Lo que se hizo fue calcular las estadísticas asociadas con el promedio de gastos que se espera de los 28799 clientes. La media que fue de $113.59 se utilizo para indicar la cantidad que un cliente gastara después de recibir la promoción. Se asumió que el 25% de los $113.59 ($28.40) representaría el beneficio. Luego el beneficio asociado con el cliente, es el beneficio

[19] Abreviatura tomada de sus siglas en ingles.

esperado por la vista menos el costo asociado con el envió del correo, o sea $2, lo que daría $26.40. [Lar06]

Tabla 7. Estadísticas asociadas con el monto promedio de gasto por visita para todos los clientes.

Población	28799
Media	113.588
Mínimo	0.490
Máximo	11,919.880
Desviación estándar	86.981
Mediana	92.000

Fuente: [Lar06]

3. **Falso negativo (FN):** un faso negativo indica que se fallo el contacto con un cliente que respondería positivamente a la promoción. Este error es muy caro, por lo tanto se debe tratar de minimizar al máximo cometerlo. El costo en el que se incurre se da cuando no se envía la promoción a una persona que respondería afirmativamente y que gastaría el promedio establecido en la tabla 1 de $113.59, por lo cual se perdería el beneficio de $28.40 asociado con ese cliente. [Lar06]

4. **Falso positivo (FP):** un falso positivo significa que se contacte un cliente que no respondería a la promoción. Lo cual no es muy costoso. El costo asociado de contactar un cliente que no respondería es de $2 por el envió del correo. [Lar06]

Con estos valores se puede construir la tabla de costo/beneficio para esta promoción para la cadena RK Clothes. Esta tabla puede ayudar a decidir cual modelo será seleccionado como el más optimo.

Tabla 8.Costo/Beneficio para el problema de promoción de la cadena RK Clothes

Resultado	Clasificación	Respuesta actual	Costo	Razón
Verdadero negativo	No respondió	No respondió	$0	No hubo contacto, no hay perdida de ganancia
Verdadero positivo	Respondió	Respondió	-$26.4	Ganancia estimada menos costos de envió de correo
Falso negativo	No respondió	Respondió	$28.4	Perdida de beneficio.
Falso positivo	Respondió	No respondió	$2	Materiales asociados al envió del correo.

Fuente: [Lar06]

2.2.2. Comprensión y preparación de los datos a utilizarse:

El conjunto de datos de esta cadena de tiendas contiene información de 28799 clientes en los siguientes campos:

- ID del cliente: es única, es la identificación del cliente encriptada.
- Código postal
- Numero de vistas para hacer compras.
- Cantidad promedio de gasto por visita
- Cantidad gastada en cada una de las 4 franquicias.
- Cantidad gastada en el mes pasado, los pasados tres meses y los pasados 6 meses.
- Cantidad gastada el mismo periodo el año anterior.
- Porcentaje del margen bruto.
- Numero de promociones de mercadeo archivadas.
- Numero de días desde que el cliente ha sido registrado.
- Numero de días entre compras.
- Porcentaje de descuento a las compras del cliente.
- Numero de diferentes clases de productos comprados.

- Numero de cupones usados por el cliente.
- Numero total de productos individuales comprados por el cliente.
- Numero de tiendas donde el cliente compra.
- Numero de promociones enviadas por correo el año pasado.
- Numero de promociones a las que respondió el cliente el año pasado.
- Porcentaje de promociones respondidas el año pasado.
- Uniformidad de producto. (Puntuación baja = diversos patrones de gastos)
- Tiempo medio entre visitas.
- Cluster del estilo de vida de los clientes.
- Porcentaje de devoluciones.
- Bandera: tarjeta de crédito del cliente.
- Bandera: numero de teléfono valido archivado.
- Bandera: compra por la Web.
- 15 variables que suministran el porcentaje de gastos de los clientes en clases especificas de ropa, incluye suéteres, sostenes tejidos, vestidos tejidos, blusas, chaquetas, pantalones formales, pantalones casuales, camisas, vestidos, trajes, joyas, artículos de moda, abrigos, la línea de colección, y también una variable que muestra la marca escogida, la cual esta encriptada.
- Variable objetivo: respuesta a la promoción.

Estos datos se basaron en una campaña similar, conducida el año pasado. Esta información se uso para desarrollar los modelos de clasificación para la campaña de mercadeo de este año. Para la fase de comprensión de los datos el grupo del proyecto utilizo métodos estadísticos gráficos y descriptivos para conocer más los datos. [Lar06]

Una de las variables, cluster del estilo de vida de los clientes, contiene la segmentación del mercado de estos clientes. Son 50 segmentos nombrado del 1

al 50. Los 6 clusters que representan los estilos de vida más comunes en el conjunto de datos son: [Lar06]

1. **Cluster 10:** hogar, dulce hogar. Familias con un ingreso y educación medio-alto, administradores/profesionales, técnicos/ventas.
2. **Cluster 1:** corteza superior. Familias metropolitanas, familias con altos ingresos y educación, propietarios de casas, administradores/ profesionales.
3. **Cluster 4:** éxito en la mitad de la vida. Familias con una altos niveles de educación y altos ingresos administradores/profesionales, técnicos/ventas.
4. **Cluster 16:** familias rurales. Familias grandes que viven en áreas rurales, tienen niveles de educación media, ingresos medios.
5. **Cluster 8:** hombres de acción y agitadores. Solteros, parejas, estudiantes y recién graduados, alta educación e ingresos administradores/profesionales, técnicos/ventas.
6. **Cluster 15:** gran comienzo. Jóvenes, solteros y parejas. Educación media-alta, ingresos medios, algunos son arrendatarios administradores/profesionales, técnicos/ventas.

A simple vista se nota que la cadena de tiendas atrae a personas acaudaladas y con altos niveles de educación, se ve que el cluster 1 es el de las personas más ricas y es el segundo más predominante entre los clientes de esta cadena. [Lar06]

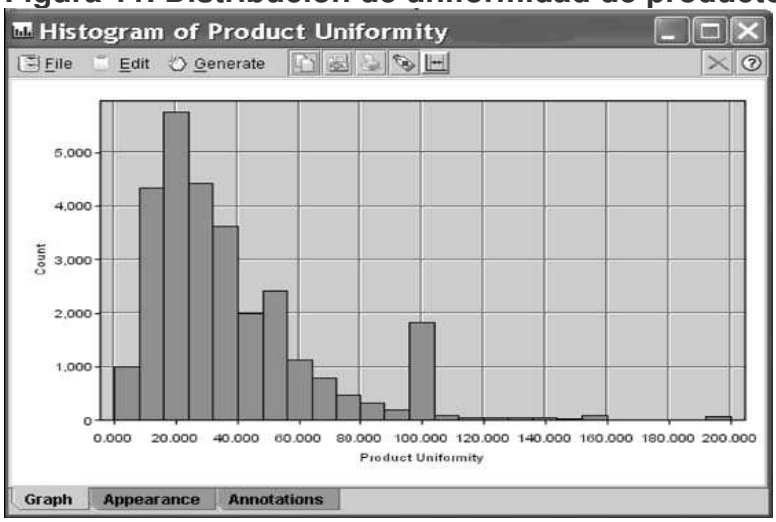
El caso de la variable ID del clientes, es único para cada cliente, pero no arroja información importante para predecir que clientes respondan probablemente a la campaña de mercadeo mediante correo directo, por lo cual es omitida para futuros análisis. [Lar06]

Transformaciones de las variables para conseguir normalidad y simetría

Para que los métodos y modelos de minería de datos trabajen mejor, se tiene que contar con variables que estén normalmente distribuidas o que al menos sean simétricas. Por lo cual se aplican ciertas transformaciones a las variables numéricas que lo requieren para tratar de cumplir con estas condiciones. Los métodos utilizados para las transformaciones según el libro Data Mining Methods y Models son la transformación del logaritmo natural, la transformación de la raíz cuadrada y la transformación BOX-COX. Como las variables que se están trabajando solo contienen valores positivos se aplico la transformación del logaritmo natural. Sin embargo para las variables que contenían el valor cero, se aplico la transformación de la raíz cuadrada, pues se sabe que al aplicar la transformación de LN(x) cuando x = 0, es indefinido.

Muchos de los campos numéricos están sesgados, por lo que no es muy recomendable utilizar las variables sin antes aplicar algunas de estas transformaciones. Por ejemplo la siguiente figura muestra la distribución de la variable uniformidad de producto, la cual toman grandes valores para clientes que compran solo unas pocas clases de ropa (Ej.: blusas, pantalones) y valores pequeños para clientes que comprar diferentes clases de ropas. Pero esta figura la cual esta sesgada a la derecha muestra que muchos clientes tienen un valor bajo de uniformidad de producto, mientras unos pocos tienen grandes valores; estos clientes con grandes valores para esta variable tienden a comprar uno o dos diferentes clases de ropas. [Lar06]

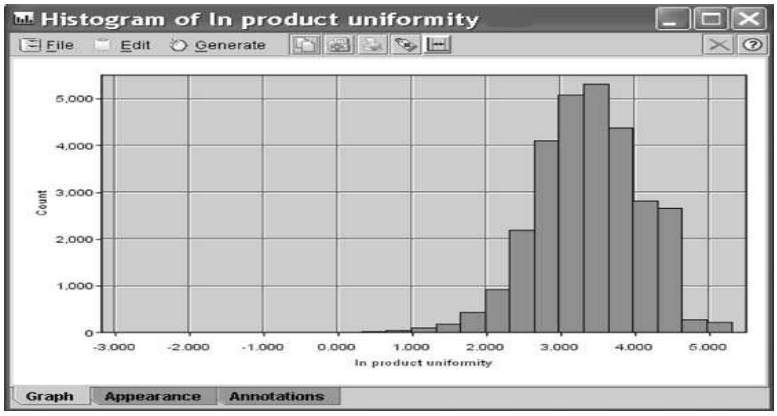
Figura 11. Distribución de uniformidad de producto



Fuente: [Lar06]

La siguiente figura muestra la distribución de la variable uniformidad de producto después de la transformación del logaritmo natural. Aunque no se logra obtener una normalidad perfecta, si se logra tener una distribución menos sesgada, permitiendo la aplicación de los modelos de minería de datos. [Lar06]

Figura 12. Distribución de la variable uniformidad de producto después de la transformación.



Fuente: [Lar06]

Para las variables que indican un porcentaje, también aplicaron transformaciones, pero esta vez se aplico la transformación de la raíz cuadrada. [Lar06]

Algunas variables, toman un carácter dicotómico, lo que incentivo a los investigadores del proyecto a obtener una variable bandera. Este fue el caso de la variable compradores de blusas. [Lar06]

Esta variable bandera indico por ejemplo que el 58% de los clientes han comprado una blusa en algún instante de tiempo. [Lar06]

Las variables banderas fueron construidas para las otras 14 variables de ropa que indican porcentaje. [Lar06]

Estandarización y variables bandera

Para Larose "Cuando hay grandes diferencias en la variabilidad, entre las variables numéricas, el analista de datos necesita aplicar una estandarización". Las transformación que ya se aplicaron ayudaron en parte a reducir las diferencia de variabilidad entre las variables, pero siguen existiendo diferencias considerables. Para ilustrar esto, se ve que la desviación estándar para la variable cantidad gastada en los pasados 6 meses (sqrt[20]) es 10.02, mientras que la desviación estándar para la variable, número de cupones usados (sqrt) es 0.735. [Lar06]. Para evitar esa gran variabilidad entre estas variables se hizo necesario normalizar o estandarizar. En este caso se escogió estandarizar los campos numéricos, de tal forma que todos tengan una media de 0 y una desviación estándar de 1. Para cada variable, esto fue hecho restando la media de la variable y luego se dividió por la desviación estándar para obtener lo que se llama Z calculada. Las variables resultantes son nombradas con una Z (Ej. Z (sqrt) # de cupones usados). [Lar06]

[20] Indica que la variable fue transformada mediante la transformación de la raíz cuadrada.

Las variables banderas fueron construidas con la necesidad de mostrar ciertas situaciones en particular. Un ejemplo es que se logro captar mediante estadísticas que la mayoría de los clientes no han gastado dinero en ninguna de las tiendas el mes pasado. Las siguientes variables también tienen una variable bandera asociada: [Lar06]

- Gasto en la tienda AM (una de las 4 franquicias), para indicar que clientes gastan dinero en esta tienda en particular.
- Gasto en la tienda PS
- Gasto en la tienda CC
- Gasto en la tienda AX
- Gastos en los últimos 3 meses.
- Gastos en los últimos 6 meses.
- Gastos en el mismo periodo del año pasado.
- Devoluciones, para indicar que clientes han devuelto mercancía
- Porcentaje de respuesta. Para indicar que clientes han respondido a una campaña de marketing anteriormente.
- Descuento, para indicar que clientes ha comprado mercancía que ha sido rebajada.

Obteniendo nuevas variables

"La fase de preparación de los datos le ofrece a la minería de datos, la oportunidad de clarificar relaciones y obtener nuevas variables que pueden ser muy útiles para el análisis". Por ejemplo considere las siguientes tres variables: cantidad gastada en el mes pasado, cantidad gastada en los tres meses pasados y cantidad gastada en los seis meses pasados. Claramente se ve que la primera variable esta contenida en las otras dos, por lo cual esta siendo contada 3 veces. Como se quiere evitar esto, el paso a seguir fue obtener 2 nuevas variables que

ayudaran a resolver este problema, las cuales se verán en la siguiente tabla. [Lar06]

Tabla 9.Nuevas variables obtenidas de gastos.

Variable Obtenida	Formula
Cantidad gastada en los meses previos 2 y 3.	Cantidad gastada en los últimos 3 meses – cantidad gastada en el ultimo mes
Cantidad gastada en los meses previos 4, 5 y 6	Cantidad gastada en los últimos 6 meses – cantidad gastada en los últimos 3 meses.

Fuente: [Lar06]

Se omitieron las variables: cantidad gastada en los tres meses pasados y cantidad gastada en los seis meses pasados y se trabajara con: gastada en el mes pasado, Cantidad gastada en los meses previos 2 y 3, y Cantidad gastada en los meses previos 4, 5 y 6. [Lar06]

Explorando las relaciones entre las variables predictoras y la respuesta

El siguiente paso fue averiguar variable por variable la asociación entre las variables predictoras y la variable objetivo (respuesta a la campaña de mercadeo). Lo ideal era hacerlo con cada variable, pero como esta tarea es un poco tediosa, porque el conjunto de datos era muy grande. Sin embargo si se utilizaron ciertos caminos para examinar aquellas variables mas útiles para la predicción. [Lar06]

La herramienta que muestra Larose para examinar cuales variables eran más útiles fue la de examinar los coeficientes de correlación para cada predictor con la variable respuesta y se seleccionaron para los futuros análisis aquellas variables que tienen la mayor correlación absoluta, o sea, $|r| \geq 0.30$.

La siguiente tabla muestra la lista de variables con las correlaciones absolutas más altas con la variable objetivo, respuesta.

Tabla 10. Variables con la correlación absoluta más grande con la variable objetivo (respuesta)

Variable	Coeficiente de correlación	Relación
Z LN Tiempo promedio entre visitas	-0.431	Negativa
Z LN visitas de compras	0.399	Positiva
Z LN # productos individuales comprados	0.368	Positiva
Z LN total neto de ventas	0.336	Positiva
Z LN promociones respondidas el año pasado	0.333	Positiva
Z LN # de diferentes clases de productos.	0.329	Positiva
Z LN # de cupones usados	0.322	Positivo
Z LN días entre compras	-0.321	Negativa

Fuente: [Lar06]

La relación entre la variable Z LN Tiempo promedio entre visitas y la variable respuesta; indica que el porcentaje de respuesta a la campaña de promoción decrece, en la medida en que el tiempo promedio entre visitas se incrementa. [Lar06]

Todas las relaciones de las variables Z LN visitas de compras, Z LN # productos individuales comprados, Z LN total neto de ventas y Z LN # de diferentes clases de productos, con la variable respuesta, muestra que a medida que estas variables se incrementan, la variable respuesta también se incrementa. Esto no sorprende, puesto que los clientes que hacen compras en los almacenes a menudo, compran diferentes productos, gastan grandes cantidades de dinero, compran diferentes clases de productos, pueden estar interesados en responder a la campaña de mercadeo. [Lar06]

Cuando la variable numero de promociones respondidas anteriormente se incrementaba, también se observaba un incremento en la variable respuesta. Lo mismo ocurrió en el caso de la variable, numero de cupones usados. Pero cuando

había un incremento en la variable numero de días entre compras, se percibía un decremento en la variable respuesta. [Lar06]

Se observa que el comportamiento de las variables respecto a la variable respuesta se ajusta a lo que se ve en la tabla de correlaciones mostrada anteriormente.

Es necesario retomar el comportamiento de la variable porcentaje gastado en blusas, pero esta vez relacionándola con la variable respuesta, cuando la primera aumentaba, el porcentaje de respuesta decrecía. Este comportamiento no es solo para las blusas, prevalece entre todas las variables de porcentajes de ropa. Esto parece indicar es que los clientes que se concentran en un tipo particular de prenda, comprando uno o dos tipos de prendas, tienden a tener una porcentaje de respuesta mas bajo. [Lar06]

En el caso de la variable uniformidad de producto, el más alto porcentaje de respuesta a la campaña, se da con los clientes que tienen una menor uniformidad de producto, esto es para los clientes que tienen hábitos de compras más diversos, en otras palabras, aquellos que compran diferentes tipos de prendas. [Lar06]

En la examinación que se hizo de la relación entre la respuesta y las variables bandera, se obtuvo una mayor asociación respecto a la campaña con las variables: comprador por Web, tenedor de tarjeta de crédito, cantidad gastada el mes pasado y cantidad gastada en el mismo periodo del año anterior. Las proporciones[21] de los clientes que responderían a la campaña, condicionadas por el valor de la bandera fueron los siguientes: [Lar06]

[21] Esta proporciones se encuentran en el libro Data Mining Methods and Models

- Los tenedores de tarjetas de crédito tienen una probabilidad de cerca de tres veces más que los que no tienen tarjeta de crédito de responder a la promoción. (28.066% vs 9.376%).
- Los compradores por Web tienen una probabilidad también cercana a 3 veces más, de que los que no utilizan este servicio, de responder a la campaña. (44.852% vs 15.247%)
- Los clientes que hicieron una compra el mes pasado tienen una probabilidad 3 veces mas alta de que respondan a la campaña (33.642% versus 11.981%).
- Aquellos que hicieron compras el año pasado en el mismo periodo de tiempo, tienen el doble de probabilidad de responder a la campaña, de aquellos que no lo hicieron (27.312% vs 13.141%).

En cuanto a la variable de clusters de estilos de vida no parecen haber diferencias sustanciales en cuanto a la respuesta a la campaña, entre los clusters. [Lar06]

Investigando la estructura de correlación entre los predictores

En esta etapa lo que se trato de investigar fue la correlación entre las variables predictoras, pues esto puede ser peligroso para el desarrollo del modelo. Lo que se hizo fue sacar los coeficientes de correlación entre parejas de variables y mirar cuales eran las correlaciones mas fuertes. La siguiente tabla muestra el conjunto de parejas con la correlación en valor absoluto más grandes, entre las variables predictoras numéricas. [Lar06]

Tabla 11. Valores de correlación más altas entre variables predictoras.

Predictor	Predictor	Correlación
Z LN visitas de compras	Z LN # de diferentes clases de productos	0.804
Z LN visitas de compras	Z LN # de productos individuales comprados	0.860
Z LN promociones archivadas	Z LN promociones enviadas el	0.890

	años pasado	
Z LN total neto en ventas	Z LN # de diferentes clases de productos	0.859
Z LN total neto en ventas	Z LN # de productos individuales comprados	0.907
Z LN días entre compras	Z LN tiempo promedio entre visitas.	0.847
Z LN # de diferentes clases de productos.	Z LN # de productos individuales comprados	0.930

Fuente: [Lar06]

Para las variables categóricas se utilizo otro método para examinar la relación entre estas variables con la variable respuesta. Este método es llamado tabulación cruzada. Como se tenía interés por saber la relación entre respuesta a la promoción y dos tipos de clientes: aquellos que han comprado suéteres y aquellos que hicieron una compra el pasado mes. Se construyo una matriz de tabulación cruzada, donde las celdas representan el valor medio de la variable objetivo (respuestas). Puesto que el objetivo representa una variable dicotomica, las medias representan proporciones [Lar06]. Esta matriz se muestra a continuación.

Figura 13. Tabulación cruzada para el gasto del pasado mes vs compra de suéteres, donde el valor de las celdas representa los porcentajes de respuesta a la promoción.

Fuente: [Lar06]

Así en la tabulación cruzada se puede ver que los clientes que ni han comprado suéteres, ni que hicieron una compra el semestre pasado, solamente tienen una probabilidad de responder a la campaña de mercadeo mediante correo directo.

Por otro parte los clientes que cumplen con las condiciones de las dos variables tienen una probabilidad de responder positivamente a la promoción. Por ultimo si un cliente cumple una de las 2 condiciones de las variables, el gasto hecho el mes pasado es un poco más significativo a responder a la promoción que los que han comprado una blusa. (0.194 versus 0.146 de probabilidad, respectivamente). [Lar06]

Para mas información sobre las fases de entendimiento y preparación de los datos se puede dirigir al libro Data Mining Methods and Models que es donde se encuentra este caso de estudio.

2.2.3. Modelado y evaluación:

Aunque la exploración inicial de los datos dio elementos para comprender mejor los datos, ya era necesario pasar a la etapa de modelado y aplicar un conjunto de algoritmos de minería de datos, que mostrara resultados más concretos. La estrategia[22] que siguieron las personas que estaban desarrollando este proyecto para el modelado fue la siguiente:

- Dividir el conjunto de datos en un conjunto de datos para entrenamiento y en otro para pruebas.
- Suministrar una lista de las entradas de todos los modelos.
- Aplicar el análisis de componente principal[23], para saber si hay multicolinealidad.
- Aplicar análisis de cluster y perfilar concretamente los clusters resultantes.
- Equilibrar el conjunto de datos de entrenamiento para suministrar a los algoritmos con un número similar de registros de aquellos que responden y no responden a la campaña.

[22] Esta estrategia fue definida en [Lar06]

[23] El Análisis del Componente Principal (ACP) (en inglés, PCA) es una técnica que usada para simplificar un conjunto de datos. Ver capitulo 1 de [Lar06].

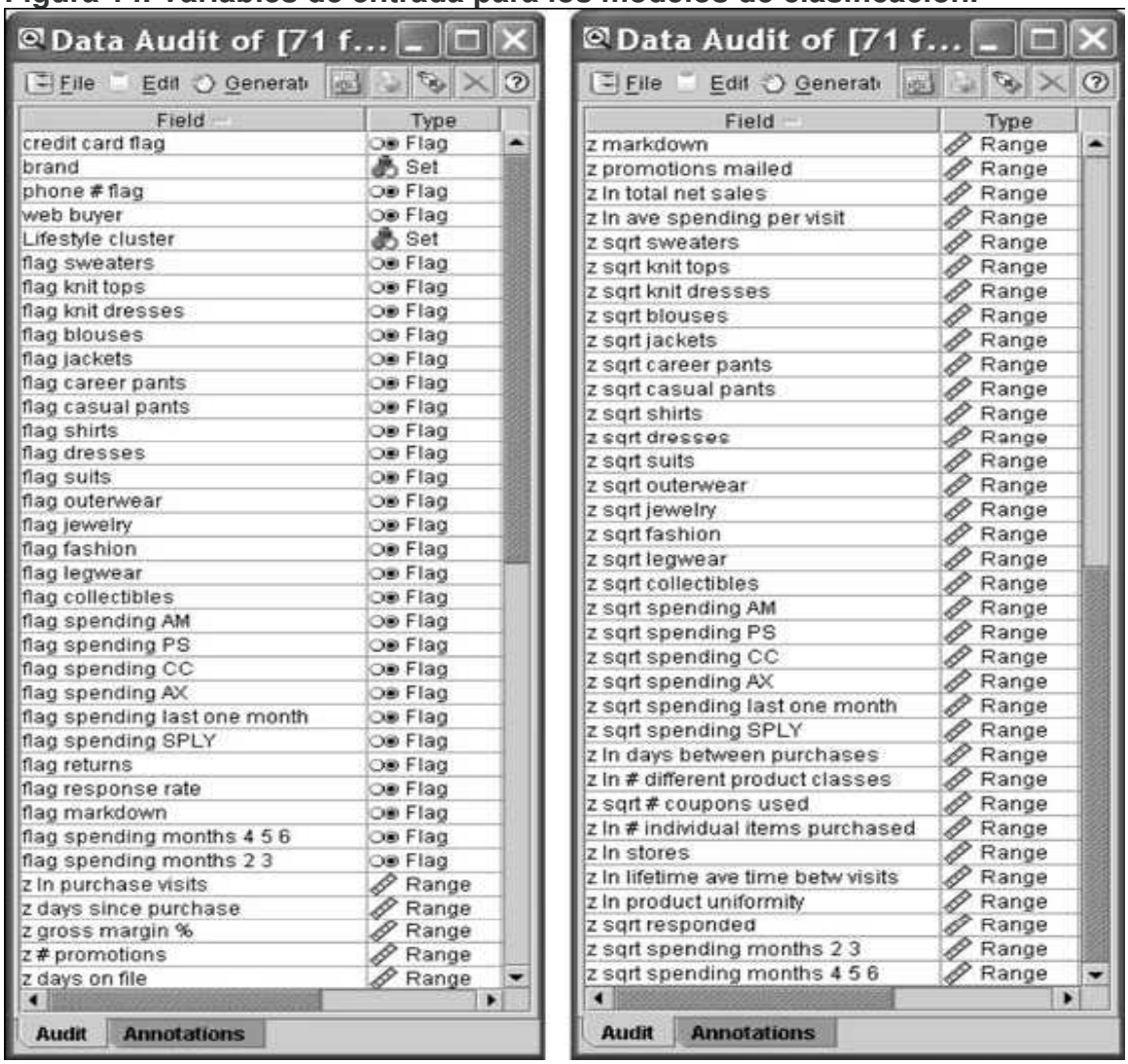
- Establecer una línea de base de rendimiento del modelo en términos del beneficio esperado por cliente contactado, para calibrar el rendimiento de los modelos candidatos.
- Aplicar los siguientes algoritmos de clasificación al conjunto de datos de entrenamiento:
    - Árboles de regresión y de clasificación (CARTs)
    - Algoritmo C5.0 de árboles de decisión.
    - Redes neuronales
    - Regresión logística.
- Evaluar cada uno de estos modelos usando el conjunto de datos de pruebas.
- Aplicar costos de error en la clasificación[24] costos en línea con la tabla de costo/benéfico definida en la fase de entendimiento del negocio.
- Aplicar sobre balance como un substituto para los costos de error en la clasificación, y encontrar el índice más eficaz de sobre-balance.
- Combinar las predicciones de los 4 modelos de clasificación usando votación de modelos.
- Comparar el rendimiento de los modelos que usan análisis de componente principal, con los modelos que no lo usan y plantear el role de cada tipo de modelo.

Lo primero que se hizo, como se mencionaba anteriormente fue dividir el conjunto de datos en un conjunto de datos para entrenamiento y en otro para pruebas; para hacer esto se utilizo la herramienta Clementine, con el cual se decidió que el 75% de los datos se utilizaran en entrenamiento y el 25% restante en pruebas. [Lar06]

El siguiente paso fue escoger las variables de entradas para los modelos de clasificación. [Lar06]

[24] En ingles se conoce como misclassification costs.

Figura 14. Variables de entrada para los modelos de clasificación.



Fuente: [Lar06]

En estas entradas están incluidas las variables derivadas, variables transformadas y las variables que no necesitaron ser transformadas. Todas las variables numéricas han sido transformadas y estandarizadas, muchas de las variables banderas fueron obtenidas o derivadas de otras, y solo las variables categóricas: marca y cluster de estilo de vida permanecen. Cuando el análisis de componentes y el de cluster sean desarrollados se indicaran para que modelos estos deben ser utilizados para las entradas. [Lar06]

Análisis de componente principal

Retomándose la tabla que muestra la pareja de correlaciones más fuertes encontradas entre los predictores. Hay que tener en cuenta que una fuerte correlación entre los predictores lleva a la multicolinealidad, lo que no es algo muy recomendable, porque puede causar inestabilidad y llevar a resultados incoherentes. Dependiendo del objetivo primario del objetivo del negocio o del problema de investigación, los investigadores pueden decidir sustituir los componentes principales por una colección particular de predictores correlacionados. [Lar06]

- Si el objetivo primario de la investigación o del problema de negocio concierne solamente a la estimación, predicción o clasificación de la variable objetivo, sin ningún interés en cualquiera de las características predictoras, la sustitución de los componentes principales por la colección de los predictores correlacionados no es estrictamente requerida.[Lar06]
- Sin embargo si el objetivo principal del análisis es evaluar o interpretar el efecto de los predictores individuales en la respuesta o para desarrollar un perfil de los que pueden responder probablemente, basados en las características de los predictores, la sustitución de los componentes principales por la colección de los predictores correlacionados es fuertemente recomendada. [Lar06]

Por consiguiente parte de la estrategia a seguir fue reportar dos tipos de los mejores modelos; uno (que no contiene componentes principales) para usar solamente en la predicción del objetivo, y otro (que contiene componentes principales) para todos los otros propósitos. [Lar06]

Tabla 12. Componentes de carga de los dos principales componentes extraídos del conjunto de datos de entrenamiento.

	Componente	
	1	**2**
Z LN # de productos individuales comprados	0.915	
Z LN # de diferentes clases de productos.	0.887	
Z LN visitas de compras	0.858	
Z LN tiempo promedio entre visitas	-0.858	
Z LN total de ventas netas	0.833	
Z promociones enviadas por correo		0.944
Z # de promociones		0.932

Fuente: [Lar06]

Componente principal 1: hábitos de compra. Este componente consiste de los más importantes hábitos de compra de los clientes en general. Aquí están incluidos el total del número de productos comprados, el número de diferentes tipos de prendas compradas, el número de diferentes veces que los clientes vienen a la tienda a comprar algo, y el monto total de dinero gastado por el cliente. Todas estas variables esta relacionadas positivamente con las otras. Por el contrario la variable promedio de tiempo entre visitas también esta incluido en este componente, pero esta negativamente correlacionada con las otras variables, puesto que tiempos más largos entre visitas podría estar presumiblemente correlacionado con otros hábitos de compras. Se podría esperar que este componente pudiera ser un fuerte indicativo de la respuesta a la campaña de mercadeo mediante correo directo. [Lar06]

Componente principal 2: contactos de promoción. Este componente consiste solamente de dos variables, el número de promociones enviadas por correo el año pasado y el total del número de promociones de marketing archivadas. Nótese que en este componente no hay información sobre la respuesta de los clientes a estos contactos de promoción. Así, es poco claro si este componente estará asociado con la respuesta a la promoción. [Lar06]

El componente principal extraído del conjunto de datos de entrenamiento se valido por comparación con el componente principal extraído del conjunto de datos de pruebas. La tabla a continuación muestra los componentes de carga para el componente principal extraído de las 7 variables correlacionadas mostradas anteriormente, pero esta vez usando el conjunto de datos de prueba. Una comparación entre las tablas 12 y 13 muestra que los componentes de carga, aunque no son idénticos, aun así, son suficientemente similares para confirmar que los componentes extraídos son validos. [Lar06]

Tabla 13. Componentes de carga de los dos principales componentes extraídos del conjunto de datos de pruebas.

	Componente	
	1	2
Z LN # de productos individuales comprados	0.908	
Z LN # de diferentes clases de productos.	0.878	
Z LN visitas de compras	0.858	
Z LN tiempo promedio entre visitas	-0.867	
Z LN total de ventas netas	0.828	
Z promociones enviadas por correo		0.942
Z # de promociones		0.928

Fuente: [Lar06]

Análisis de clustering: algoritmo de clustering BIRCH.

El siguiente paso fue aplicar el algoritmo de BIRCH. [Lar06]. Este algoritmo requiere solo una pasada a través del conjunto de datos y así representar una solución escalable para conjuntos de datos muy grandes. El algoritmo tiene dos pasos principales y por lo cual es llamado clustering de dos pasos en la herramienta Clementine. En el primer paso el algoritmo preagrupa los registros en grandes números de pequeños subclusters construyendo un árbol de clusters característicos. En el segundo paso, el algoritmo luego combina esos subclusters en clusters de alto nivel, los cuales representan la solución al este algoritmo de clustering.

“Un beneficio de la implementación de este algoritmo es que a diferencia del clustering por K-means y por kahonen, es que el analista no necesita especificar el número deseado de clusters.” Para este caso de estudio el algoritmo retorno K = 3 clusters. [Lar06]

Los resultados que arrojo la herramienta fueron los siguientes: el cluster 2 con 8183 registros seguido por el cluster 3 y luego por el 1 con 7891 y 5666 registros, respectivamente. [Lar06]

Cluster 1: compradores que gastan moderadamente [Lar06]

- Este cluster tiene la más alta proporción de clientes que han comprado alguna vez un traje.
- La proporción de los que ha comprado alguna vez pantalones formales es 6 veces mas alta que en el cluster 2.
- El total de ventas netas para este cluster es moderada, situado no muy lejos de la media global.
- La uniformidad de producto es alta, lo que significa que estos compradores tienden a focalizarse en tipos particulares de ropa.
- Los hábitos de compra globales, sin embargo, no indican que estos son los clientes más leales, puesto que las visitas de compras y la cantidad gastada el mes pasado es baja, mientras que el tiempo entre visitas es alto. También este cluster no ha tendido a responder a las promociones el año pasado.

Cluster 2: compradores casuales que gastan poco [Lar06]

- Este cluster tiene las ventas totales netas mas bajas, con una media cercana a una desviación estándar, debajo del promedio global.

- Comparado con el cluster 1, este cluster tiende a comprar ropa mas casual, teniendo mas del doble de proporción de compras de pantalones casuales, y la mas alta cantidad de gasto en suéteres.
- Este cluster no esta interesado en trajes, solo el 0.1% ha comprado uno.
- Este cluster es similar al cluster 1 en algunos aspectos, como el numero bajo de visitas de compras, el bajo gasto en el mes pasado, una uniformidad de producto alta, un tiempo grande entre visitas, y el bajo porcentaje de respuestas a promociones pasadas.

Cluster 3: frecuente, altos gastos, compradores sensitivos [Lar06]

- La media de visitas entre compras y la media del total de ventas netas están cada una a aproximadamente una desviación estándar sobre el promedio global, lo que significa que este cluster representa compradores frecuentes que tienden a gastar grandes cantidades.
- Estos compradores tienen una baja uniformidad de producto, lo que significa que no se focalizan en algún tipo de ropa en particular.
- Estos compradores son sensitivos, al menos cerca del 90% de ellos ha respondido a una campaña de mercadeo en el año anterior.
- Una mayoría de estos compradores, tienen una tarjeta de crédito la cual figura en el expediente, a diferencia de los otros 2 clusters.
- Este cluster compra en línea, en un porcentaje 4 veces mayor que los otros 2.

Basado en estos resultados, los modelos de clasificación a continuación tendrán las siguientes entradas:

Modelo de colección A: (incluye análisis de principal componentes) modelos apropiados para perfilar clientes, análisis de variables y predicción. [Lar06]

- Las 71 variables listadas en la figura 14, menos las 7 variables de la tabla 11, usadas para la construcción de los componentes principales.
- Los dos principales componentes construidos usando las variables de la tabla 11
- Los cluster descubiertos por el algoritmo BIRCH de dos pasos.

Modelo de colección B: (no incluye PCA): modelos usados para la predicción de un objetivo. [Lar06]

- Las 71 variables listadas en la figura 14.
- Los cluster descubiertos por el algoritmo BIRCH de dos pasos.

Equilibrando el conjunto de datos de entrenamiento

Para lo modelos de clasificación en donde una de las clases de variables objetivo tienen una baja frecuencia relativa que otras clases, el equilibrio es una actividad recomendable. La idea que es que el conjunto de datos de entrenamiento contenga registros de diferentes tipos. [Lar06]

Un beneficio del equilibrio de los datos es suministrar a los algoritmos de clasificación con un rico balance de registros para cada resultado de la clasificación, de modo que cada algoritmo tenga una oportunidad de aprender sobre los distintos tipos de registros. Para este caso de estudio, el conjunto de datos de entrenamiento contiene 18.129 (83.4%) clientes que no han respondido a una campaña de mercadeo mediante correo directo y 3.611 (16.6%) clientes que si han respondido. Aunque es posible proceder directamente con los algoritmos de clasificación con este grado de desequilibrio, es recomendable que el balance sea aplicado, así la clase minoría contiene al menos 25% de los registros, y quizá a lo sumo 50%, dependiendo del problema especifico. [Lar06]

En este caso de estudio, se aplico un balanceo para lograr obtener una distribución aproximado de 50%-50% para las clases de los que responden y los que no responden a la campaña. Para hacer esto se utilizo un modulo de Clementine para retener aproximadamente el 20% los registros de aquellos que no responden, seleccionados aleatoriamente. El resultado del balance fue el siguiente: 3686 (50.5%) de registros de personas que no responden, y 3611 (49.5%) de personas que responden. [Lar06]

Por otra parte el conjunto de datos de prueba nunca se balancea. El conjunto de datos de prueba representa nuevos datos que los modelos aun no han visto. Ciertamente los datos del mañana no serán balanceados para los modelos de clasificación, es por eso que este conjunto de datos en si mismo no debe ser balanceados. La idea es que cuando se apliquen los modelos al conjunto de datos de pruebas, trabajen simulando el mundo real. [Lar06]

Estableciendo la línea de base de rendimiento del modelo

Para poder calibrar el rendimiento de los modelos candidatos, se necesito establecer un estándar de comparación, con aquellos modelos con los que se puedan comparar. Estos estándares de comparación se convierten en una especie de línea de base de rendimiento de modelos simples. Dos de estos simples modelos son: (1) “no enviar la promoción de mercadeo a nadie” y (2) “enviar la promoción de mercadeo a todo el mundo”. [Lar06]

Es claro que para aplicar estos modelos no es necesario aplicar minería de datos, pero si se quiere. El uso de la minería de datos se da con el fin de obtener modelos que funciones mejor que estos modelos de línea base, con gran margen de ganancia, pues esto es lo que justifica un proyecto de este tipo. [Lar06]

En la tabla 8, donde se establecían los costos y beneficios de este caso de estudio. Aplicando esos costos y beneficios a estos dos modelos de base, se obtienen para el conjunto de datos de entrenamiento (5908 respuesta negativas y 1151 respuestas positivas) las medidas de rendimiento mostradas en la siguiente tabla. [Lar06]

Tabla 14.Medidas del rendimiento de los 2 modelos línea de base.

	Costo TN	**Costo TP**	**Costo FN**	**Costo FP**	**Porcentaje error global**	**Costo global.**
Modelo	$0	-$26.4	$28.4	$2		
No enviar a nadie	5908	0	1151	0	16.3%	$32,688.40 ($4.63 por cliente)
Enviar a todo el mundo	0	1151	0	5908	84.7%	–$18,570.40 (–$2.63 por cliente)

Fuente: [Lar06]

Una comparación de los errores globales podría indicar una preferencia por el modelo de “Enviar a todo el mundo”. Sin embargo los proyectos de minería deben ser evaluados en término de factores reales como lo son los costos y beneficios. En este caso el modelo “no enviar a nadie” le esta costando a la empresa $4.63 por cliente, de perdida de beneficios. De los clientes en este conjunto de datos el 16.3% podría responder positivamente a la campaña. Por consiguiente este modelo debe considerarse como un completo fracaso. Por otra parte el modelo “enviar a todo el mundo” esta actualmente dejando ganancias a la compañía, un estimado de $2.63 por cliente. Esta estadística de “por cliente” contiene todos los clientes en el conjunto de datos, incluido los que no responden. Por consiguiente es el modelo “enviar a todo el mundo” el que se considera como la línea base, y es la ganancia de $2.63 por cliente la definida como estándar de comparación de beneficio que cualquier modelo candidato debe mejorar. [Lar06]

Modelo de colección A: usando los componentes principales

Se comenzó el modelado aplicando los cuatro algoritmos mencionados anteriormente para obtener los modelos de clasificación al conjunto de datos usando componentes principales. Los resultados obtenidos se muestran en la tabla 15.

Tabla 15. Resultados del rendimiento de los modelos de clasificación usando componentes principal

	Costo TN	**Costo TP**	**Costo FN**	**Costo FP**	**Porcentaje error global**	**Costo global por cliente.**
Modelo	$0	-$26.4	$28.4	$2		
Redes neuronales	4694	672	479 9.3%	1214 64.4%	24%	-$0.24
CART	4348	829	322 6.9%	1560 65.3%	26.7%	-$1.36
C5.0	4465	782	369 7.6%	1443 64.9%	25.7%	-$1.03
Regresión logística	4293	872	279 6.1%	1615 64.9%	26.8%	-$1.68

Fuente: [Lar06]

Nótese que los porcentajes indicados en las columnas FN y FP representan el porcentaje falso negativo y el porcentaje falso positivo, respectivamente. Es decir, el porcentaje FP = FP/FP+TP y el porcentaje FN = FN/FN+TN. El modelo de regresión logística funciona mejor que los otros tres, con una media de beneficio estimada de beneficio $1.68 por cliente. Sin embargo, es un punto discutible que ninguno de los modelos se acerca al estándar de comparación mínimo de $2.63 por cliente establecido para el modelo “enviar todos”. [Lar06]

¿Pero porque el rendimiento de estos modelos es tan pobre? La respuesta es que no se aplicaron costos de error en la clasificación. Para desarrollar modelos candidatos que se evaluaran usando una matriz de costo/beneficio, se debió buscar como encajar esos costos dentro de los modelos. En Clementine, dos algoritmos de clasificación están equipados con mecanismos explícitos para definir

costos de error en la clasificación asimétrica: C5.0 y CART. Por consiguiente, el siguiente paso a seguir fue desarrollar modelos de árboles de decisión usando costos de error en la clasificación en C5.0 y CART. Se procedió a definir el costo de tomar una decisión falsa negativa en $28.4 y el costo de tomar una decisión falsa positiva en $2, no hay mecanismo para definir el beneficio de una decisión verdadera positiva en $26.4 y la verdadera negativa en $1.0. Se debería notar que usando estos valores para definir los costos de error en la clasificación, es equivalente a fijar el costo del falso negativo en $14.2 y del falso positivo en $1. [Lar06]

Infortunadamente, la aplicación de estos costos resultantes en los modelos CART y C5.0 clasifico a todos los clientes como que responderían a la campaña, similar al modelo de "enviar a todos". Evidentemente la combinación del equilibrio que se hizo al 50% y estos fuertes costos de error en la clasificación fueron demasiado costosos para que los dos modelos predijeran negativamente. Por consiguiente los costos de error en la clasificación fueron reducidos del radio 14.2-1.0 al radio 10.0-1.0, con un costo falso negativo igual $10 y un costo falso positivo igual a $1.0. [Lar06]. Las medidas de rendimiento resultante se muestran en la tabla 16

Tabla 16. Resultados de rendimiento para los modelos de clasificación CART y C5.0 usando costos de error en la clasificación 10-1.

	Costo TN	**Costo TP**	**Costo FN**	**Costo FP**	**Porcentaje error global**	**Costo global por cliente**
Modelo	$0	-$26.4	$28.4	$2		
CART	754	1147	4 0.5%	5154 81.8%	73.1%	–$2.81
C5.0	958	1153	8 0.9%	5050 81.5%	71.7%	–$2.81

Fuente: [Lar06]

Inesperadamente, con la aplicación de costos de error en la clasificación en la etapa de la construcción del modelo, el beneficio global por cliente ha saltado por

más de un dólar. Ambos modelos el CART y C5.0 funcionaron ahora, mejor que el modelo de base. [Lar06]

Sobre balance como un reemplazo para los costos de error en la clasificación

La tabla 16 no contiene los modelos con redes neuronales o con regresión logística, esto porque Clementine no tiene un método especifico para aplicar costos de de error en la clasificación para estos algoritmos. No obstante hay un método alternativo efectos de decisión similares a aquellos suministrados por los costos de error en la clasificación**.** Este método alternativo es el sobre balance. [Lar06]

La tabla 17 contiene los resultados de rendimiento para una serie de modelos de redes neuronales, si usar componentes principales, y con varios niveles de balanceo o equilibrio. [Lar06]

Tabla 17. Resultados de rendimiento para modelos de redes neuronales para varios niveles de balance y sobre balance.

	Costo TN	**Costo TP**	**Costo FN**	**Costo FP**	**Porcentaje error global**	**Costo global por cliente**
Modelo	$0	-$26.4	$28.4	$2		
Sin Balance 16.3%–83.7%	5865	124	1027 14.9%	43 25.7%	15.2%	$3.68
Con Balance 50%–50%	4694	672	479 9.3%	1214 64.4%	24%	-$0.24
Sobre balance 65%–35%	1918	1092	59 3%	3990 78.5%	57.4%	-$2.72
Sobre balance 80%–20%	1032	1129	22 2.1%	4876 81.2%	69.4%	-$2.75
Sobre balance 90%–10%	592	1141	10 1.7%	5316 82.3%	75.4%	-$2.72

Fuente: [Lar06]

En la tabla se puede ver que los modelos que están sobre balanceados cada uno funciona mejor que el modelo de línea base, aun así a ninguno de estos modelos se les ha aplicado costos de error en la clasificación directamente. [Lar06]

El rendimiento mas optimo fue obtenido con redes neuronales usando un índice de sobre balance de 80%-20%. Lo que se hizo después fue comparar este rendimiento con el de los otros 3 algoritmos usando el mismo radio. [Lar06]

Tabla 18. Resultados del rendimiento de los 4 algoritmos usando un índice de sobre balance de 80%-20%, omitiendo los cluster de estilos de vida.

	Costo TN	**Costo TP**	**Costo FN**	**Costo FP**	**Porcentaje error global**	**Costo global por cliente**
Modelo	$0	-$26.4	$28.4	$2		
Red neuronal	885	1132	19 2.1%	5023 81.6%	71.4%	-$2.73
CART	1724	1111	40 2.3%	4184 79%	59.8%	-$2.81
C5.0	1467	116	35 2.3%	4441 79.9%	63.4%	-$2.77
Regresión logística	2389	1106	45 1.8%	3519 76.1%	50.5%	-$2.96

Fuente: [Lar06]

Para la construcción de estos modelos se excluyo la variable cluster de estilos de vida, pues parece no estar estrictamente asociada con la respuesta o no respuesta a la campaña, por lo que se omitió esta campaña y se corrieron los modelos. Los resultados se muestran en la tabla 18. [Lar06]

Sin embargo el mejor modelo es la regresión logística usando un balance de 80%-20%, sin componentes principales, y excluyendo la variable cluster de estilo de vida. Obsérvese que sin la aplicación del sobre balance como un sustituto a los costos de error en la clasificación, no se tendría acceso al modelo de regresión logística. Este modelo provee un beneficio estimado por cliente de $2.96, lo que representa una mejora del 45% sobre los modelos que aplicaron costos de error

en la clasificación ($2.81), directamente comprado con el modelo de base ($2.63). [Lar06]

Combinación de modelos: Votación

Los analistas con el fin de obtener mejores resultados combinaron modelos con el fin de unir fortalezas y pulir las debilidades que tiene cada uno por aparte. [Lar06]

Un método para combinar modelos es usando la votación simple. Para cada registro, cada modelo proporciona una predicción de respuesta (1) o no respuesta (0), después se cuentan los votos de cada registro. Por ejemplo en este caso se están utilizando 4 modelos de clasificación al problema de respuesta a la promoción. Por lo tanto, los registros pueden recibir votos de 0 a 4 para predecir la respuesta. En este caso, a nivel global, se puede predecir una respuesta positiva a la promoción basados en uno de los siguientes 4 criterios: [Lar06]

A. Enviar una promoción, solo si los 4 modelos predicen respuestas.
B. Enviar una promoción, solo si 3 de los 4 modelos predicen respuestas.
C. Enviar una promoción, si al menos 2 de los modelos predicen respuestas.
D. Enviar una promoción, si uno de los modelos predice respuesta.

Claramente el criterio A tendería a proteger contra los falsos positivos, dado que los 4 algoritmos de clasificación, tendrían que estar de acuerdo en una predicción positiva de acuerdo a este criterio. Similarmente el criterio D tiende a proteger contra los falsos negativos, donde al menos uno de los algoritmos necesitaría predecir una respuesta positiva. Cada uno de estos 4 criterios en efecto define un modelo de combinación cuyo rendimiento puede ser evaluado. La siguiente tabla muestra los resultados de rendimiento para cada uno de los 4 modelos combinados. La mejor combinación de modelos es el modelo definido por el criterio B, “enviar una promoción, solo si 3 de los 4 modelos predicen respuestas”.

Este modelo tiene una representación alternativa, enviar una promoción, solo si la mayoría de los modelos predicen respuesta. [Lar06]

Tabla 19. Resultados de rendimiento para los 4 modelos de conteo de votos usando un índice de sobre balance de 80%-20% omitiendo la variable cluster de estilo de vida.

	Costo TN	**Costo TP**	**Costo FN**	**Costo FP**	**Porcentaje error global**	**Costo global por cliente**
Modelo	$0	-$26.4	$28.4	$2		
Enviar una promoción, solo si los 4 modelos predicen respuestas	2772	1067	84 2.9%	3136 74.6%	45.6%	-$2.76
Enviar una promoción, solo si 3 de los 4 modelos predicen respuestas	1936	1115	36 1.8%	3972 78.1%	56.8%	-$2.9
Enviar una promoción, si al menos 2 de los modelos predicen respuestas	1207	1135	16 1.3%	4701 806%	66.8%	-$2.85
Enviar una promoción, solo si 3 de los 4 modelos predicen respuestas	550	1148	3 0.5%	5358 82.4%	75.9%	-$2.76

Fuente: [Lar06]

Se ve que ninguno de los modelos supero, el resultado que se obtuvo con el modelo de regresión logística usando un índice de sobre balance de 80%-20%, y omitiendo los cluster de estilos de vida, por lo que se pueden descartar estos resultados. [Lar06]

Según Larose “una de las desventajas de la combinación de modelos es la dificultad a la hora de interpretar los resultados.”

Modelo de colección B: Modelos sin PCA.

Finalmente, se examinaron los modelos que no incluyen componentes principales. En lugar de eso, siguen usando las variables que se muestran en la tabla 11 y así no serán usadas para cualquier propósito excepto para la predicción de la variable objetivo. Por otra parte, puesto que el conjunto de variables es altamente predictivo de la respuesta a la campaña, se espera que los modelos sin PCA funcionen mejor que los modelos en términos de la predicción a la respuesta.

La estrategia[25] a seguir en esta parte fue seguir trabajando como con los modelos con PCA, con una adición especial: [Lar06]

1. aplicar modelos CART y C5.0, usando costos de error en la clasificación y un índice de balance de 50%.
2. aplicar los 4 modelos de clasificación con sobre balance del 80%.
3. combinar los 4 modelos de clasificación usando votación.
4. combinar los 4 modelos de clasificación usando probabilidad media de respuesta.

Se comenzó aplicando los algoritmos de árboles de decisión, CART y C5.0 usando costos de error de clasificación de 14.2-1.0 y el índice de balance de 50%-50%. Los resultados obtenidos se muestran en la tabla 20. Los dos modelos funcionan mejor que los mejores modelos son PCA, con un beneficio estimado por persona de $3.04 y $3.01 comprado con los $2.96 para el modelo de regresión logística con PCA. [Lar06]

[25] Estrategia definida en [Lar06].

Tabla 20. Resultados de rendimiento para los modelos de clasificación CART y C5.0 usando costos de error en la clasificación 14-2.

	Costo TN	Costo TP	Costo FN	Costo FP	Porcentaje error global	Costo global por cliente
Modelo	$0	-$26.4	$28.4	$2		
CART	1645	1140	11 0.7%	4263 78.9%	60.5%	-$3.01
C5.0	1562	1147	4 0.3%	4346 79.1%	61.6%	-$3.04

Fuente: [Lar06]

Luego se aplico el sobre balance como un substituto para los costos de error en la calificación, justo como se hizo en los modelos con PCA. La tabla 21 muestra los resultados del rendimiento para los 4 algoritmos, usando un índice de sobre balance del 80%. [Lar06]

Tabla 21. Resultados del rendimiento de los 4 algoritmos usando un índice de sobre balance de 80%-20%.

	Costo TN	Costo TP	Costo FN	Costo FP	Porcentaje error global	Costo global por cliente
Modelo	$0	-$26.4	$28.4	$2		
Red neuronal	1301	1123	28 2.1%	4607 80.4%	65.7%	-$2.78
CART	2780	1100	51 1.8%	3128 74%	45%	-$3.02
C5.0	2640	1121	30 1.1%	3268 74.5%	46.7%	-$3.15
Regresión logística	2853	1110	41 1.4%	3055 73.3%	43.9%	-$3.12

Fuente: [Lar06]

Nótese, la extensa disparidad en el rendimiento de los modelos. En este caso, el modelo C5.0 es el ganador con un beneficio estimado de 3.15 por cliente, representando el mejor rendimiento de predicción para un modelo individual en este caso de estudio. El modelo de regresión logística no esta muy lejos, le sigue con un beneficio de $3.12, y la red neuronal muestra el peor rendimiento. [Lar06]

El siguiente paso fue combinar los 4 modelos, primero usando el método de la votación, las métricas de rendimiento se muestran en la siguiente tabla para los 4

métodos de conteo de votos, donde una vez mas se usa un índice de sobre balance del 80%. [Lar06]

Tabla 22. Resultados de rendimiento para los 4 modelos de conteo de votos, usando un índice de sobre balance de 80%-20% para modelos sin PCA.

	Costo TN	Costo TP	Costo FN	Costo FP	Porcentaje error global	Costo global por cliente
Modelo	$0	-$26.4	$28.4	$2		
Enviar una promoción, solo si los 4 modelos predicen respuestas	3307	1065	86 2.5%	2601 70.9%	38.1%	-$2.9
Enviar una promoción, solo si 3 de los 4 modelos predicen respuestas	2835	1111	40 1.4%	3073 73.4%	44.1%	-$3.12
Enviar una promoción, si al menos 2 de los modelos predicen respuestas	2537	1133	18 0.7%	3551 75.8%	50.6%	-$3.16
Enviar una promoción, solo si 3 de los 4 modelos predicen respuestas	1075	1145	6 0.6%	4833 80.8%	68.6%	-$2.89

Fuente: [Lar06]

Los resultados de los modelos combinados pueden ser un poco sorprendente, dado que el modelo de combinación, enviar una promoción solo si al menos 2 de los modelos predicen respuestas, ha funcionado mejor que todos los modelos de clasificación individuales con un beneficio promedio global por cliente de $3.16. Esto representa la sinergia de la estrategia de combinación de modelos, donde la combinación de los modelos es en cierto sentido mayor que la suma de sus partes. Aquí, el beneficio mas grande es obtenido cuando al menos dos modelos

están de acuerdo en enviar la promoción a un receptor potencial. El método de combinar modelos mediante votación ha dado mejores resultado que con los modelos individuales. [Lar06]

Combinar modelos usando probabilidad media de respuesta

La votación no es el único modelo para combinar modelos. El método de votación representa para cada modelo, una decisión de alza o de baja, de un extremos al otro, sin el cuidado de medir la confianza en la decisión. [Lar06]

Afortunadamente tales medidas de confianza están presentes en Clementine, pero con un poco de diferencia. Para cada resultado de los modelos, Clementine no solo reporta la decisión sino que también reporta un campo continuo que esta relacionado a la confianza del algoritmo en esta decisión. Cuando se usa este campo continuo, se obtiene una nueva variable que mide para cada registro la probabilidad que este cliente en particular responda positivamente a la promoción. La variable se obtiene de la siguiente forma: [Lar06]

- Si la predicción es positiva, luego la probabilidad de respuesta = 0.5 + (confianza reportada)/2.
- Si la predicción es negativa, luego la probabilidad de respuesta = 0.5 - (confianza reportada)/2.

Para cada modelo, el modelo de probabilidad de respuesta (MRPs por sus siglas en ingles. [Lar06]) Fueron calculados usando esta formula. Luego la media MRP fue encontrada dividiendo la suma de los MRPs por 4.

El analista puede definir bandas que partan el conjunto de datos de acuerdo a varios valores de MRP. Cabe recordar que el error falso negativo es 14.2 veces

que el error falso positivo, por lo que se tendió a fijar estas particiones en un punto bajo. [Lar06]

La partición de los registros se hizo de acuerdo al siguiente criterio: MRP < 0.85 y MRP > 0.85. Puesto que es cerca de este valor que la proporción de respondedores positivos comienza a incrementar rápidamente. [Lar06]

Sin embargo como se muestra en la tabla 23. El modelo basado en tal partición es sub-optimo, dado que permite muchos falsos positivos. Como resultado, la partición optima esta cerca del 50% de probabilidad. [Lar06]

Tabla 23. Métricas de rendimiento para los modelos definidos por partición de varios valores de MRP.

	Costo TN	**Costo TP**	**Costo FN**	**Costo FP**	**Porcentaje error global**	**Costo global por cliente**
Modelo	$0	-$26.4	$28.4	$2		
Partición: MRP < 0.95 MRP ≥ 0.95	5648	353	798 12.4%	260 42.4%	15%	+$1.96
Partición: MRP < 0.85 MRP ≥ 0.85	3810	994	157 4%	2098 67.8%	31.9%	-$2.49
Partición: MRP < 0.52 MRP ≥ 0.52	2738	1121	30 1.1%	3170 73.9%	45.3%	-$3.1736
Partición: MRP < 0.51 MRP ≥ 0.51	2686	1123	28 1%	3222 74.2%	46%	-$3.1744
Partición: MRP < 0.46 MRP ≥ 0.46	2493	1129	22 0.9%	3415 75.2%	48.7%	-3.1666

Fuente: [Lar06]

Se puede ver que el modelo de combinación continúa definida en la partición MRP = 0.51 es el mejor modelo para predecir la respuesta a la campaña de mercadeo mediante correo directo. Este modelo provee un beneficio estimado de $3.1744 por cada promoción enviada. Esto comparado con el modelo de base tiene una diferencia de 20.7% o 54.44 centavos por usuario. Por ejemplo si se quieren

enviar 100.000 correos a los clientes. El beneficio estimado se incrementa en $54440. Este incremento del beneficio es debido a la disminución en costos asociados al envió de promociones a clientes que no responderían. [Lar06]

Para mas información sobre la fase de modelado y evaluación de este caso de estudio se puede recurrir a la fuente. [Lar06].

2.2.4. Despliegue y utilización de los resultados de minería de datos:

Como se mencionaba en la etapa de entendimiento del negocio, un proyecto de este tipo que trata el problema de un negocio, hace necesario para los gerentes medir el éxito de los modelos en términos económicos; por lo que se establecieron los costos y beneficios de acuerdo a las circunstancias propias del oficio.

Esta tabla de costo/beneficio, se utilizo para medir las ganancias por usuarios que dejaba cada modelo de predicción. Fueron muchos los modelos construidos, pero el modelo que arrojaba las mejores ganancias fue el modelo con la partición MRP < 0.51, MRP > 0.51, construido con el método de combinación de modelos usando la probabilidad media de respuesta.

Este modelo mejora en un 20.7% al modelo de base, lo que representa grandes ganancias para la cadena de tiendas RK Clothes, pues se identifico de una forma mas precisa las personas que responderían a la campaña de promoción utilizando correo directo, y se focalizaron las campañas de marketing de una manera mas concreta.

A parte de las ganancias que se obtuvieron por utilizar la minería de datos para predecir la respuesta que tendrían los clientes a la campaña de mercadeo, estos recursos también le permitieron a la organización mejorar en otros aspectos y

comenzar otros proyectos utilizando minería de datos con el animo de innovar y generar ventajas competitivas en otras áreas.

Por ejemplo la cadena de almacenes, comenzó a tener relaciones más rentables con los mejores clientes. Conociendo el perfil del cliente mas propenso a comprar los productos. Y se logro enfocar las campañas de marketing para conseguir más respuestas y un mayor retorno de las inversiones hechas, al realizar este proyecto. [CRM05]

También se logro identificar los principales grupos de clientes y se entendió de cierta forma su comportamiento, con lo que se comenzaran a desarrollar otras campañas que fueron exitosas para la organización. Al identificasen los perfiles de los clientes, se pudieron crear interacciones más precisas y rentables. [CRM05]

Se conocieron las necesidades de los clientes de forma que se incremento el beneficio adoptando estrategias de Cross – and – up Selling. Una vez que se estableció la relación con el cliente, se incremento su valor ofreciendo productos y servicios adicionales que le interesasen a este. Así los clientes tenían un mayor grado de satisfacción con la cadena, y se reforzó la decisión de estos de volver a comprar ropa en alguno de los almacenes. También se identificaron los productos que se compraban juntos, de forma que se repitió la promoción a clientes que tenían características similares. [ESP03]

Otro punto en el que se mejoro como resultado de este proyecto, fue en los puntos de contacto con el cliente. Se mejoro la página Web, con herramientas que permitieron hacerla lo suficientemente inteligente como para personalizar los contenidos para cada visitante y presentarles las ofertas más atractivas. [CRM05]

Se incremento la rentabilidad durante todo el ciclo de vida de los clientes reduciendo el coste de captación, maximizando su crecimiento y extendiendo la relación, lo más posible. [CRM05]
La empresa mejoro sus sistemas CRM, para centrar su atención en el sitio correcto, sobre la experiencia total de los clientes.

Según la pagina Web[26] de CRM con SPSS “Aplicar un CRM inteligente permite actuar sobre las oportunidades más rentables, que de otra manera pasarían inadvertidas. Con un análisis tan potente como el que da el CRM se tiene el conocimiento necesario de los clientes para responder a las cuestiones estratégicas que se plantearon para mejorar la rentabilidad de sus relaciones.”

Se estableció un sistema en las tiendas de la cadena, que aprende de cada interacción con el cliente, para mejorar las relaciones con este, pero en tiempo real.

[26] http://www.spss.com/la/soluciones/crm.htm

2.3. SEGMENTACIÓN DE CLIENTES PARA CREAR ESTRATEGIAS DE MERCADEO Y VENTAS POR SEGMENTOS

Este caso trata de encontrar los diferentes grupos de clientes que tiene un banco.

Se trabajo la tarea de segmentación y se utilizo la herramienta DB2 Intelligent Miner for Data de IBM.

2.3.1. Comprensión del negocio o del tema a tratar:

El banco de Québec[27] es uno de los 10 bancos más grandes de Norteamérica y siempre se ha caracterizado por estar un paso por delante de sus más cercanos competidores. Para continuar con esta ventaja quieren mejorar su área de inteligencia de negocios.

Como se sabe que la cultura de un banco y la forma de pensar de sus empleados esta muy influenciada por el esquema de segmentación que el banco adopta para su mercado, se hizo importante que este esquema estuviera alineado con las necesidades de los clientes, y que fuera un factor determinante para dar ventaja competitiva única y duradera al banco. [SLB+01]

Esto se hace esencial en un mercado tan competido como el que se ha venido dando en los últimos años. La necesidad de tratar a un cliente de una forma más individualizada, de maximizar su satisfacción y su rentabilidad se hace cada vez más imperiosa para tratar de ganar en ese mercado. Darles a los clientes el mismo servicio significa tratarlos de una forma diferente, pero este tipo de interacción no es viable económicamente en mercados tan grandes. La solución a este problema es agrupar clientes con comportamientos similares y desarrollar

[27] Para ver la información real de la organización base, se puede dirigir al documento: Bank of Montreal becomes master of its destiny with IBM scoring tool. [BOM01]

estrategias específicas de ventas y mercadeo a ese grupo que hacerlo de una forma individual. [SLB+01]

El problema radica en que la segmentación tradicional ha envuelto grupos de clientes, en pequeños grupos de clusters usando reglas del negocio, basadas en unas cuantas variables. Las más usadas son la edad, el ingreso, y una medida para el grado de envolvimiento con la compañía. [SLB+01]

La dificultad radica en el desarrollo y entendimiento de estas reglas, y que básicamente los resultados que arroja esta técnica aunque parecen convincentes, no son lo mas esperado para una inversión tan alta de dinero y tiempo. [SLB+01]

El desafió es desarrollar un proyecto de minería de datos que permita hacer una segmentación de los clientes; que le permita al banco lograr sus acometidos.

Esta segmentación consiste en tener grupos de clientes con necesidades similares en relación al banco. Esta segmentación puede reflejar el orden natural del mercado, y en consecuencia la fuerza de ventas y mercadeo estará en capacidad de desarrollar estrategias más eficientes. [SLB+01]

La pregunta que se trato de responder es: ¿Cuáles son las características del comportamiento de los clientes del banco?

2.3.2. Comprensión y preparación de los datos a utilizarse:

Los datos requeridos para la segmentación de los clientes fueron agrupados en las siguientes categorías: [SLB+01]

- **Datos demográficos del cliente:** son los datos individuales del cliente. Algunos ejemplos son edad, sexo, estudios, entre otras.
- **Datos de filiación:** describe la dimensión y calidad de relación que se tiene con un cliente en particular. Esto incluye datos sobre tenencias, propiedad de producto, utilización de producto, canal de utilización, fecha de establecimiento, quejas, advertencias, etc.

  Adicionalmente, estos datos pueden incluir información sobre ventas previas, y actividades de marketing, preferencias de los clientes relacionadas con el banco, entre otras.

- **Datos de transacciones:** describe el número y tamaño de las transacciones hechas por un cliente.

- **Datos adicionales:** cubren otros datos que pueden ser útiles para el entendimiento de la interacción con el cliente. Estos datos son: segmentación actual, datos económicos generales, relación con otras compañías o grupos, entre otras.

Datos sugeridos para la segmentación

La siguiente lista contiene las variables que podrían ser utilizadas para la segmentación. Esta lista no es muy exhaustiva, pero propone algunas variables que podrían ser apropiadas para dar un buenos resultados. Claro que estos datos se fueron ajustando sobre el tiempo para refinar el modelo y recoger nuevos datos sobre los clientes.

Tabla 24. Variables a usar

	Descripción de la variable	Nombre de la variable
1	Identificación del cliente	Customer_id
2	Identificación del grupo familiar	Household_id

3	Edad	Age
4	Sexo	Sex
5	Dirección	Location
6	Ingreso anual	Annual_income
7	Dueño de automóvil	Car_owner
8	Dueño de casa	Home_owner
9	Estado civil	Marital_status
10	Hijos	Has_children
11	Tiempo como cliente	Time_as_customer
12	Total depósitos	Total_deposits
13	Total deudas	Total_liabilities
14	Fondos disponibles totales	Total_disposable_funds
15	Tiene atrasos en pagos	Has_defaulted
16	Numero de pagos automáticos	No_automatic_payments
17	Salario depositado en el banco	Salary_deposited_in_bank
18	Dueño de producto	
19	Cuenta corriente	Has_Current_account
20	Tiene productos de ahorro	Has_Saving_products
21	Tiene productos de préstamo	Has_Loan_products
22	Tiene productos de inversión	Has_Investment_products
23	Tiene productos de pensión	Has_Pension_products
24	Tiene fondo mutualista	Has_Mutual_funds
25	Tiene tarjeta de crédito	Has_Credit_card
26	Tiene productos de la juventud	Has_Youth_products
27	Numero de transacciones	Number_of_transactions
28	Numero de transacciones usando:	
29	Uso del cajero	Teller_usage
30	Uso de la Terminal bancaria	ATM_usage
31	Uso de la Web del banco	Web_bank_usage
32	Uso del call center	Uses_call_center
33	Uso del kiosco	Uses_kiosk
34	Numero de transacciones de la cuenta actual	Number_of_current_acc_trans
35	Valor promedio de las transacciones	Avg_value_of_trans
36	Valor promedio de las transacciones de crédito	Avg_value_of_credit_trans
37	Valor promedio de las transacciones de debito	Avg_value_of_debit_trans
38	Segmentación actual	Current_Segmentation
39	Puntuación del crédito	Credit_score
40	Medida de rentabilidad.	Profitability_measure

Fuente: [SLB+01]

Entender la red de relaciones sobre el grupo familiar de un cliente, que si tiene una cuenta con otro familiar, si tiene una hipoteca con su esposa, una cuenta

comercial; es un factor significativo en la segmentación de la base de datos, porque el banco esta haciendo campañas de marketing a los grupos familiares en vez de hacerlo a una sola persona. [SLB+01]

Para hacer que la interpretación fuera más fácil para las personas, se incorporaron convenciones del negocio en los datos. Para lograr esto se tuvo que discretizar las variables, ya que esto me da una forma de ampliar los datos con el conocimiento del negocio. [SLB+01]

Como los datos residen en los sistemas de producción internos o en la datawarehouse, para los propósitos de la minería de datos que tiene el banco, estos datos se tuvieron que transformar, para que pudieran ser accedidos por la herramienta de minería.

Cuando se hace segmentación de clientes, todos los datos deben usarse en su totalidad en vez de usar una muestra. Cuando se usan muestras se puede perder detalles y un numero importante de nichos en la base de clientes puede que no sea descubierto. [SLB+01]

2.3.3. Agregación y transformación de los datos

Se hizo necesario producir un número de agregaciones a las variables de filiación y de transacciones; estas agregaciones fueron añadidas al modelo de datos.

Para los datos de filiación se necesito agrupar los productos en un pequeño número de grupos, que representan los diferentes tipos de productos que se ofrecen en el banco. Los productos del banco son relativamente homogéneos dentro de los diferentes grupos de productos. A parte de la identificación de grupos de productos se presento también la idea de crear un indicador de pertenencia para productos que se saben están relacionados a una preferencia

especifica o a un tipo de cliente. Estos productos pueden estar asociados con características específicas como prestamos para construcción, tarjetas de crédito, préstamos para estudio, comercio electrónico, entre otras. [SLB+01]

Los datos transaccionales fueron agregados sobre diferentes espacios de tiempo. Dependiendo del tipo de producto y el tipo de canal, la longitud de la espacio de tiempo puede variar de 1 a 12 meses. Un buen punto de partida fue hacer agregaciones para todas las variables de 1, 3, 6 y 12 meses. [SLB01]

2.3.4. Evaluación de los datos

Esta evaluación se hizo con el ánimo de encontrar cualquier anomalía dentro de los datos, y conocer la calidad de los datos antes de comenzar el proceso de minería. Básicamente lo que se hizo fue buscar variables que no tuvieran asignados algún valor, valores extremos y signos de correlación.

Este proceso también sirvió para tener cierto conocimiento y entender los datos que se tenían y también permito encontrar una forma apropiada de discretizar[28] las variables.

Básicamente esta etapa cubrió dos pasos los cuales se van a describir a continuación:

a) Búsqueda de valores extremos (outlying)

Se utilizo estadística descriptiva (histogramas y diagramas de círculos) para buscar básicamente aquellos valores extremos, variables con una distribución

[28] Reducir el número de valores de un atributo continuo dividiendo el rango del atributo en intervalos

inusual de los datos, que pudieran indicar un error sistemático en el proceso de obtención de los datos. [SLB+01]

Otra forma de búsqueda de posibles anomalías fue inspeccionando visualmente la distribución de las variables; en donde se encontraron errores tan simples como que un cliente pueda tener una edad de 189 años o variables que no tenían ningún valor asociado. [SLB+01]

Cuando se encontraban con este escenario, el procedimiento que se siguió fue reemplazar ese campo vació con un valor por defecto, que fuera valido, esta simple tarea evito caer en errores desde el comienzo. Pero también era importante guardar la información de cuales variables tenia un valor ausente; por lo que en el caso de las variables numéricas se crearon otras variables que sirvieran para guardar estos datos (Por ejemplo se utilizaba la letra M para ausente y la letra P para poblada). Cuando se trataba de variables categóricas se asignaba un indicador que me indicaba este tipo de situaciones. (Por ejemplo: “missing”). [SLB+01]

La importancia de las variables que no tienen ningún valor adjunto es más evidente en relación con los datos demográficos y aquellos de conexión e interacción. Cuando este tipo de variables tenia valores extraños le indicaban al banco que no tenían una relación con el cliente muy desarrollada. [SLB+01]

Las inspecciones visuales fueron complementadas con datos estadísticos como el mínimo, el máximo, el promedio y la desviación estándar. Si se encontraban errores estos eran corregidos mediante reemplazamiento o ajuste. [SLB+01]

b) Identificación de problemas con las variable

En este paso se removieron las variables redundantes o duplicadas. Entendiendo como redundantes aquellas que están altamente correlacionadas con otras variables y el cual si se usaban ambas variables en la segmentación no aportarían información adicional y hubieran podido distorsionar el resultado. Lo que se quiere evitar es que si un número de variables están altamente correlacionadas, no se usen siempre las mismas y se dejen por fuera otras que pueden aportar mas información. [SLB+01]

Se trataron de evitar las correlaciones combinando el conocimiento del negocio y el sentido común, esto fue de suma importancia porque con las variables categóricas no existe un método estadístico simple que permitiera identificar correlaciones. [SLB+01]

Este proceso de validación de datos aunque es tedioso es importante hacerlo porque permite obtener mejores resultados en la construcción del modelo.

2.3.5. Modelado y evaluación:

La técnica de minería utilizada para la extracción de las características de los clientes es la técnica de la segmentación o clustering. La ventaja de usar esta técnica es que no requiere un conocimiento a priori de los resultados para obtener los clusters.

Las técnicas de clustering trabajan agrupando clientes similares, tratando de maximizar las diferencias entre los diferentes grupos de usuarios. [SLB+01]

Cabe anotar que las diferentes técnicas de cluster producen diferentes vistas de los clientes. Como se podían obtener diferentes modelos de las diferentes

técnicas, lo que se trato de realizar fue un consenso entre las técnicas que se utilizaron, en donde se compararon los resultados obtenidos y se observo donde estaban de acuerdo y en desacuerdo. [SLB+01]

Las dos técnicas de clustering que se utilizaron fueron clutering demográfico[29] y clustering neural.

Clustering demográfico y clustering neural

Estas dos técnicas tienen diferentes formas de obtener los clusters de los datos y son por lo tanto apropiadas para usar bajo diferentes circunstancias.

El clustering demográfico ha sido inicialmente desarrollado para trabajar con datos demográficos que son generalmente grandes cantidades de variables categóricas. Por esta razón es que la técnica trabaja mejor con conjuntos de datos formados por este tipo de variables. Cuando trabaja con variables continuas la técnica las trata como si fueran variables categóricas. En contraste el cluster neural trabaja con variables numéricas continuas y mapea las categóricas a este tipo de variables. Sin embargo las dos técnicas pueden utilizarse como complemento la una de la otra y validar los resultados obtenidos. [SLB+01]

Otra diferencia entre estas técnicas es que para el clustering neural, el usuario tiene que especificar el número de clusters que él desea obtener. Mientras que con el clustering demográfico el numero de clusters es creado automáticamente basándose en las especificaciones del usuario y en el grado de similitud que tienen los registros dentro de un cluster. [SLB+01]

[29] También conocido como clustering basado en el algoritmo de los vecinos más cercanos utilizando un valor condorcet. Para mas información ver la referencia [BAB+01]

Similitud es un simple concepto que tiene un valor entre 0 y 1. Si el valor es 1 esto indica que los clientes en un cluster son idénticos. Si el valor es 0, entonces el cliente no concuerda con ese cluster. [SLB+01]

La similitud entre dos clientes es calculada comparando cada atributo de la variable del cliente, y se le asigna una puntuación de cuanto concuerdan las variables. Las puntuaciones son sumadas y divididas por el número de variables comparadas. Esto para el caso de las variables categóricas. Para las variables continuas se usan técnicas como la desviación estándar, para expresar la diferencia en términos de esta. [SLB+01]

Otro factor importante que se tiene que considerar en el algoritmo de clustering, es que si un cliente tiene una similitud aceptable con un grupo existente de clientes; esto no indica que automáticamente ese cliente debe ser asignado a ese cluster. Puede haber otros grupos en donde el cliente encaje mejor. El clustering demográfico encuentra una óptima combinación de clientes que maximiza la similitud de todos los clientes con cada cluster, mientras al mismo tiempo maximiza las diferencias entre los clusters resultantes. Para tener esto, se debe intentar maximizar el cálculo de un valor estadístico, el cual es llamado valor condorcet. [SLB+01]

Aplicación de la técnica de minería

Como la aplicación de la técnica del clustering es un proceso interactivo donde se trata de probar distintas combinaciones de variables y diferentes formas de discretizar las variables. La aplicación de la técnica consistió en usar las variables correctas, encontrar el número correcto de clusters mediante el algoritmo de cluster demográfico y luego comparar esos resultados con los arrojados por el cluster neural. [SLB+01]

Para la construcción del modelo de minería, se trato de reducir el número de variables, usando aquellas que fueran más significativas. Las razones para hacer es que la interpretación del negocio de los segmentos resultantes se hace más fácil con menos variables. Segundo, porque había que asegurarse de que los segmentos resultantes reflejen verdaderamente las similitudes y diferencias entre los usuarios. [SLB+01]

La reducción del número de variables se hizo con varios métodos, se escogieron y se quitaron aquellas variables que tenían el valor condorcet mas bajo y de las cuales se veía que no aportaban información relevante al negocio. [SLB+01]

Las variables que se iban descartando eran agrupadas como un conjunto de variables suplementarias, las cuales podrían ser utilizadas en algún momento para asistir en la interpretación de los resultados. [SLB+01]

Para encontrar el número apropiado de cluster, usando la técnica de segmentación demográfica, se empezó con un umbral de similitud de 0.5 el cual dio una primera impresión del comportamiento de los clientes. Este valor relativamente bajo del umbral, arrojo un numero pequeño de clusters, pero muy grandes en tamaño. [SLB+01]

El problema de haber escogido un umbral de similitud tan bajo es que este modelo dio como resultado clusters muy heterogéneos, y esto lo indica el valor condorcet[30] el cual es muy bajo. [SLB+01]

Se fueron obteniendo clusters más homogéneos, incrementando el umbral de similitud hasta que el valor condorcet alcanzara un nivel aceptable[31]. Tanto como

[30] No se sabe cual fue este valor; pues no esta especificado en las fuentes bibliograficas [SLB+01]
[31] Un valor mínimo de 0.65 se considera como aceptable para este caso. Ver [SLB+01]

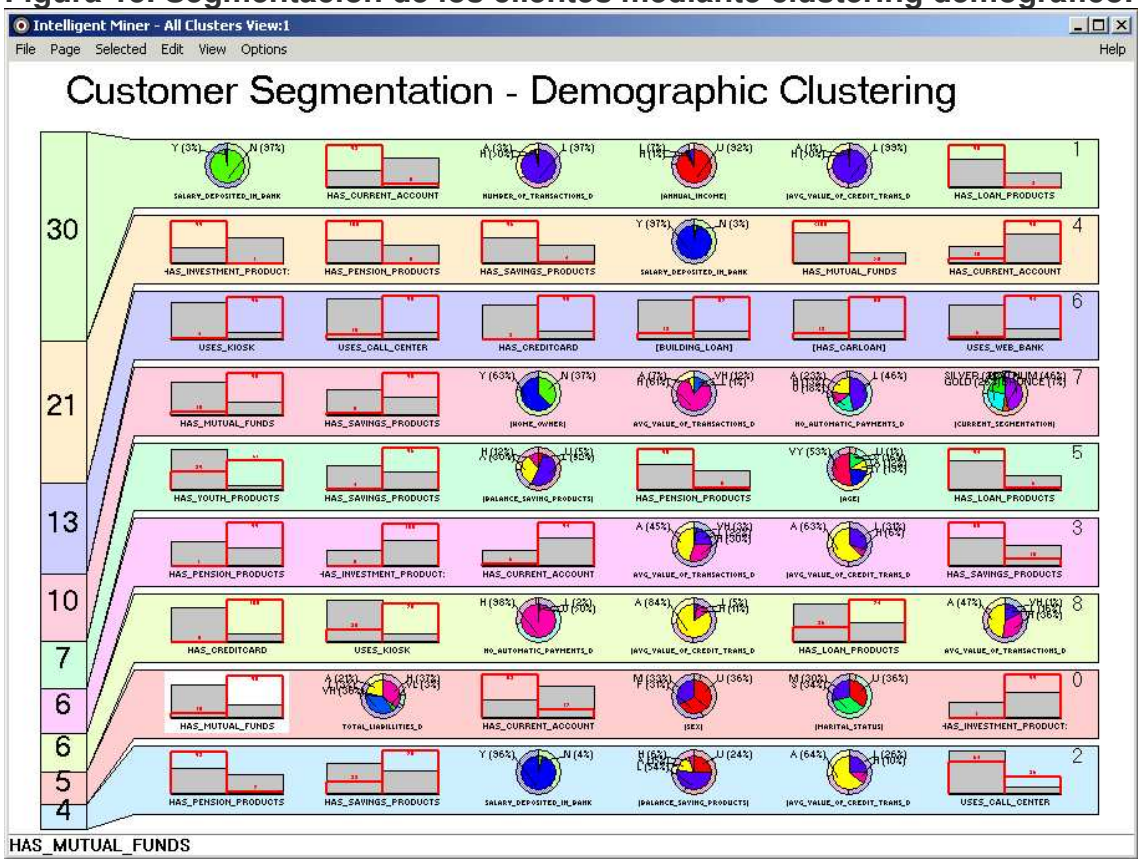
se fue incrementando el umbral de similitud, se fueron obteniendo mas clusters, ya que mas nichos se fueron singularizando.

El tamaño de los nichos en los que se concentrarían en el banco fue establecido en un mínimo del 2% de la población total, esto porque eran segmentos donde se podrían aplicar estrategias de ventas y mercadeo que si eran viales económicamente. Esto fue apropiado durante algunos pasos del análisis. Sin embargo, también desde el punto de vista de ventas y mercadeo el manejo de más de 9 clusters seria demasiado engorroso, y con este tamaño mínimo de los clusters se podrían obtener un número más grande de segmentos. Por lo cual lo que se hizo fue limitar el número máximo de clusters obtenidos a un número más manejable[32]. Para lograr esto se unieron los clientes de los clusters mas pequeños con los mas grandes, tratando de que fueran lo mas similar posible, hasta que se formaran los 9 clusters. [SLB+01]

El resultado obtenido para la técnica del clustering demográfico se muestra a continuación.

[32] El numero de clusters con los que se trabajaría seria 9.

Figura 15. Segmentación de los clientes mediante clustering demográfico.

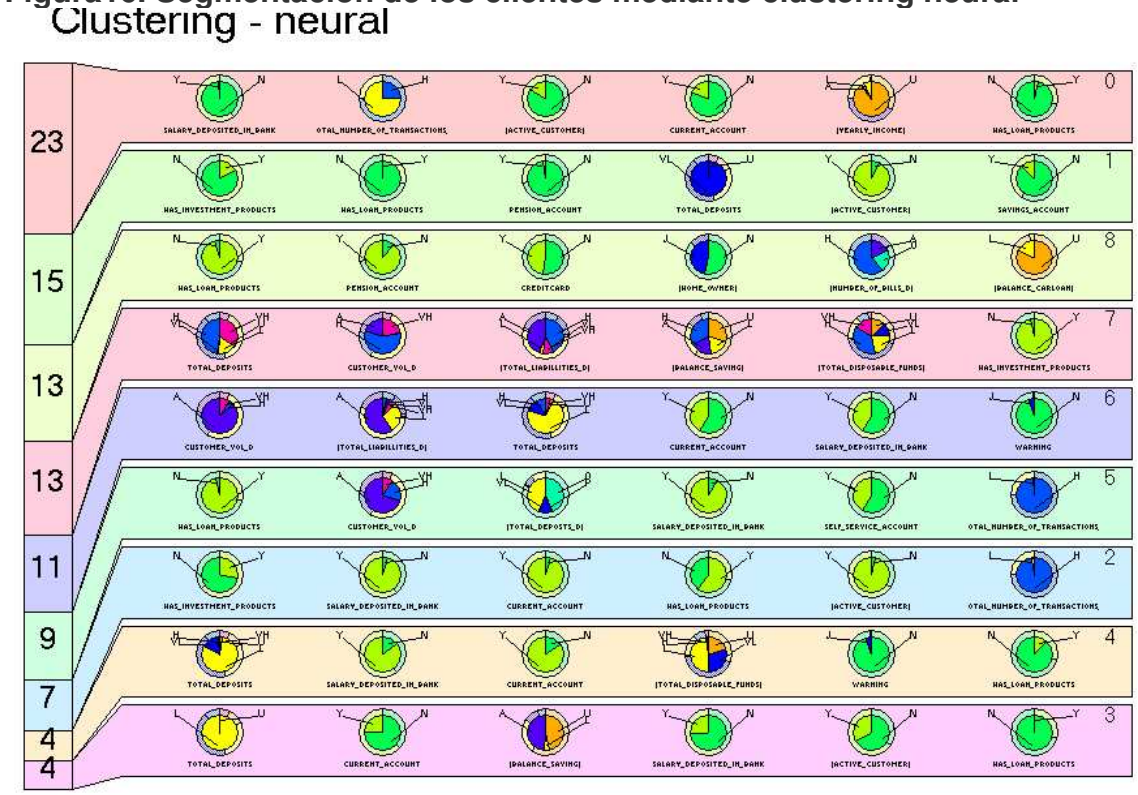


Fuente: [SLB+01]

Los clusters obtenidos se muestran separados por franjas horizontales y ordenadas por tamaño, el tamaño del cluster esta indicado en la parte izquierda y las variables en cada franja están ordenadas de izquierda a derecha según la importancia que tengan en el cluster.

Después de muchas iteraciones y de haber obtenido un número de clusters apropiados para trabajar, se procedió a realizar la comparación con el clustering neural, el cual requirió que se especificaran el número de clusters esperados; como se había mencionado anteriormente. Como el número determinado de clusters que se obtuvo con el clustering demográfico fue de 9, este fue el valor que se utilizo para ejecutar la técnica de clustering neural y posteriormente hacer la comparación. Los resultados obtenidos fueron los siguientes. [SLB+01]

Figura16. Segmentación de los clientes mediante clustering neural



Fuente: [SLB+01]

Este resultado fue en muchas formas muy similar al obtenido con la técnica del clustering demográfico, aunque no exactamente el mismo. Estos dos resultados son validos y mas allá de los análisis esas diferencias radican en que hay clientes que no encajan fácilmente en ninguno de los nueve clusters. [SLB+01]

2.3.6. Despliegue y utilización de los resultados de minería de datos:

Antes de que se entrara a hacer uso de los resultados obtenidos, se realizo una evaluación de estos. El criterio que se siguió fue observar si los resultados si tenían sentido desde el punto de vista del negocio, la forma en que se hizo esto

fue examinar cada segmento y darles un nombre. Los nombres asignados a los clusters en orden según el tamaño fueron los siguientes: pasivo/inactivo, modesto, súper estrella, gris dorado, muy jóvenes - activos, activos, solicitantes de prestamos, inversionista puro y jóvenes estrellas. [SLB+01]
Después de hacer esta valuación se procedió a sacar el verdadero significado que tienen estos resultados para el negocio.

La forma en como se asignaron los nombres a los clusters fue observando la distribución de las variables para los clientes dentro de cada cluster y luego se comparo con el total de la población. El enfoque se dio principalmente en aquellas variables que tenían la mayor discrepancia con la distribución de esas mismas variables, pero en toda la población. [SLB+01]

A continuación se mostrara la interpretación realizada a algunos de los clusters.

Clientes pasivos/inactivos:

Por lo general este tipo de clientes son aquellos que solo tiene unos pocos productos con el banco y que no hacen muchas interacciones con la compañía, o que simplemente están inactivos. [SLB+01]

Las características interesantes de este segmento son las siguientes:

- Solo una pequeña parte de los clientes tienen depositado su salario en el banco.
- Los clientes tienen pocos productos, por lo general uno solo.
- Los productos de los que son propietarios son: pensión, inversiones y productos jóvenes.
- Los clientes tienen muy poquitas transacciones.

- Una parte relativamente grande de los clientes es que son clientes jóvenes o muy jóvenes.
- Una parte relativamente grande de los clientes son nuevos clientes.

Los clientes que no tienen depositado su salario en el banco es porque tienen una relación mas cercana con otro banco que le provee estos servicios financieros. Combinado con el hecho de que son clientes que solo tienen un producto sea de pensión, inversión, producto joven o un préstamo de construcción, esto indica que este tipo de clientes: son muy jóvenes, son compradores, son nuevos clientes o se han ido del banco. Los clientes que abandonan el banco por lo general se quedan con alguno de los productos que se mencionaban anteriormente ya que son productos duros o indeseables de mover. Dirigir estrategias y productos hacia este tipo de usuarios puede resultarle al banco más barato que conseguir nuevos clientes. [SLB+01]

Este segmento es típicamente muy diverso y dinámico, porque contiene personas con un comportamiento consistente con un número diferente de etapas en la relación con el banco. [SLB+01]

Clientes gris dorados:

Este cluster contiene personas de la tercera edad, que no tienen préstamos, pero si grandes depósitos en fondos mutualistas, en productos de ahorro o inversión. [SLB+01]

Las características importantes de este segmento son:

- Los clientes tienen muy pocos productos de préstamo.
- Los clientes tienen relativamente grandes cantidades depositadas.

- Los productos típicos de este cluster son: fondos mutualistas, productos de inversión y de ahorro.
- Los clientes hacen un gran número de transacciones.
- Las transacciones son hechas por los canales tradicionales. (cajeros y terminales bancarias. Lo contrario a banco Web, call center y kioscos).
- Los clientes tienen un número relativamente bajo de pagos automáticos.
- Los clientes tiene varias posesiones.

Estos clientes son muy lucrativos, representan un muy bajo riesgo para el banco por sus relativamente grandes cantidades depositadas. Para estos clientes las estrategias de negocio es reducir frecuentemente el uso de aquellos canales manuales costosos por el desarrollo de servicios alternativos. Una estrategia agresiva de precios no es muy crucial, ya que por lo general son clientes muy leales. [SLB+01]

Resumen de los clusters

Para el análisis de los clusters restante se mostrara un resumen con la visión general de cada segmento.

Tabla 25. Visión general de los clusters.

Segmento	Nombre del segmento	Tamaño relativo	Parte del beneficio
0	Inversionista puro	4.55	9
1	Pasivo/inactivo	33.45	2
2	Estrellas jóvenes	3.74	8
3	Activos	6.24	9
4	Modestos	20.68	8
5	Muy jóvenes – activos	6.78	7
6	Súper estrellas	12.97	31
7	Gris dorado	9.60	17

8	Solicitantes de prestamos	5.53	9

Fuente: [SLB+01]

También se mostrara la penetración de cada producto en cada cluster.

Tabla 26. Penetración de los productos por clusters y en la población global.

Nombre del segmento	Cuenta corriente	Productos de ahorro	Productos de préstamo	Productos de inversión	Productos de pensión	Fondos mutualistas	Tarjetas de crédito	Productos para jóvenes
Inversionista puro	12	2	1	99	8	90	2	1
Pasivo/inactivo	5	33	3	58	49	3	0	17
Estrellas jóvenes	88	78	73	58	7	45	45	22
Activos	94	18	42	100	99	32	16	4
Modestos	90	2	52	0	1	0	2	20
Muy jóvenes – activos	66	95	8	2	2	1	2	61
Súper estrellas	94	78	44	94	92	67	93	3
Gris dorado	97	95	19	99	69	90	1	0
Solicitantes de prestamos	90	5	79	8	4	3	99	14
Global	62	40	32	54	44	24	18	16

Fuente: [SLB+01]

Cada número del cuadro representa el porcentaje de personas que tienen determinado producto en el cluster. El último renglón representa el porcentaje de personas que tienen el producto pero para toda la población. Este tipo de información sirvió de herramienta para analizar las distintas estrategias a seguir en los distintos cluster. Por ejemplo mirar que oportunidades de Cross selling aplicar en que sectores, o saber que la venta de tarjetas de crédito en el segmento “gris dorado” probablemente no sea una buena estrategia. [SLB+01]

Tradicionalmente los jóvenes han sido vistos como un grupo homogéneo, pero en realidad este es un grupo muy diverso. Según las graficas mostradas anteriormente, los clientes jóvenes están distribuidos en todos los segmentos, lo que permitirá diseñar estrategias específicas de negocio para este importante grupo de clientes y asegurarse de que en el futuro se conviertan en buenos clientes para el banco. [SLB+01]

Para el caso de los cluster 0, 2, 4 y 6, se desarrollaron estrategias de retención de los clientes, pues son los cluster que cuenta con una mayor parte de beneficio según se vio anteriormente. [SLB+01]

Para el cluster de clientes activos se aplicaron estrategias de Cross- y up- selling de tarjetas de crédito. Ya que es un cluster que se caracteriza porque no usan mucho este tipo de tarjetas. Para los segmentos de clientes que poseen muchos productos también se aplicaron estrategias de Cross selling para los pocos productos que no poseen. [SLB+01]

Para los cluster más especializados, como los inversores puros o los solicitantes de préstamos la estrategia a seguir fue comercializar productos que se adapten al perfil de comportamiento de estos y a la promoción de productos que aumente el grado de fidelidad de estos clientes. [SLB+01]

Estas fueron algunas de las estrategias de ventas y mercadeo que se realizaron a partir de la segmentación resultante, y de la información que arrojaron las variables dentro de cada cluster.

Los resultados de este proyecto también provocaron diversas reacciones en los ejecutivos del banco y permitió comprobar varias cosas: [SLB+01]

1. Se tuvo una excelente visualización de los datos lo que permitió tener análisis más significativos.
2. Se refino la segmentación original.

Basados en los resultados de este caso de estudio, muchas oportunidades de minería de datos fueron descubiertas y posteriormente realizadas. Algunos de estos proyectos fueron los siguientes: [SLB+01]

- Varios modelos predictivos para realizar campañas mediante correo directo.
- Trabajos adicionales en la segmentación, usando datos mas detallados sobre el comportamiento de los clientes.
- Uso de algoritmos de asociación dentro de los segmentos encontrados.
- Integración de los resultados de la segmentación, lo que permitió al sistema de administración de la información, a las directivas y a los empleados analizar y tomar decisiones en base a esta nueva información.
- Como el modelo de segmentación fue implementado en el sistema de base de datos, se implementaron aplicaciones como las soluciones personalizadas en la Web, soluciones de flujo de trabajo y soluciones para el manejo de campañas. Además esto permitió que el modelo trabajara en tiempo real.
- Con esta segmentación el banco dio el primer paso para predecir el comportamiento de sus clientes. Después de haber clasificado, ordenado y ubicado sus clientes en grupos afines, el banco utilizo estos resultados para la implementar otro proyecto que le permitiera saber que tipos de productos venderles a que clientes y como.

2.4. ASOCIACIÓN DE PRODUCTOS DE LA CANASTA DE MERCADO PARA ANALIZAR EL COMPORTAMIENTO DE LOS CLIENTES

Este caso se analiza el comportamiento de los clientes de una cadena de supermercados de acuerdo a sus compras.

Se trata la tarea de asociación y se utilizó la herramienta de SAS, llamada Enterprise Miner.

2.4.1. Comprensión del negocio o del tema a tratar:

La cadena de supermercados KENDALL es uno de los negocios de retail más grandes del sur de Italia, cuenta con más de 4000 empleados distribuidos en todos sus supermercados y tiendas, y quiere hacer un análisis de las cestas de mercado de sus clientes.

Una cesta de mercado de un cliente, muestra los hábitos de compra de ese cliente en un instante de tiempo dado. Una lista completa de las compras hechas por un conjunto grande de clientes, provee mucha mas información y describe la parte mas importante de un negocio de retail – cual es la mercancía que los clientes están comprando y cuando. [Bel04]

Cada cliente compra diferentes clases de productos, en diferentes cantidades y en diferentes épocas. El análisis de las cestas de mercado usa información sobre que clientes compran para obtener un análisis mas detallados de quienes son ellos, y porque hacen ciertas compras. Además este análisis también permite conocer que productos tienden a ser comprados juntos y cuales serian los mas sensibles cuando se realizaran promociones. [Bel04]

La idea de esta cadena de supermercados es desarrollar un proyecto de minería de datos para hacer este tipo de análisis, con estas herramientas, las cuales pueden se mas potentes y productivas de las que se tienen actualmente, y así tratar de obtener mejores resultados. A parte de querer conocer solo las características de los clientes y de que productos tienden a venderse en conjunto, también quieren tener ideas sobre como hacer el diseño de las nuevas tiendas que tienen pensado abrir, determinar los productos que mas se venderán y definir el tema de los cupones de promoción, las campañas de promoción, etc. [Giu03]

"El objetivo del análisis es individualizar las más frecuentes combinaciones de productos comprados por lo clientes, que sirvan para identificar patrones." [Giu03]

2.4.2. Comprensión y preparación de los datos de los datos a utilizarse:

El conjunto de datos fue sacado de la base de datos de esta cadena de supermercados, la cual agrupa los datos y transacciones hechas por sus 37 sucursales. En cada tienda las transacciones registradas son aquellas hechas por alguien que tiene una de las tarjetas, que proporciona la cadena para el manejo de fidelidad con los clientes. Cada tarjeta tiene un código que identifica las diferentes características del cliente, incluyendo características personales importantes como el sexo, cumpleaños, cumpleaños del cónyuge, número de hijos, profesión y educación. [Giu03]

La tarjeta permite seguir al analista el comportamiento de compra de su dueño: cuantas veces va al supermercado o a la tienda en un periodo dado, que compra, si sigue las promociones, etc. El propósito aquí es considerar los datos de transacciones con productos, con el ánimo de analizar las asociaciones entre estos productos. Por consiguiente no se considerara la influencia de las variables demográficas sobre las promociones. [Giu03]

El conjunto de datos disponibles esta organizado en una colección de 37 bases de datos transaccionales, una por cada tienda. Para cada tienda, hay una unidad estadística (una fila en la base de datos) que corresponde al código de la tarjeta de fidelidad y aun producto comprado. Para cada código de tarjeta puede haber más de un producto, y en la fila, el mismo código de la tarjeta puede aparecer más de una vez, cada vez correspondiendo a una visita en particular de la tienda. [Giu03]

El periodo considerado consistió de 75 días entre el 2 de enero y el 21 de abril de 2001. [Giu03]

El número total de productos disponibles en las tiendas es de cerca de 5000, sin tener en cuenta las marcas, formato y tipo específico (Ej. Peso, color, tamaño). Los productos están usualmente agrupados en categorías. El número total disponible de categorías en los supermercados es de aproximadamente 493. [Giu03]

Para mas claridad se limito el análisis a solo 20 categorías, que corresponden a aquellas que son las más vendidas. En la tabla que se mostrara a continuación se listaran esas categorías de productos, junto con su frecuencia de ocurrencia y el nombre. [Giu03]

Tabla 27. Categorías de los productos y su frecuencia de ocurrencia.

	Categoría	Frecuencia de ocurrencia
1	Pasta	4541
2	Leche	4428
3	Agua	3452
4	Bizcochos	3371
5	Café	2703
6	Galletas	2680
7	Yogurts	2493
8	Vegetales congelados	2484
9	Atún	2464
10	Cerveza	1970

11	Salsa de tomate	1958
12	Colas	1883
13	Arroz	1822
14	Jugos	1681
15	Galletas saladas	957
16	Aceite	775
17	Pescado congelado	759
18	Helado	445
19	Mozarela	432
20	Carne enlatada	177

Fuente: [Giu03]

Estas categorías son utilizadas para generar una base de datos de transacciones. Esta base de datos presenta para cada tarjeta y cada fecha de compra, una lista de los productos que han sido puestos en la cesta. La base de datos de transacción puede ser adecuadamente, expresada como una matriz de datos, donde cada fila representa una transacción de la persona que posea una tarjeta, las columnas son variables binarias que representan si o no cada producto especifico ha sido comprado al menos una vez en esa transacción. Esta matriz fue llamada base de datos de los propietarios de las tarjetas. El número total de transacciones fue de 46727, que corresponde al número de registros en esta base de datos. [Giu03]

2.4.3. Análisis exploratorio de los datos

Para entender las asociaciones entre los 20 productos o categorías escogidas, se tuvo que considerar la construcción de 190 tablas de contingencia[33] de dos direcciones. La tabla 28 muestra una de esas tablas, la cual fue usada para estudiar la asociación entre los productos helado y colas. [Giu03]

[33] Este tipo de tablas se emplean para registrar y analizar la relación entre dos o más variables, habitualmente de naturaleza cualitativa -nominales u ordinales-.

Tabla 28. Ejemplo de una tabla de contingencia de dos caminos y el cálculo de los cocientes de las probabilidades.

HELADO Frecuencia Porcentaje Porcentaje de columna Porcentaje de fila	COLAS 0	 1	 TOTAL
0	41179 88.13 89.60 98.57	4779 10.23 10.40 96.56	45958 98.35
1	599 1.28 77.89 1.43	170 0.36 22.11 3.44	769 1.65
TOTAL	41778 89.41	4949 10.59	46727 100

	Valor	Limites de confianza al 95%	
Cocientes de las probabilidades[34]	2.4455	2.0571	2.9071

Fuente: [Giu03]

En cada celda de la tabla de contingencia, se tiene la frecuencia absoluta, la frecuencia relativa (como un porcentaje) y la frecuencia condicional por fila y por columna. Debajo de la tabla se reporta la medida de asociación, los cocientes de probabilidades entre las dos variables, junto con el correspondiente intervalo de confianza. Una asociación es juzgada como significativa si el valor 1 es externo al intervalo de confianza. De acuerdo a los resultados mostrados en la tabla se puede decir que hay una fuerte asociación positiva entre los dos productos. Hay que recordar que el tamaño de la muestra es absolutamente grande (46727 transacciones), por consiguiente aun un cociente de probabilidad pequeño puede ser significante. [Giu03]

La tabla 29 muestra el resumen de los cocientes de probabilidades más grandes de los 190 posibles.

[34] También conocidos en ingles como Odds Ratio. Es la razón entre la probabilidad de que un evento suceda y la probabilidad de que no suceda

Tabla 29. Cocientes de probabilidad más grandes entre parejas de productos y el correspondiente intervalo de confianza

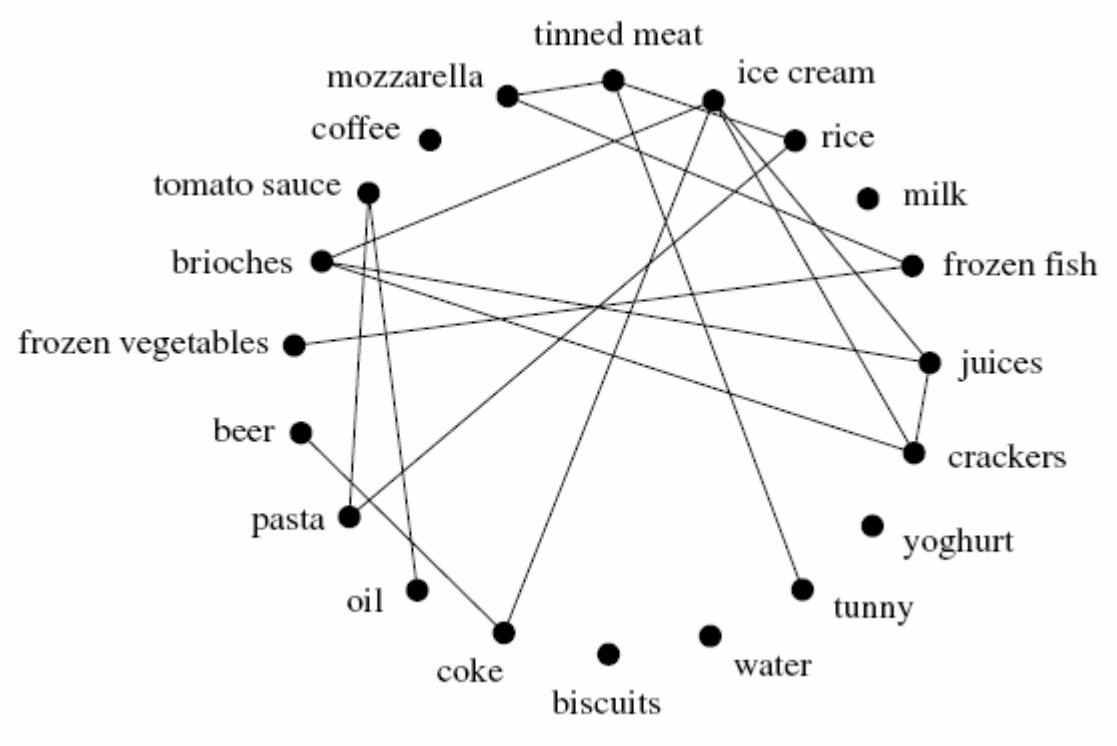
Producto 1	Producto 2	Cociente de probabilidad	Intervalo de confianza	
Carne enlatada	Atún	5.0681	3.9101	6.5689
Carne enlatada	Mozarela	4.8847	2.9682	8.0386
Vegetales congelados	Pescado congelado	3.3610	2.9521	3.8265
Colas	Cerveza	2.8121	2.6109	3.0289
Galletas	Jugos	2.8094	2.6094	3.0248
Jugos	Helado	2.5333	2.1018	3.0534
Colas	Helado	2.4455	2.0571	2.9071
Salsa de tomate	Pasta	2.3773	2.2446	2.5179
Galletas saladas	Helado	2.2839	1.7061	3.0574
Galletas	Galletas saladas	2.2833	2.0276	2.5713
Carne enlatada	Arroz	2.1433	1.4762	3.1120
Arroz	Pasta	2.1129	1.9618	2.2756
Galletas	Helado	2.0211	1.7178	2.3781
Galletas saladas	Jugos	2.0486	1.7633	2.3800
Pescado congelado	Mozarela	2.0785	1.4721	2.9347
Aceite	Salsa de tomate	2.0713	1.8318	2.3420

Fuente: [Giu03]

Las asociaciones más grandes fueron detectadas entre los productos carne enlatada y atún, carne enlatada y mozarela, pescado congelado y vegetales congelados. En todos estos casos se observa que la pareja de productos son considerados como comida rápida. Luego sigue la relación entre dos bebidas: colas y cerveza. En general las asociaciones que se muestran en la tabla 3 parecen razonables desde el punto de vista del tema. En el cálculo de los cocientes de probabilidad, cada par de variables son consideradas independientes de las 18 restantes. Es posible relacionarlas dibujando un grafo cuyos nodos son las variables binarias del producto. Una arista es dibujada entre un par de nodos si el correspondiente cociente de probabilidad es significativamente diferente de 1, en otras palabras si el intervalo de confianza no contiene el valor 1. Dibujar este tipo de grafos puede ser muy útil en la fase exploratoria de los datos. [Giu03]

Es difícil visualizar un grafo con al menos 190 relaciones. Por lo tanto se mostrara un grafo que solo presenta las asociaciones positivas con un cociente de probabilidad de mayor que 2. Esto reduce el número de enlaces en el grafo. [Giu03]. La siguiente figura es ejemplo de ello.

Figura 17. Grafo que muestra las asociaciones positivas más fuertes entre productos.



Fuente: [Giu03]

Nótese que los productos asilados (leche, biscochos, agua, café y yogurt) de los otros, no están asociados positivamente con ninguno. Los otros productos están relacionados directa e indirectamente. Es posible individualizar al menos 3 grupos mediante las conexiones de los enlaces, pero también por las relaciones lógicas entre productos. Estos grupos son muy interesantes porque identifican medianamente el comportamiento de compra de los clientes. Hay un grupo con 5 nodos: atún, carne enlatada, mozarela, vegetales congelados y pescado congelado. Estos nodos relacionados altamente con otros, corresponden a los

productos de comida rápida: fácil y rápida de preparar. Un segundo grupo contiene 4 nodos: arroz, pasta, salsa de tomate y aceite. Este grupo puede ser identificado como comida comprada para platos comunes (comunes en los estándares de la región del mediterráneo). Un tercer grupo contiene otros 6 productos: cerveza, colas, jugos, helados, galletas y galletas saladas. Todos se relacionan con los productos que se consumen fuera de las comidas regulares. Este grupo parecer ser el menos homogéneo de los 3. [Giu03]

Hasta el momento no se ha detectado ninguna asociación negativa. Esto tiene importantes implicaciones, por ejemplo una promoción en pasta presumiblemente incrementara las ventas de este producto, pero es poco probable que decrementen las ventas de otros productos como arroz o agua. Las asociaciones negativas son raramente consideras en un análisis de la cestas del mercado. [Giu03]

2.4.4. Modelado y evaluación:

Modelos log-lineales

Lo modelos log-lineales son muy usados en la minería de datos descriptiva, estos investigan la asociación entre las variables consideradas. Ajustar un modelo log-lineal para las 20 variables binarias puede requerir muchos parámetros para ser estimado. Además el grafo de independencia condicional puede ser difícil de interpretar. [Giu03]

La figura 17 sugiere la existencia de 5 nodos asilados que pueden ser estimados independientes de los otros: leche, biscochos, agua, café y yogurt. Por lo cual se trato de adecuar un modelo grafico log-lineal a las restantes 15 variables, con el propósito de chequear si los resultados del análisis exploratorio pueden ser

confirmados. Se construyo una tabla[35] que presenta la máxima probabilidad estimada de los parámetros del modelo log-lineal, con estimaciones hasta el segundo orden, medida en una tabla de contingencia de 15 direcciones correspondientes a las 15 variables consideradas. [Giu03]

De esa tabla resultante, se vio que todas las interacciones mostradas en la tabla 29, siguen siendo fuertemente significativos, excepto por las relaciones arroz y pasta, galletas y helado, galletas saladas y jugos; que tienen un cociente de probabilidad estimado levemente mas bajo de 2. Además hay otras 14 asociaciones positivas (fuertes): 9 de ellas son de carne enlatada con: colas, galletas saladas, jugos, aceites, salsas de tomate, cerveza, vegetales congelados, pescado congelado y helado; otras 3 de ellas son de helado con: vegetales congelados, arroz, pescado congelado; las dos restantes son de: arroz con: mozarela y galletas saladas. La diferencia de esta tabla que presenta las probabilidades máximas estimadas de los parámetros log-lineales con la tabla 29 es que se tomaron en cuenta las dependencias condicionales entre las variables y además se encontraron interacciones que han sido más significativas. [Giu03]

Reglas de asociación

La forma mas común de hacer un análisis de los datos de las cestas de mercado es usar reglas de asociación. Se considero empezar con un escenario simple. Considérense los productos helado y colas. El orden no es relevante al estudiar la asociación entre 2 productos. Las reglas de asociación se fueron construyendo de la siguiente forma. Se saca el valor de soporte[36] de la regla “si helado, luego colas”. [Giu03]

[35] No mostramos esta tabla, pero se puede consultar en la fuente [Giu03].
[36] El soporte (support) de una regla es la frecuencia relativa que indica la proporción de transacciones en donde la regla es observada. Es obtenido dividiendo el número de transacciones donde se satisface la regla, por el del número de transacciones. Es una medida simétrica.

Soporte (helado → colas) = 170 / 46727 = 0.0036

Este valor indica un bajo soporte para la regla. Esto significa que estos dos productos son compradas juntas ocasionalmente. El soporte corresponde a uno de las 4 frecuencias que se muestran en la tabla de contingencia (tabla 28). Un soporte de 0.036 indica que solo el 0.36% de las transacciones consideradas tenían ambos productos en la cesta. [Giu03]

La confianza[37] de una regla, aun cuando es calculada para una asociación, donde el orden no importa, depende del cuerpo y la cabeza de la regla (medida asimétrica):

Confianza (helado → colas) = 170 / 769 = 0.22.

El cual corresponde a la frecuencia condicional de la segunda fila de la tabla 28 donde colas = 1. Y

Confianza (colas → helado) = 170 / 4949 = 0.034.

El cual corresponde a la frecuencia condicional de la segunda columna de la tabla 28 donde helado = 1. La frecuencia en el primer caso, indica la proporción, de aquellos que también compran helado, de los que también compran colas; en el segundo caso, indica la proporción, de aquellos que también compran colas, de aquellos que también compran colas. [Giu03]

La elevación[38] es una medida normalizada de una regla, de gran interés. Toma la confianza de una regla y la relaciona con el soporte de la cabeza de la regla. [Giu03]

[37] La confianza (confidence) de una regla (A→ B) expresa una frecuencia relativa, que indica la frecuencia (o probabilidad) de ocurrencia de B, condicionado a que A sea verdadero. Es una medida asimétrica

Elevación (helado → colas) = 0.22 / 0.11 = 2
Elevación (colas → helados) = 0.034 / 0.017 = 2

Ya se analizaron los resultados que arroja una sola regla, pero no se pueden sacar conclusiones certeras de estos, hay que conocer los resultados que ofrecen las demás reglas, para poder hacer un análisis mas completo y sacar verdaderas conclusiones que permitan conocer mejor los hábitos de los consumidores. Para descubrir esto y obtener una descripción más comprensiva de las reglas de asociación, la cadena hizo el modelo con todas las reglas de asociación posibles. [Giu03]

Los resultados de las reglas más relevantes fueron organizados en distintas tablas[39] según el criterio con el cual se estuviera midiendo la relevancia de las reglas.
Una primera tabla presenta en segundo orden, reglas de asociación con el valor de soporte más grande de las 190 reglas posibles. Para cada regla de esta tabla se muestra le medida de elevación, el soporte, y la confianza, así como también la frecuencia de ocurrencia, que es la frecuencia absoluta de la regla. [Giu03]

Según los resultados que se obtuvieron, se vio que la regla (helado, colas) no aparece entre las más frecuentes. Otras reglas tienen una medida de soporte más grande y la regla con el soporte mas grande es leche → pasta, que aparece en aproximadamente el 50% de las transacciones. Esta es seguida por biscochos → pasta, leche → biscochos, agua → pasta y leche → agua; todas ocurren en aproximadamente el 39% de las transacciones. [Giu03]

---

[38] La medida de elevación (lift) de una regla, es un radio entre la frecuencia relativa de ambos productos ocurriendo conjuntamente, y la frecuencia relativa del mismo evento pero asumiendo que los dos productos son independientes.
[39] El resultado de estas tablas no se muestra totalmente, solo mostraremos un subconjunto de estos, donde se muestran los más significativos. Para ver la tabla completa se puede ir a la fuente [Giu03]

Tabla 30. Cocientes de probabilidad más grandes entre parejas de productos y el correspondiente intervalo de confianza

	Relaciones	Elevación	Soporte (%)	Confianza (%)	Frecuencia de ocurrencia	Regla
1	2	1.13	49.84	75.86	3359	Leche → pasta
2	2	1.13	49.84	73.97	3359	Pasta → leche
3	2	1.18	39.9	79.77	2689	Bizcochos → pasta
4	2	1.18	39.9	59.22	2689	Pasta → biscochos
5	2	1.21	39.86	60.66	2686	Leche → biscochos
6	2	1.21	39.86	79.68	2686	Biscochos → Leche
7	2	1.15	39.72	77.55	2677	Agua → pasta
8	2	1.15	39.72	58.95	2677	Pasta → agua

Fuente: [Giu03]

Cabe anotar que como la tabla anterior muestra las reglas con la mayor medida de soporte y esta es una medida simétrica, es obvio que la regla reciproca este siempre adyacente a la otra. [Giu03]

La segunda tabla muestra en segundo orden, reglas de asociación con el valor de confianza más grande de las 190 reglas posibles. Esta tabla muestra que por ejemplo la regla arroz → pasta, tiene una confianza igual a 90.18%. Esto significa que si una transacción contiene arroz, esta también tendrá pasta cerca del 90% de las veces. Por otra parte la regla pasta → arroz no se encuentra entre los resultados mas significativos; esta regla tiene una confianza de 36.18%. Esto puede ser interpretado diciendo que si una transacción contiene pasta también tendrá arroz el 36.18% de las veces. [Giu03]

Tabla 31. Reglas con la mayor medida de confianza.

	Relaciones	Elevación	Soporte	Confianza	Frecuencia de	Regla

			(%)	(%)	ocurrencia	
1	2	1.34	24.38	90.18	1643	Arroz → pasta
2	2	1.32	2.34	89.27	158	Carne enlatada → pasta
3	2	1.32	25.86	89.02	1743	Salsa de tomate → pasta
4	2	1.30	9.88	87.75	666	Pescado congelado → pasta
5	2	1.27	9.85	85.68	664	Aceite → pasta
6	2	1.26	5.46	85.19	368	Mozarela → pasta
7	2	1.29	5.61	84.94	378	Helado → leche
8	2	1.29	9.56	84.85	644	Pescado congelado → leche

Fuente: [Giu03]

Una tercera tabla reporta las reglas con los valores más altos de elevación de las 190 posibles reglas. Nótese que las reglas pescado → mozarela y helado → galletas saladas vienen primero, ambas con un valor de elevación de 2.36; también se puede ver que la regla colas → helado esta bien ubicada. [Giu03]

Tabla 32. Reglas con la mayor medida de elevación.

	Relaciones	Elevación	Soporte (%)	Confianza (%)	Frecuencia de ocurrencia	Regla
1	2	2.36	1.71	15.15	115	Pescado congelado → mozarela
2	2	2.36	1.71	26.62	115	Mozarela → pescado congelado
3	2	2.36	2.21	33.48	149	Helado → galletas saladas
4	2	2.36	2.21	15.57	149	Galletas saladas → helado
5	2	2.17	1.62	14.36	109	Pescado congelado → helado
6	2	2.17	1.62	24.49	109	Helado →

						pescado congelado
7	2	1.98	3.65	55.28	246	Helado → colas
8	2	1.98	3.65	13.06	246	Colas → helado

Fuente: [Giu03]

Esta tabla puede ser útil para comparar con la tabla de cocientes de probabilidades (tabla 29), que fue usada para construir el grafo exploratorio. Comparando las dos tablas se vio que productos similares aparecen en ambas, excepto por pasta que aparece solo en la tabla 29. Sin embargo las parejas son en algunos casos diferentes. Los productos aislados que se muestran en la figura 1-leche, biscochos, agua, café y yogurt, tienen una alta medida de soporte pero una baja medida de elevación, por consiguiente no se encuentran en los resultados de las reglas que tienen la mayor medida de elevación. [Giu03]

También se consideraron asociaciones con más de 2 productos. Los resultados que se obtuvieron considerando las reglas con un mayor soporte, revelaron que aparecían algunas de las reglas de la tabla 30, en los primeros lugares según este criterio. En lugares secundarios se veían algunas reglas de orden 3, que tenia por lo general productos que se consumen fuera de las comidas regulares y combinaciones de productos aislados como pasta, leche y biscochos. [Giu03]

Para asociaciones de cuarto orden[40] los valores de confianza fueron más bien altos. Por ejemplo una transacción que contiene salsa de tomate, aceite y galletas saladas, también tendrá pasta; adicionalmente el valor de confianza para esta regla es del 100%. Sin embargo el soporte de esta regla es de solo el 1.41%. [Giu03]

[40] En la fuente bibliografica [Bel04] aparece una tabla que muestra el resumen de los resultados para este tipo de asociaciones.

Otra metodología que se siguió para la construcción de las reglas se baso en árboles de decisión. Se escogió la categoría de pastas, pues es el tipo de productos más frecuentes y la cabeza mas frecuente en las reglas de asociación. Se construyo un árbol donde la variable pasta se utilizo como variable objetivo y los demás productos como los predictores. Entre los caminos que conducen a los nodos terminales, se consideraron aquellos caminos donde todas las variables tenían el valor 1. Estos caminos corresponden a las reglas con una alta confianza. Usando un árbol CHAID se obtuvieron las siguientes reglas: [Giu03]

- Atún y salsa de tomate → pasta
- Salsa de tomate y arroz → pasta
- Arroz y biscochos → pasta

Y sus respectivas medidas de interés son:

Elevación 1.41, confianza 95.24%, soporte 14.84%
Elevación 1.44, confianza 96.80%, soporte 12.14%
Elevación 1.40, confianza 94.23%, soporte 18.43%

Nótese que las tres reglas tienen valores de confianza altos. Esto se esperaba, ya que los modelos basados en árboles tratan de desarrollar las mejores reglas predictivas para la variable objetivo. [Giu03]

Comparación de los modelos

Como la idea de buscar patrones y reglas es muy reciente, hay muy poco consenso en la literatura de minería de datos en como medir sus rendimientos. Una idea es medir la utilidad de los patrones en términos de cuan interesante o inesperados son para el analista. [Giu03]

Las medidas de interés para las reglas también pueden ser usadas para evaluar el rendimiento. En este caso se consideraron las medidas: soporte, confianza y elevación para validar el conjunto de reglas. Pero las necesidades del usuario son las que imperan al decidir cual de estas utilizar para escoger un conjunto de reglas. El soporte se puede usar para evaluar la importancia de una regla en términos de su frecuencia en la base de datos; la confianza se puede usar para investigar las posibles dependencias entre las variables; y la elevación puede ser usada para medir la distancia de la situación de independencia. [Giu03]

Finalmente, un conjunto de reglas tienen que ser evaluadas sobre su capacidad para responder a las necesidades u objetivos del análisis. Aquí los objetivos primarios son reorganizar el diseño de ventas de las terminales y planear promociones para incrementar los ingresos. Una vez las asociaciones han sido identificadas, es posible organizar promociones. Por ejemplo, poniendo un producto en promoción, se incrementan las ventas de los productos asociados. [Giu03]

Los cocientes de probabilidades y los modelos log-lineales también pueden ser empleados para determinar una estructura de asociación global entre las variables de compra, para este caso hay ciertas medidas estadísticas para evaluar la calidad del modelo. Sin embargo estas tienen un propósito diferente. Los árboles de decisión también pueden ser vistos como un modelo global[41] capaz de reproducir una estructura de asociación. Sin embargo las reglas de asociación[42] son mejores para este tipo de casos porque son más fáciles de detectar y de interpretar. Los modelos log-lineales y los árboles pueden ser mejor utilizados cuando se tiene el tiempo y el conocimiento necesario para implementar un modelo global. [Lar05]

[41] Un modelo es una descripción global de o explicación de un conjunto de datos, tomando un alto nivel de perspectiva. Para este caso de estudio los modelos globales son los modelos log-lineales y los árboles de decisión.
[42] Las reglas de asociación son utilizadas para descubrir patrones locales en los datos.

2.4.5. Despliegue y utilización de los resultados de minería de datos:

Con los resultados arrojados por estos modelos, donde se pudo conocer cuales son los tipos de productos más vendidos, aquellos que por lo general se venden juntos, y las características de los clientes. La cadena de supermercados KENDALL emprendió una serie de acciones para mejorar los niveles de ganancia de la compañía, donde participaron desde los gerentes de departamentos hasta los vendedores. Algunas de las acciones emprendidas fueron las siguientes:

- Se mejoro el diseño de la tienda virtual. Se desarrollo de forma que cuando un cliente esta utilizando este servicio, se le mostrara una lista de productos en los que posiblemente este interesado. [SBA+01]

  Para esto la cadena puso en práctica otro proyecto que recomendara automáticamente a los clientes que estuvieran comprando, usando un sistema de compras remoto. En este caso la cadena animaba a los clientes a usar asistentes personales digitales (PDAs) para crear y transmitir sus órdenes a la tienda, la cual se encargaba de organizar los pedidos y de repartirlos. [SBA+01]

  Este sistema recomendador suministra una fuente alternativa de nuevas ideas para los clientes, que visitan las tiendas con menos frecuencia. Estas recomendaciones fueren generadas usando las técnicas de minería de datos descritas anteriormente, emparejando los productos con los clientes, basados en las solicitudes esperadas del producto y en los gastos previos del cliente. [SBA+01]

  Una prueba de este sistema fue hecho usando clientes de 2 tiendas, cada una con más o menos 1000 clientes. Las recomendaciones que hacia el

sistema solo las hacia con las categorías de productos que se trabajaron. [SBA+01]

La prueba fue conducida en 2 fases. Durante la primera fase de la prueba, usando solo una de las tiendas y una versión primaria del sistema, un total de 97 órdenes completas fueron procesadas de usuarios que usaban una PDA. De esas, 6 ordenes (6.1%) contenían al menos un producto escogido para la lista de recomendación. Es importante anotar que la lista de recomendada, contiene productos que no han sido comprados por el cliente previamente). El objetivo de este recomendador de productos es incrementar el ingreso comparable a las compras espontáneas que un comprador puede hacer mientras camina a través de la tienda o después de recibir algún tipo de propaganda por correo. Por esta medida los resultados del recomendador inicial fueron algo decepcionantes: el incremento del ingreso fue 0.3% superior al ingreso generado por la compra de productos que un cliente hace personalmente de acuerdo a su lista de compras. [SBA+01]

Cuando el sistema fue evolucionando, el equipo de desarrollo noto que las distribución del gasto en las diferentes categorías de productos fue diferente de los productos comprados de las lista de recomendada en comparación de la lista de compras que cada cliente mantiene. Estos resultados se interpretaron y dan a entender, que hay un conjunto de categorías en donde las recomendaciones son más bienvenidas que otras, y una serie de entrevistas con algunos clientes corroboro esta interpretación. Los clientes quieren recomendaciones más interesantes. Por lo cual este sistema se enfatiza en las categorías que han sido consideras como las más atrayentes para los clientes. La meta fue crear un conjunto de productos que fueran los más bienvenidos. Otra de las fuentes para

incluir otros productos en la lista de recomendaciones, fue utilizar los nuevos productos que ingresaban a los supermercados. [SBA+01]

En la segunda fase del estudio, el nuevo sistema recomendador fue utilizado esta vez en las dos tiendas, y un nuevo conjunto de clientes fueron adicionados a la prueba. Para la primera tienda, la fracción de ordenes que contenía al menos un producto recomendado incremento del 6.1 al 7.7% y el incremento del ingreso incremento del 0.3 al 0.5%. Para la segunda tienda sin embargo, la respuesta fue mucho mayor, y el 25% de las ordenes incluían al menos una recomendación, con un margen de ganancias de 1.8%, lo que se considera un numero respetable, dado los apretados márgenes de beneficio del supermercado. [SBA+01]

Los resultados obtenidos con estas pruebas, muestra claramente que este sistema de recomendaciones puede dar una muy buena impresión del negocio, y a parte, puede influir positivamente en los márgenes de ganancia.

- Se sacaron nuevas promociones y descuentos. Cambiando un cupón para que se incluya un segundo producto que un cliente compraría, de esta forma se incrementara las ventas sin costos promociónales adicionales.

- Mejores ubicaciones de los productos en las plantas físicas. Se ubicaron los productos que tenían una fuerte relación de compra mas cercanamente, para tomar ventajas de la correlación natural entre los productos. [Giu03]

- Se implementaron estrategias de ventas cruzadas. Se les ofrecía a los clientes de acuerdo a sus compras, otros productos en los que podrían estar interesados.

2.5. MANEJO DE LAS RELACIONES CON EL CLIENTE (CRM) UTILIZANDO TECNICAS DE MINERIA DE DATOS.

En este caso se estudia el comportamiento de compras de los clientes de una compañía que vende mercancía por correo.

En términos de minería de datos este es un problema de clasificación predictiva.

2.5.1. Comprensión del negocio o del tema a tratar:

La compañía XYZ la cual vende mercancía por correo en Italia, quiere emprender un plan de mejora sobre su sistema CRM. “En este caso de estudio se observan los aspectos estadísticos del sistema de manejo de las relaciones con el cliente”. [Giu03]

En un contexto como este, las compañías tienen como primer objetivo impulsar la fidelidad de los clientes, para obtener de ellos tanto valor como sea posible. La necesidad de tener la fidelidad de los clientes motiva a las compañías a conocerlos bien. Una forma de hacer esto es administrando y procesando la base de datos de los clientes. Los métodos de minería de datos representan un acercamiento valido para la extracción información valiosa e importante de la base de datos, y luego usar esta información para manejar las relaciones con los clientes existentes y los potenciales. Las compañías personalizan cada vez más sus servicios para satisfacer a cada cliente. [Giu03]

“El objetivo es estudiar el comportamiento de compra de los clientes de la compañía, y en particular, entender desde el principio, que factores pueden crear un comprador ocasional o un comprador leal. Esto puede indicar que clientes

serán realmente rentables y donde concentrar cualquier esfuerzo de marketing." [Giu03]

2.5.2. Comprensión y preparación de los datos de los datos a utilizarse:

Los datos que se consideraron para comenzar el proyecto de minería; fue la información de referencia de la población, y los clientes actuales de la compañía. Esta información esta distribuida a través de 3 bases de datos diferentes, que contienen la lista de los clientes y sus características, la lista de órdenes recolectadas por las agencias locales, luego transmitidas a la compañía, y la lista de las órdenes de compra transmitidas por las agencias a la compañía. Las 3 bases de datos contienen variables de los clientes, principalmente socio-demográficas y variables de comportamiento. Estas variables se refieren a las maneras en las cuales se ha establecido el primer contacto comercial. (Ej. El numero de productos comprados y el método de pagos) [Giu03]

Para lograr conseguir las metas de este análisis, es conveniente analizar un grupo homogéneo de consumidores. Esto es, analizar el comportamiento en un cierto plazo de tiempo de personas cuyo primer contacto ocurrió casi al mismo tiempo. Esto elimina efectos perjudiciales, debido a los cambios estructurales en la economía, o en la estructura de la compañía. [Giu03]

Primero se consideraron los clientes que habían ingresado en la base de datos entre 1992 y 1996, el número total era demasiado grande, igual a 210 085. Por consiguiente era demasiado costoso en tiempo y en dinero analizar todo este conjunto de datos, por lo cual se tomo una muestra estratificada y se analizo ésta. Se tomaron el mismo número de clientes de cada espacio de tiempo sobre el periodo completo de tiempo, esta muestra contenía un total de 2470 clientes. [Giu03]

Finalmente, como los datos estaban esparcidos en las tres bases de datos, se procedió a construir una base de datos, donde se organizo toda la información que se requería. Después de un largo proceso de administración de la base de datos se obtuvieron las siguientes variables. [Giu03]

Tabla 33. Variables a utilizar.

Estado del mercadeo.	Dimensión de la compra.
Si el cliente esta activo.	Edad.
Si el cliente esta tiene deudas	Área de residencia
Numero total de pedidos.	Sexo
Fecha del primer pedido	Si el primer pago esta dentro del plazo
Fecha del ultimo pedido	Primer cantidad gastada.
Monto total de los pedidos	Numero de productos en el primer pedido.
Monto total pagado	
Balance actual	
Si los pagos han estado atrasados	
Tiempo de espera entre el primer y segundo pedido.	
Cantidad actual de los plazos	
Numero residual de plazos.	

Fuente: [Giu03]

Análisis exploratorio de los datos

Antes de comenzar a hacer el análisis de los datos, se necesita identificar la variable respuesta, definir las variables predictivas y las posibles transformaciones que se les tiene que hacer. El objetivo principal es clasificar a los clientes en 2 categorías: aquellos que solamente hacen un pedido y aquellos harán más pedidos. Esta variable binaria se puede nombrar Y, y puede ser deducida de la variable “Numero total de pedidos”, la cual se encuentra especificada en la tabla 33. Se fijara Y = 0; si el numero de pedidos es igual a uno y Y=1; si el numero de pedidos es mayor a uno. Por consiguiente un cliente se percibe como leal, si ha hecho al menos 2 pedidos. La distribución de la variable respuesta se puede ver en la siguiente tabla.

Tabla 34. Distribución de la variable respuesta.

Modalidad	Frecuencia absoluta	Frecuencia relativa (%)
Y = 0	1457	59.71
Y = 1	1013	40.29

Fuente: [Giu03]

Se consideraron después las variables predictoras. Aparentemente parece ser importante considerar variables que describen como fue establecido el primer contacto establecido con la compañía, así como las variables socio-demográficas disponibles de los clientes como: edad, sexo, área de residencia y la dimensión de la correspondiente agencia. [Giu03]

Las siguientes tablas muestra la distribución condicional de la variable respuesta en las variables socio-demográficas.

Tabla 35. Distribución condicional de la variable respuesta en las variables predictoras sociodemográficas.

Sexo	Y = 0	Y = 1
Masculino	61.04%	38.96%
Femenino	57.88%	42.12%

Área	Y = 0	Y = 1
Norte	55.40%	44.60%
Centro	58.22%	41.78%
Sur	62.73%	37.27%

Edad	Y = 0	Y = 1
15 – 35	68.80%	31.20%
36 – 50	53.44%	46.56%
51 – 89	60.42·	39.58%

Dimensión	Y = 0	Y = 1
Pequeña	60.39%	39.61%
Mediana	56.95%	43.05%
grande	62.11%	37.89%

Fuente: [Giu03]

De estas tablas se pudieron sacar las siguientes conclusiones: [Giu03]

- Sexo: a primera vista, se ve que no es una variable que influencie mucho la variable respuesta, y no hay una diferencia substancial en la distribución de los hombres y la de las mujeres.
- Área de residencia: se puede ver como la probabilidad condicional de Y = 1 decrece cuando el área cambia de norte a centro y del centro al sur. Esta parece ser una variable predictiva.

- Edad: esta variable puede ser un predictor relevante, con la probabilidad de Y = 1 aumenta sensiblemente con la edad.
- Dimensión de la agencia: esta representa la única información que se puede usar para reconstruir la locación de la agencia, una variable importante que no se tiene. La dimensión de la agencia subdivide las agencias en tres clases, según la base del numero de clientes a los que se les presta el servicio: si el numero es menor que 15, es considerada pequeña, si el numero esta entre 15 y 30, es considerada mediana; y si el mayor de 30 (hasta un máximo de 60), es considerada grande. Según la tabla 35 se puede ver que la probabilidad condicional de Y = 1 es mas baja para las agencias mas grandes. Las agencias medias parecen mostrar la más alta probabilidad condicional.

Además de las variables socio-demográficas; también se tienen las variables de comportamiento que refieren a las primeras órdenes de pedido de los clientes: [Giu03]

- Plazo: una variable binaria que indica si la primer compra es pagada en plazos (1) o no (0). Este indica la duración de la relación entre el cliente y la compañía. Si una persona paga en plazos, el contacto con la compañía tendera a ser mas largo.
- Primer cantidad gastada y número de productos en el primer pedido: estas dos variables cuantitativas parecen ser particularmente informativas sobre el comportamiento del cliente en el primer contacto comercial.

Estas dos variables cuantitativas no serán transformadas. Para ayudar a la interpretación se van a transformar a binarias las variables cualitativas edad, área y dimensión de la agencia, todas con tres niveles. Lo que da un total de de 9 variables binarias.

2.5.3. Modelado y evaluación:

Los siguientes fueron los tipos de algoritmos utilizados para tratar de conseguir el mejor modelo.

Modelos de regresión logística

Habiendo seleccionado las variables predictoras; se necesita encontrar las que pueden predecir con mas efectividad la variable respuesta. El primer modelo es la regresión logística. Con el propósito de escoger un modelo, se siguió paso a paso el procedimiento, basado en la diferencia de desviación G^2, con un nivel de significancia de 0.05. La siguiente tabla muestra los resultados obtenidos:

Tabla 36. Modelo de regresión logística seleccionado.

	Estimados	Stderr	Wald	Pr>Chi-square	Cocientes de probabilidad
Intercepto	0.3028	0.1248	108.93	<0.0001	-
Edad 15-35	-0.5440	0.1367	15.84	<0.0001	0.580
Plazos	1.6107	0.1371	137.98	<0.0001	5.006
Numero de productos	0.3043	0.0465	42.78	<0.0001	1.356

Fuente: [Giu03]

Se obtuvo un modelo con 3 variables que afectan significativamente a Y: la variable binaria plazos, con una fuerte asociación positiva medida por un cociente de probabilidad de aproximadamente 5; la variable edad, o mas precisamente, la clase mas joven de esta variable, con un efecto negativo determinado por un cociente de probabilidad de aproximadamente 0.580; y la variable numero de productos con una asociación positiva un poco suave, expresado por un cociente de probabilidad de aproximadamente 1.356. Como esta variable es discreta, el efecto debe ser interpretado diciendo que un incremento unitario en el numero de

productos determina un incremento en la probabilidad de que Y = 1 de aproximadamente 1.356. La regla discriminante logística para este caso de estudio permitió distinguir los clientes a priori que son rentables (y = 1), de aquellos que son menos rentables, por consiguiente se pueden idear diferentes formas de apuntar hacia los clientes. [Giu03]

Para este modelo lo que se necesita saber cuando un usuario hace un primer pedido, son las 3 siguientes cosas: si son jóvenes (variable A), si pagan en plazos (Variable B) y cuantos productos van a pedir (Variable C). Donde t_a, t_b, t_c. Son los parámetros estimados de estas tres variables cuyos valores se muestran en la tabla 36; y donde t es el parámetro estimado del intercepto. Un cliente será rentable, si la probabilidad estimada de hacer pedidos más de una vez es mayor que 0.5 y esto corresponde a chequear si la desigualdad $t + t_a*A + t_b*B + t_c*C > 0$ es verdadera. Por ejemplo si un cliente no es joven, paga en plazos y compra un producto, el provecho es $(-1.3028 \times 1) + (1.6107 \times 0) + (0.3043 \times 1) = 0.6122 > 0$. Si un cliente no es joven, no paga en plazos y compra solo un producto, entonces es probablemente menos rentable; $-1.3028 + 1.6107 + 0.3043 = -0.9985$. [Giu03] "El modelo de regresión logística puede por consiguiente suministrar un simple mecanismo de puntuación para cada cliente, lo que puede ser usado para tomar decisiones." [Giu03]

Redes de función de base radial

Para un modelo de red neuronal, se escogió la función de base radial (RBF) con un nodo escondido, esto es, porque puede haber una estructura de vecindario en el espacio de la variable de entrada. Se consideraron 13 variables predictoras: plazo, primer cantidad, gastada, numero de productos, centro, edad51_89, edad36_50, dimensión grande, dimensión pequeña, edad 15_35, sexo, norte, sur e islas, dimensión pequeña, a parte de la variable respuesta Y. Como una función de la combinación de las variables de entrada, se tomo una función de base radial

gaussiana con igual ancho y con igual altura. La función de activación para el nodo escondido es la función identidad y la función de activación para el nodo de salida es la función softmax[43], así que se obtienen los valores de salida normalizados correspondiente a la probabilidad estimada de Y = 1. [Giu03]

Modelos de árboles de clasificación

Se comenzó comparando dos modelos de árboles CART, basados en la entropía y en la impureza de Gini[44]. El mejor modelo es el basado en la impureza de Gini. Los resultados del mejor árbol están basados en un algoritmo de poda que dirige a un óptimo número de nodos terminales. Esto se hace reduciendo el radio de error en la clasificación.

El árbol de clasificación más optimo que se obtuvo, se describe en término de 11 reglas de asociación[45], apuntando hacia las hojas, que toman a 1465 clientes en el conjunto de datos de entrenamiento y los divide en 11 grupos objetivo, cada uno con una probabilidad estimada diferente de hacer un pedido otra vez (Y = 1). [Giu03]

Cada regla es enunciada según la trayectoria que conduce del nodo de la raíz al nodo terminal. Pero la lista de condiciones que expresa una regla se escribe en orden inverso, así que los nodos más lejos de las hojas, estén más cercanos a estas en la regla. La siguiente regla tiene la medida de soporte más grande, con aproximadamente 48.3% de clientes que siguen esta regla: [Giu03]

IF (375000 ≤ FIRST AMOUNT SPENT < 2659000) AND
(INSTALMENT = 0), THEN (Y =0)

[43] La función softmax permite que las múltiples salidas de la red puedan ser interpretadas como probabilidades a posteriori, de forma que la suma de todas ellas es igual a 1.
[44] La entropía y la impureza de Gini, son criterios que se utilizan para medir la impureza del nodo de un árbol.
[45] Las reglas se pueden encontrar en el capitulo 10 de la fuente de este caso. [Giu03]

Esta regla corresponde a una hoja obtenida de partir todas las observaciones por plazos, y luego por la desigualdad 37500 ≤ Primer cantidad gastada < 2659000. Las personas que cumplen las condiciones de esta regla son estimados como no rentables, pues la probabilidad estimada de Y = 1 es solo de 18.6%. Esto explica porque la cabeza de la regla es Y = 0. En general, la cabeza de la regla obedece la regla clásica discriminante: si la probabilidad cabida es menor al 50% luego Y = 0; de otra forma Y =1. [Giu03]

Los árboles de clasificación proporcionan así una regla discriminante inmediata, basada en la división de las variables predictoras. Para asignar un nuevo cliente, se comienza en la raíz y se toma el camino correspondiente a las características del cliente; luego se ve si la hoja termina da una probabilidad mayor que 50% a Y = 1. La diferencia con la regresión logística es que la regla discriminante es una declaración jerárquica lógica (basada en la división de los datos) mejor que un puntaje aditivo (basado en todos los datos). Las variables que parecen relevantes para la clasificación son pagos, numero de productos y edad que corresponden a las variables significativas en el modelo de regresión logística. El árbol también asigna relevancia a la primera cantidad gastada y al ara geográfica. [Giu03]

Modelo de los vecinos más cercanos

En un modelo de los vecinos más cercanos, el principal parámetro a escoger es el ancho o distancia K; esta establece el tamaño del vecindario de las variables predictoras que se usara para predecir Y. Se inicio con un valor muy grande, K = 732, que corresponde a la mitad del total de observaciones en el conjunto de datos de entrenamiento. Luego se intento con un valor mas bajo, K = 10. Resulta que K = 10, es una mejor opción, en termino del índice de error de clasificación: para K = 732 es 0.41, para K = 100 es 0.328 y para K = 10 es 0.316. [Giu03]

El modelo de los vecinos más cercanos trabaja mejor cuando las observaciones están bien separadas en el espacio de las variables predictoras y cuando los correspondientes grupos son bastante puros. En una situación ideal, las observaciones se deben partir en regiones sin traslapo, posiblemente de un tamaño pequeño, y cada región debe contener observaciones con un valor similar de la variable respuesta (0 o 1). [Giu03]

Comparación de modelos

Primero se compararon los modelos en términos de matrices de confusión obtenidas del conjunto de datos de validación. Se escogió un umbral de corte de 50%, y los errores son obtenidos sobre esa base. La siguiente tabla muestra la matriz de confusión para el modelo final de regresión logística. [Giu03]

Tabla 37. Matriz de confusión para el modelo de regresión logistica.

		Predecido	
		0	1
Observado	0	48.02	10.91
	1	22.92	18.14

Fuente: [Giu03]

En esta tabla y en las siguientes, las frecuencias son expresadas como porcentajes. La tabla muestra que el modelo predice como no rentable (Y = 0 predecido) clientes que de hecho son rentables (Y = 1 observado) en el 22.92% de los casos; este es el error tipo 1. Por otra parte, predice como rentables (Y = 1 predecido) aquellos que no son (Y = 0 observado) en 10.91% de los casos, este es el error tipo 2. [Giu03]

Si el modelo de regresión logística conduce a una regla discriminante valida, depende de las evaluaciones de mercadeo los costos relativos de los dos errores. Usualmente, si un cliente es clasificado como rentable, una campaña directa de correo es dedicada a estos por mail, llamadas telefónicas, etc. Por lo contrario los clientes que no son rentables no se tendrán en cuenta en la campaña. Por consiguiente los costos del error tipo 1 depende de la probabilidad de perder clientes que no han sido clasificados, aunque estos puedan ser rentables; el costo del error tipo 2 es el dinero asignado por la compañía para seguir clientes que probablemente no merecen atención. [Giu03]

La tabla 38 muestra la matriz de confusión para el modelo redes de función de base radial (RBF).

Tabla 38.matriz de confusión para el modelo redes de función de base radial (RBF).

		Predecido	
		0	1
Observado	0	47.34	11.60
	1	20.87	20.19

Fuente: [Giu03]

Nótese que el índice de error de clasificación para el modelo RBF es aproximadamente del 32.47%, mas pequeño que el de la regresión logística, pero mas grande que el de los arboles CART. La probabilidad de los dos errores están desbalanceados, así como los de la regresión logística, por lo cual se pueden sacar las mismas conclusiones que con la regresión. Sin embargo la ligera mejora que da este modelo no justifica el incremento de la complejidad del modelo y la dificultad de interpretación comprado con la regresión logística. [Giu03]

La tabla 39 muestra la matriz de confusión para el modelo de árboles CART.

Tabla 39. Matriz de confusión para el modelo del árbol CART.

		Predecido	
		0	1
Observado	0	43.52	15.42
	1	14.32	26.74

Fuente: [Giu03]

El índice de error de clasificación para el árbol de clasificación es ligeramente mas bajo que para el modelo de regresión logística, 29.74% contra 33.83%. Además, las probabilidades de los dos tipos de errores están mejor balanceados. El modelo del árbol deberá por consiguiente ser escogido en la ausencia de información de costos de los dos errores, o cuando los costos son mas o menos equivalentes. [Giu03]

La tabla 40 muestra la matriz de confusión para el modelo de los vecinos más cercanos.

Tabla 40. Matriz de confusión para el modelo de los vecinos más cercanos

		Predicho	
		0	1
Observado	0	41.34	17.60
	1	12.14	28.92

Fuente: [Giu03]

Resulta que el modelo de los vecinos mas cercanos tiene el mismo índice de error de clasificación del modelo del árbol, 29.74%, y por ello es bueno en general. Pero las probabilidades para los errores tipo 1 y tipo 2 son ligeramente mas desbalanceados, y el modelo de los vecinos mas cercanos tiene la probabilidad de error tipo 1 mas bajo entre los modelos considerados. Por lo tanto, si los costos

del error tipo 1 son mayores que los costos del error tipo 2, el modelo de los vecinos más cercanos debería ser escogido. Si los costos del error relativo no están en consideración, los modelos del vecino más cercano y el modelo del árbol CART pueden ser escogidos, pues reducen al mínimo el índice de error de clasificación sobre el conjunto de datos de validación. [Giu03]

Se ve que los árboles de clasificación y el modelo de los vecinos más cercanos son las mejores herramientas para esta tarea predictiva. Sin embargo hay una diferencia de interpretación que debe ser tomada en cuenta. Los árboles de clasificación producen reglas que son fáciles de interpretar en términos lógicos, pero el modelo de los vecinos mas cercanos no entregan reglan explicitas. Estos modelos de los vecinos mas cercanos son herramientas no parametricas, por lo tanto tienen las ventaja de no requerir fuertes asunciones para el modelado, pero son mas duros de interpretar. Escoger entre las dos herramientas depende de quien va a utilizar los resultados; si los resultados van para estadísticos o para expertos de IT. Los árboles son probablemente mejores para expertos en el negocio quienes les gustaría una representación mas amigable para el usuario de lo que esta pasando realmente. [Giu03]

2.5.4. Despliegue y utilización de los resultados de minería de datos:

Primero que todo se pudo llegar a conocer aquellos clientes que se pueden considerar fieles según las condiciones de la empresa. Con esta información se pudieron crear relaciones mas rentables con los clientes vía electrónica en tiempo real.

Se comenzó otro proyecto que también involucraba minería de datos para realizar una mejor segmentación de los clientes considerados como fieles, a los cuales se les realizaron estrategias de mercadeo y ventas, que ayudara a afianzar la

relación con el usuario, lo que hizo que se incrementaron las ventas y por ende la rentabilidad de la empresa.
El sistema CRM cuenta con la ayuda de la minería de datos para conocer cuales son las necesidades más probables de cada cliente; la idea es tratar de personalizar al máximo el tratamiento que se le da al cliente. [CRM05]

El hecho de contar con la capacidad analítica que da la minería de datos, se generaron modelos específicos que se ajustaban a los problemas de negocio y a los datos que se tenían. Los beneficios de adoptar estos modelos de minería fueron los siguientes: [CRM05]

- Minimizaron los costos, captando a los clientes de una forma más eficaz.
- Ampliaron los beneficios, evitando que los clientes se fueran con la competencia.

Se dirigieron campañas de marketing solamente sobre la gente con más probabilidades de responder y llegar a convertirse en clientes rentables. Se trato de conocer el perfil de los clientes mas propensos a comprar los productos de la empresa, de esta forma se enfocaron esas campañas de marketing para conseguir el mayor retorno de inversión. [CRM05]

Se identificaron aquellos clientes con más posibilidades de abandonar la empresa y utilizar a la competencia y porque razón haría esto. Se tomaron las medidas necesarias para tratar de evitar esta deserción. Algunas de estas medidas fueron por ejemplo hacer promociones especiales que daban mejores beneficios entre mas antiguo fuera el cliente, mayores facilidades de pago, regalos en ocasiones especiales, etc.

Para las empresas es más costoso tratar de adquirir nuevos clientes que tratar de retener los que ya se tienen; esta es la razón de que se implementaran estas estrategias de fidelización. [CRM05]

Este nuevo CRM permitió actuar sobre las oportunidades más rentables del mercado, y a la vez se tuvo el conocimiento necesario de los clientes para responder a las cuestiones estratégicas que se plantearon para mejorar la rentabilidad de las relaciones. [CRM05]

Un proyecto específico que se siguió fue el recomendar productos que los clientes compraran con mayor probabilidad y de esta forma incrementar las ventas. Para esto generaron modelos de cestas de compras para observar que productos eran comprados juntos. Con estas recomendaciones no solo se dieron cuenta de que las ventas se incrementaron, sino que también se solidificaron las relaciones con los clientes.

Se incluyeron nuevos productos en el portafolio de la empresa, basados en las nuevas necesidades y deseos de los clientes, incrementando la satisfacción de estos y aumentando la vida de cada cliente.

En general, se puede ver que la mayoría de los casos tratados en este capitulo tratan problemas de negocios de empresas de comercio, donde el objetivo principal era aumentar las ganancias de la organización. Aunque también se pudo estudiar el caso del banco que pretendía hacer una mejor segmentación de los clientes, pero apuntando a enfilar y mejorar sus estrategias de mercadeo y ventas hacia estos segmentos, lo que le representara mejores dividendos. Según esto y lo visto a través de la investigación se puede notar que aunque la minería de datos es aplicable en muchos campos, las principales aplicaciones que se han hecho son en el campo de las empresas de retail y en el sector financiero y de seguros.

También se puede observar un patrón en las etapas de comprensión y preparación de los datos en todos los casos, del énfasis que se hace por escoger correctamente los datos a utilizar, de transformarlos a un formato único, de corregir errores que tengan, mirar que datos faltan, observar la distribución de las variables, y en general de utilizar un conjunto de herramientas como la estadística y la misma intuición de las personas para tratar de tener una buena fuente de datos con los cuales los algoritmos que se van a utilizar para el modelamiento arrojen resultados creíbles y óptimos para el negocio. Estas etapas se caracterizaron por su dificultad al momento de construirlas porque en la mayoría de las fuentes no se especificaba la forma como se hicieron y además implicaban tener un alto grado de conocimientos estadísticos, los cual hacia mas difícil comprenderlas y formarlas.

Encontrar la forma como las empresas despliegan y utilizan los resultados obtenidos con la minería de datos, también fue un poco difícil, pues algunas veces se encontraba solo el ciclo de minería hasta la parte de modelado y otras veces se tenía que armar por completo el caso. El camino que se siguió para completar el ciclo fue utilizar diversas fuentes para resumir un conjunto de posibles aplicaciones que tuvieran estos resultados en el problema de negocio que se estaba tratando.

3. PRINCIPALES HERRAMIENTAS DE MINERÍA DE DATOS

Antes de realizar un análisis de las herramientas que se considera hoy día como las más representativas a nivel mundial y en Colombia, es conveniente presentar una metodología que sirva como referente para la elección de una herramienta de Minería de Datos. Esta metodología pretende presentar un conjunto de pasos que se sugieren seguir en el proceso de búsqueda y selección de una solución de minería de datos. Así como los criterios que se podrían tener en cuenta para la evaluación de las herramientas encontradas en el medio.

3.1. METODOLOGÍA Y CRITERIOS DE SELECCIÓN DE HERRAMIENTAS DE MINERÍA DE DATOS

La siguiente tabla presenta las fases que se recomiendan seguir para un proceso de búsqueda, selección y evaluación de herramientas de minería de datos. En general, esta estructura puede ser seguida en el proceso de selección de cualquier tipo de herramienta software que se desee adquirir, lo que verdaderamente hace una diferencia entre las metodologías de selección de una herramientas de software convencional y una metodología para la selección de herramientas de minería de datos son los criterios presentados en las 4 plantillas que se proponen en la sexta fase de este proceso.

Tabla 41. Proceso de Selección Herramientas de Minería de Datos[46]

FASES	OBJETIVOS	RESULTADOS
DOCUMENTACIÓN DE LA NECESIDAD	Definición de áreas y funcionalidades de la organización que se involucraran con la herramienta Definición de objetivos que se pretenden alcanzar a través de la exploración de datos	**Documento con:** Áreas y funcionalidades definidas que se van a involucrar en el proceso de minado. Objetivos que se pretenden alcanzar con la obtención de la herramienta
ANÁLISIS DE LA NECESIDAD	Características del negocio que la herramienta debe entender: •Áreas de la organización involucradas •Procesos de negocios alcanzados •Costo máximo que se pagara por la adquisición	Documento con información necesaria para identificar los requisitos a tener en cuenta en la búsqueda de proveedores.
BÚSQUEDA EN EL MERCADO	•Identificación de proveedores en el medio •Identificación de la trayectoria de los proveedores y de las herramientas que ofrecen •Consulta con profesionales en otras organizaciones tratando de recolectar experiencias con las herramientas manejadas.	informe de proveedores
CONTACTO DE PRVEEDORES	•Contactar cada proveedor encontrado en la fase anterior •Solicitar información general de la herramienta •Reducir la cantidad de candidatos a 5	Documento con información recopilada de las herramientas basada en la información entregada por los proveedores.
ENTREVISTAS POSIBLES CANDIDATOS Y RECOPILAR INFORMACION	•Identificación de características técnicas •Identificación de casos de éxito con otras organizaciones •Descripción de módulos que la componen, funcionalidades de cada modulo	Documento con la información institucional de cada proveedor y los datos recopilados de forma homogénea de cada una de las herramientas para facilitar mas adelante la comparación de características entre ellas
CREACIÓN DE INFORME CON CRITERIOS PONDERADOS Y PUNTOS DE COMPARACIÓN COMUNES DE ACUERDO A LAS NECESIDADES DE LA EMPRESA.	Identificar criterios y asignarles pesos o ponderaciones: •Características técnico-funcionales de la herramienta. (características técnicas y funcionales de la herramienta) •Características del proveedor. o Evolución, crecimiento, facturación anual, ubicación geográfica, otros clientes experiencia) o Importante evaluar la solides del proveedor, ya que si el proveedor deja de existir, la empresa se queda con un sistema sin soporte ni posibilidad de evolucionar •Características del servicio •Características Económicas o Costos de licencia y mantenimiento de la herramienta	Plantilla con características técnico - funcionales con sus ponderaciones y sus resultados. Plantilla con Características de Proveedores, ponderaciones y resultados Plantilla con Características de Servicio, ponderaciones y resultados. Plantilla con Características económicas, ponderaciones y resultados.
EVALUACIÓN Y SELECCIÓN DE PROVEEDORES	Basado en los resultados obtenidos en la fase anterior, evaluar e identificar cual obtuvo la mayor puntuación según las ponderaciones asignadas a cada uno de los 4 grupos de características	Presentar el resultado de la evaluación realizada con ponderaciones y justificaciones de las ponderaciones a cada uno de los 4 grupos

3.1.1. Criterios para la Selección de Herramientas de Minería de Datos

A continuación se presentaran los criterios [BrG04] que se deben de tener en cuenta en el proceso de evaluación de proveedores y herramientas de Minería de datos, divididos en 4 categorías, los cuales son utilizados en la fase 6 del proceso que se presento en la tabla anterior.

[46] Construcción propia basada en información de [BrG04]

Características Técnica-Funcionales:

- Metodología / Ciclo de vida soportado por la herramienta.
- Adaptabilidad y flexibilidad para la toma de datos
    - Desde una Base de Datos: cantidad de formatos soportados para la toma de datos desde bases de datos diversas.
    - Desde Fuentes Externas: Cantidad de formatos soportados para la toma de datos.
- Facilidad para integrar diferentes técnicas de Minería de Datos
- Multi-lenguaje: permite trabajar en diferentes lenguajes (i.e. ingles)
- Cantidad de técnicas que posee para el minado de los datos
- Herramientas de visualización e informe
- Multiplataforma: posibilidad de desempeñarse en varias plataformas
- Instalación remota: permite instalación y trabajo técnico de forma remota
- Múltiples Usuarios: trabaja con una estructura que permite múltiples usuarios
- Seguridad: maneja perfiles de seguridad por tipos de usuarios
- Metodologías para copias de resguardo y recuperación de datos
- Tipos de interfaz de la herramienta
- Posee herramientas que administran las distintas versiones generadas
- Posee Documentación técnica
- Conectividad externa (i.e. Internet, EDI, etc.)
- Compatibilidad con correo electrónico

Características de Proveedores:

- Solidez del proveedor
- Ganancias
- Cantidad de empleados
- Clientes
- Perspectivas de evolución del proveedor en el mercado
- Ubicación de las oficinas

- Otras implementaciones que usen las herramientas
- Confianza (Criterio no cuantificable que queda a criterio de los miembros del equipo)

Características de Servicio

- Garantía del producto (que casos estarían cubiertos y que casos no estarían dentro de la garantía)
- Upgrade (cada cuanto tiempo sacan una nueva versión al mercado)
- Alcance de la licencia
- Soporte (existencia de un helpdesk para problemas no reportados)

Características Económicas:

- Costo de las herramientas
- Costos del Hardware
- Costos de Software
- Forma de pago de las licencias (una sola vez, solo cuando se implementa, una vez por año)
- Financiación
- Costos de Upgrade

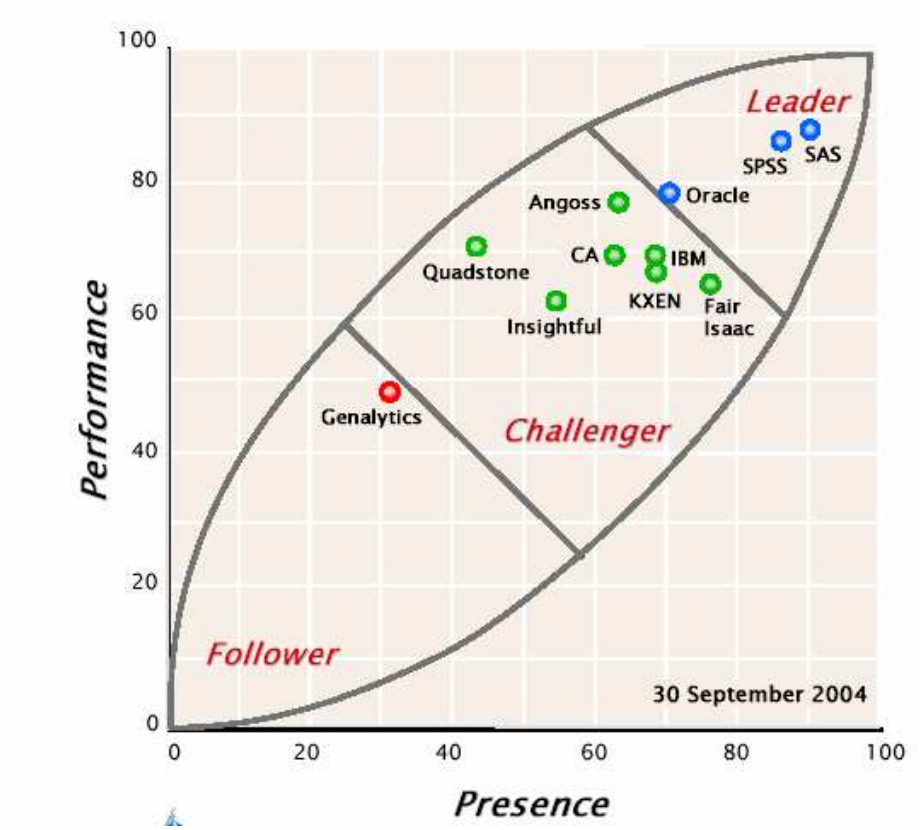
Cabe anotar, que para efectos de una evaluación, se debe asignar una ponderación a cada una de las categorías según criterio del equipo de trabajo, que sumen en total 100%.

Después de presentar el conjunto de criterios, que a consideración, son importantes tener en cuenta en el momento de analizar y seleccionar una herramienta, se presentan las principales herramientas que se tienen en el mercado, según algunos de los criterios mencionados anteriormente.

3.2. BENCHMARKING DE HERRAMIENTAS DE MINERIA DE DATOS A NIVEL MUNDIAL REALIZADO POR EMPRESAS EVALUADORAS DE TECNOLOGIAS TI

Debido a la gran cantidad de herramientas que se encuentran en el mercado de la minería de datos, se tomo como referencia para el propósito de esta investigación y el análisis de herramientas y proveedores, los resultados arrojados por la evaluación realizada por METASpectrum [Met04], la cual, muestra cuales son las principales herramientas de minería de datos a nivel mundial, teniendo como parámetros de referencia la presencia que estos proveedores tienen en el mercado y la funcionalidad de la herramienta ofrecida.

Figura 18. Principales Herramientas de Minería de Datos a nivel Mundial



Fuente: [Met04]

Este grafico presenta las herramientas evaluadas en tres cuadrantes diferentes. Estos grupos fueron definidos, primero, según el nivel de funcionalidad que posee la herramienta, el cual se mide según elementos como técnicas y tareas manejadas, métodos de visualización, etc. y segundo, según la presencia que estas tengan en el mercado a nivel mundial. Las herramientas que se encuentran en el cuadrante de los Lideres del mercado, identificados por este análisis, son los que tienen un producto estable y maduro que sobresalen en casi todas los aspectos de funcionalidad de Minería de Datos y que además, manejan una cuota grande del mercado con respecto a los otros jugadores. En algunos casos, las capacidades técnicas y funcionales no son muy superiores a las presentadas por los retadores, pero en un mercado especializado como la minería de datos, solo una pequeña porción de vendedores pueden sobresalir con este titulo de líderes en presencia y funcionamiento.

Las herramientas encontradas en el cuadrante de los retadores, se caracterizan sobre todo por un alcance levemente mas estrecho con la funcionalidad de la minería de datos y/o menos compromiso con la industria en general, a diferencia de los líderes. Por otro lado, otros de los retadores presentados en el cuadrante mágico, pueden tener un compromiso con la industria de la minería de datos pero todavía no poseen los recursos y la base de clientes necesarios para manejar parte del mercado y sobresalir en criterios de presencia.

Por ultimo encontramos a los seguidores, donde Genelytics, se presenta como único proveedor en este cuadrante. Este proveedor se reconoce como nuevo en el mercado de la minería de datos, con apenas unos cuantos clientes y con una estructura organizacional pequeña.

Las herramientas que fueron analizadas en la evaluación, serán presentadas a continuación, con un nivel de detalle mas profundo, para así identificar cuales son las características técnicas que cada una de estas tienen, los modelos de referencia que cada una utiliza para el desarrollos de un proceso de minería de datos y cuales son los campos en los que regularmente se trabajan estas

herramientas. Después de que se identifiquen las características asociadas a cada una de las herramientas, se identificaran cuales tienen presencia en el ámbito local, Colombia, y cuales son sus representantes en el país.

3.2.1. SPSS Clementine

Clementine es uno de los sistemas de minería de datos más populares del mercado, distribuido y creado por la empresa SPSS. Inicialmente esta herramienta visual fue desarrollada por ISL (integral Solution Limited) [Cle05].

Metodología de Referencia de Clementine

El diseño esta basado en la Metodología CRISP-DM

Proveedor

SPSS Inc. http://www.spss.com/clementine/

Características

- Acceso a los datos: fuentes de datos ODBC, Tablas Excel, Archivos planos ASCII y archivos SPSS
- Preprocesado de datos: *pick and mix*, muestreo, particiones, reordenación de campos, nuevas estrategias para la fusión de tablas, nuevas técnicas para recodificar, intervalos numéricos, etc.
- Técnicas de aprendizaje:
    - Árboles de decisión (C5.0 y C&RT)
    - Redes neuronales (redes de kohonen, perceptron multi capa y RBF)
    - Agrupamiento (K medias)
    - Reglas de asociación (GRI, A priori, etc)
    - Regresión lineal y logística
    - Combinación de modelos (boosting con C5.0)

- Técnicas para la evaluación de modelos guiadas por las condiciones especificadas por el experto.
- Visualización de resultados**:** esta herramienta ofrece un soporte grafico que permite al usuario tener una visión global de todo el proceso, que comprende desde el análisis del problema hasta la imagen final del modelo aprendido.
    - Para la visualización de los procesos, Clementine presenta las siguientes herramientas:
    - Gráficos estadísticos: Histogramas, diagramas de dispersión, etc.
    - Gráficos 3D y gráficos animados
    - Visualizadores interactivos de las diferentes tareas que realiza el experto.
    - Navegadores para árboles de decisión, reglas de asociación, redes neuronales de kohonen, agrupamientos, etc.
- Exportaciones: esta herramienta tiene la opción de generar automáticamente informes en formatos HTML y textos, volcado de los resultados del ejercicio de minería en la base de datos. Además, se tiene la posibilidad de exportar los modelos creados a otros lenguajes, tales como: C, SPSS, HTML, estándar PMML, SQL para árboles de decisión y reglas.

Plataformas en las que se encuentra disponible

Esta herramienta es un sistema multiplataforma y esta disponible para los siguientes sistemas:

- Windows
- Sun Solaris
- HP-UX AIX
- OS/400 [Cle05]

Algunos Campos de Aplicación

Estos son algunos de los campos en los cuales se ha aplicado Clementine como herramienta de minería de datos presentado por empresas representantes de cada campo:

- "Mantener y mejorar las relaciones con nuestros clientes"
  Hatem Labben
  Gerente de relaciones con el cliente
  Orange Communications
  Empresa de comunicaciones móviles e Internet
  Sector de las telecomunicaciones
- "Detectar patrones inusuales en las transacciones de los clientes y reportarlos a las autoridades competentes. Prediciendo lavado de dinero por parte de algunos usuarios."
  Unidad de prevención de lavados de dinero
  Banco Granahorrar, ahora BBVA
  Bogota, Colombia
  Sector Financiero
- "Pudimos sobrepasar nuestra meta original de venta – e incrementar las respuestas a las campanas enviadas por correo en un 100% "
  British telecommunications
  Sector de las telecomunicaciones

Representantes en Colombia

SPSS Andino (Colombia, Ecuador, Peru y Venezuela)
Cra 16a No. 78-11, Ofc. 501
Bogotá
Colombia
Teléfono: 57.1.6358585
Fax: 57.1.6358584
http://www.spsscolombia.com/

3.2.2. SAS Enterprise Miner

SAS Enterprise Miner, es una de las dos herramientas de Minería de Datos que proporciona SAS institute. Esta herramienta se enfoca en la minería de datos tradicional, mientras que la segunda herramienta de minería de datos que ellos manejan, *SAS Text Miner,* se enfoca en la minería de textos, la cual amplia su funcionalidad para trabajar con información contenida en archivos de texto. [Her04]

SAS Interprise Miner tiene una arquitectura distribuida, la cual permite acceder a toda su funcionalidad por medio de potentes interfaces graficas de usuario.

Metodología de Referencia

El diseño es basado en la metodología SEMMA[47] (*Sample, Explore, Modif., Model and Assess*), soportada por el propio instituto.

Proveedor

SAS Institute Inc. www.sas.com

Características

- Acceso a Datos: Formatos de archivo propio de SAS**,** Archivos de sistemas de bases de datos comerciales: *Oracle, DB2, Sybase, etc.*
- Procesado de Datos:
  - Transformaciones: Variables nuevas que se definen en base a las ya existentes aplicando una serie de operaciones de tipo matemático o estadístico.
  - Tratamiento estadístico para los valores desconocidos
  - Filtros para la eliminación de valores extremos

[47] Articulo: Metodología CRISP-DM vs. SEMMA.

- Tareas de muestreo: particionar el conjunto de datos para entrenamiento, validación y comprobación del modelo.

- Modelo:
  - Árboles de decisión: aproximaciones de los métodos CHAID, C&RT y C4.5
  - Regresión lineal y logística
  - Redes Neuronales: MLP y RBF, con sus variantes.
  - Construcción de modelos múltiples utilizando métodos *ensamble* como boosting, bagging, etc.
- Evaluación: Este modulo permite comparar la eficacia y el rendimiento entre diferentes modelos de aprendizaje.
- Visualización y presentación de resultados: La visualización de los resultados obtenidos después de aplicado el modelo son presentados:
  - Visualizador de resultados obtenidos: Gráficos estadísticos en dos o tres dimensiones, visores de árboles de decisión, diagrama ROC, etc.
  - Generador automático de informes: resume la información de un informe con formato HTML para ser visualizado en cualquier explorador.
  - Presentación de la información en lenguaje natural: Un ejemplo de este tipo de representación es, las reglas subyacentes en un árbol de decisión pueden ser expresadas por el sistema en un lenguaje natural fácil de comprender por el usuario.

Plataformas en las que se encuentra disponible:

El programa cliente y el servidor, pueden ser manejados en diferentes plataformas:

- Windows
- Linux
- Solaris

- HP-UX
- Digital Unix [Sas05]

Algunos campos de aplicación:

- "Honda mejora la seguridad y la satisfacción de sus clientes mientras se ahorra millones de dólares en costos de garantías"

  Usando SAS Enterprise Miner, para un sistema de detección temprana, honda utiliza la información que resulta de la retroalimentación realizada por los clientes y técnicos, resultando en el mejoramiento de la ingeniería para la construcción de mejores vehículos.

  American Honda

  Sector Automotriz
- "Mayores beneficios a través de puntajes de crédito"

  Después de la recuperación de la economía americana, los emisores de tarjetas de crédito se encuentran buscando áreas a cultivar y así mejorar las ganancias. Capital Card realiza estrategias de selección y de entendimiento de los riesgos que implica incursionar en nuevos segmentos de la población. Con la ayuda de SAS Enterprise Miner, analiza las grandes almacenes de datos para hacer mejores decisiones.

  Capital Card Bank

  Sector Financiero
- Trabajando con SAS, Alfonso Pedraza, profesor de operaciones y tecnología, esta introduciendo a los futuros profesionales en la ciencia de la Minería de Datos.

  Universidad de los Andes

  Bogota, Colombia

  Sector de la Educación
- SAS Enterprise Miner, ayuda a Vermont Country Store a mantener una relación intima con sus clientes, encontrando mejores caminos para enviar por correo sus catálogos a los diferentes clientes. Larry Shaw dice: " Cuando uno envía

mas de 50 millones de catálogos al año, y se puede mejorar el resultado de 1% a 3%, se puede decir que se obtuvo una gran ganancia"

Larry Shaw

Vicepresidente de Mercadeo y creatividad

Vermont Country Store

Sector Comercial (Retail y Distribución)

Representantes en Colombia:

SAS región andina

SAS Institute Colombia S.A.

Calle 114 # 9-45

Torre B Oficina 810

Teleport Business Park

Bogotá, Colombia

Teléfono: (571) 629-2525

Fax: (571) 629-2710

E-mail: infocolombia@sas.com

http://www.sas.com/offices/latinamerica/andean/

3.2.3. ODMS: Oracle Data Mining Suite (Darwin)

[Her04] Este sistema de minería de datos fue desarrollado en un comienzo por *Thinking Machines* como Darwin, ahora ha sido adquirido y es comercializado por *Oracle.* La arquitectura en la cual trabaja esta herramienta de Minería de datos es cliente – servidor y ofrece una gran versatilidad para el acceso a grandes volúmenes de datos.

La interacción con este sistema es fácil y eficaz gracias a la interfaz grafica que facilita la interactividad entre usuario y sistema.

Metodología de referencia:

Realiza una implementación completa del modelo de proceso KDD, utilizando una metodología propietaria que tiene como referente la metodología mundialmente conocida CRISP-DM.

Metodología Oracle:

1. Definición del problema
2. Preparación de los datos
    a. Acceso a los datos
    b. Muestreo de los datos
    c. Transformación de los datos
3. Construcción del modelo y evaluación
    a. Crear el modelo
    b. Probar el modelo
    c. Evaluar e interpretar el modelo
4. Ejecución del conocimiento
    a. Aplicar el modelo
    b. Reportes personalizados
    c. Aplicaciones externas [Ber05]

Proveedor

Oracle Corporation

Características

- El acceso a datos en diversos formatos: Almacenes de datos**,** Bases de datos relacionales (Oracle, SQL Server, Informix y Sybase), Archivos planos**,** Conjuntos de datos SAS
- Preprocesado de datos:
    - Muestreo de datos

- Obtención de información derivada apoyándose sobre una librería de funciones de tipo matemático, estadístico, lógico, manipulación de caracteres, etc.
- Particiones de los conjuntos de datos.

- Modelos de aprendizaje:
  - Redes neuronales para clasificación y regresión.
  - Regresión lineal (como caso particular de las redes neuronales).
  - Inferencia de árboles de decisión usando el criterio CART.
  - Vecinos más próximos.
  - Aprendizaje bayesiano.
  - Técnicas de agrupamiento (k medias y O-agrupamiento).
- Herramientas de visualización:
  - Representabilidad de los modelos inferidos: destaca su visualizador interactivo de árboles de decisión.
  - Resultados estadísticos: resultados relacionados con el propio modelo o estudios comparativos entre modelos diferentes aplicados sobre un mismo problema (histogramas, gráficos de línea, sectores, etc., en dos y tres dimensiones).
  - Importación de gráficos desde herramientas comerciales como Microsoft Excel, Microsoft Word y Microsoft Power Point. [Ber05]

Plataforma en la que se encuentra disponible:

Aplicación cliente de ODMS:

- Windows

Aplicación servidor de ODMS:

- Windows
- Sun Solaris
- PH-UX

Con Oracle Data Mining se pueden implementar estrategias para:

- Desarrollar perfiles de clientes a los cuales se quiere llegar
- Anticipar y prevenir una disminución de clientes
- Adquirir nuevos clientes e identificar aquellos clientes que tienen un mejor perfil para el negocio
- Identificar oportunidades de Cross-sell
- Detectar actividades no comunes y fraudulentas
- Descubrir nuevos grupos y segmentos
- Desarrollar perfiles de clientes
- Identificar puntos promisorios en el descubrimiento de drogas [Ber05]

Representantes en Colombia:

Oracle Colombia
Calle 100 #7A-81 - Piso 3
Bogota, Colombia
(57-1) 611-9600
http://www.oracle.com/global/lad/index.html

3.2.4. DB2 Intelligent Miner

Esta herramienta comercial con distribución Cliente/Servidor fue pensada y desarrollada por IBM para explotar los grandes sistemas de información que tienen las grandes corporaciones [IBM04].

Esta suite proporciona una amplia gama de herramientas para aplicar tareas de minería de datos tales como:

- Agrupamiento
- Asociaciones
- Patrones
- Clasificación

- Predicción
- Análisis de series temporales

IBM DB2 Intelligent Miner es el conjunto de los siguientes productos:

- Intelligent Miner Scoring
- Intelligent Miner Modeling
- Intelligent Miner visualization
- Intelligent Miner for data[48]

Metodología:

IBM Intelligent Miner utiliza una metodología propietaria que tiene como referente la metodología mundialmente conocida CRISP-DM.

1. Definir el problema del Negocio.
2. Definir el modelo de datos a utilizar
3. fuente y preprocesamiento de los datos
4. seleccionar la técnica de Minería de Datos
5. Interpretar los resultados obtenidos
6. Presentar los resultados

Proveedor:

IBM Corporación

Características técnicas

[IBM04] Acceso a Datos: DB2**,** Archivos planos**,** Variedad de datos accesibles mediante la herramienta *Datajoine*, como por ejemplo: *Oracle* y *TeraData.*

- Técnicas de aprendizaje:
    - Árboles de Decisión (versión modificada de CART)
    - K-means
    - Redes neuronales (MLP, propagación hacia atrás, RBF)
    - Regresión Lineal

[48] Articulo: Manual de referencia de Administración y Programación de DB2 Intelligent Miner v8.2 (2005)

- Herramientas de Visualización: Se utiliza la herramienta Intelligent Miner Visualization para ver y analizar los resultados obtenidos del proceso de minería de datos y de esta manera poder realizar una interpretación de los mismos. Intelligent Miner Visualization proporciona los siguientes visualizadores en Java para presentar los resultados del modelado de los datos para el análisis:
    - Visualizador de asociación
    - Visualizador de clasificación
    - Visualizador de agrupamiento
    - Visualizador de regresión

Plataformas

Aplicación Servidor:

- Windows (2000, NT, XP)
- Solaris
- AIX
- OS (390, 400)
- z/OS

Aplicación Cliente:

- Windows
- AIX

Algunos Campos de aplicación:

- Utilizando modelos de clasificación:
    - Aprobar o denegar Peticiones de seguro
    - Detectar fraudes de tarjetas de crédito
    - Identificar defectos en imágenes de componentes manufacturados
    - Diagnosticar condiciones de error
- La función de minería de agrupación se utiliza a menudo en CRM:
    - Comercialización cruzada

- Venta cruzada
- Planes de comercialización personalizados para diferentes tipos de clientes
- Toma de dediciones sobre el medio a utilizar
- Análisis de los objetivos de compra
- Muchos otros sectores

- La función de minería de asociación se utiliza para:
  - Determinar que producto o productos es probable que se encuentre en una transacción si en esta se dan determinados productos.
- La función de minería de regresión se utiliza para:
  - Determinar a cuantos clientes debe dirigirse si desea lograr un volumen de negocio determinado.
  - Determinar de cuanto debe ser el volumen de negocio a dirigir, a cierta cantidad de clientes[49]

Casos:

- El Banco de Montreal buscaba una relación más estrecha y estable con sus clientes, tener un servicio al cliente proactivo en lugar de reactivo y tener una idea clara y amplia de cada una de las relaciones con los clientes del negocio, para así entender mejor sus necesidades y deseos. Con Intelligent Miner lograron todos estos desafíos.

  Murray Sutherland

  VP Sales and Service
- En un ambiente de alta competencia y un mercado constantemente en cambio, SPARQ pasó de tener a IBM como administrador de información a tener a IBM como una herramienta de soporte de estrategias de mercadeo, realzando la capacidad de leer el pulso del mercado y tomar decisiones más eficaces para mantenerse competitivos.

  Ling Dongjie.

[49] Articulo: Manual de referencia: Administración y programación Intelligent Miner v8.2

Departamento de Mercadeo

- Con la misión de traducir todos los descubrimientos genéticos en nuevos tratamientos, TGEN requiere de poderosas capacidades de computación para manejar las enormes demandas de procesamiento de análisis genéticos avanzados. Con la adquisición del nuevo sistema que les permite tener más escalabilidad y un avanzado proceso de minería de datos, consiguieron tener como beneficio: 2 trillones de calculaciones por segundo, reduciendo el tiempo del ciclo en un 99%.
  Dr. George Poste
  TGEN institute

Representantes en Colombia:

IBM de Colombia
628-1000 en Bogotá
01 8000 917555 a nivel nacional
Fax: (57) 1-635-7611
Carrera 53 No. 100-25
Bogotá – Colombia

3.2.5. CleverPath Predictive Analysis Server

Proveedor:

Computer Asociates

Características:

[Blo04] CleverPath Predictive Server tiene una tecnología avanzada de Neugent que analiza el comportamiento basado en patrones y relaciones encontradas en datos históricos. Un ejemplo claro de esto, es que puede fácilmente aprender el

perfil de los clientes con más valor para la compañía o identificar factores con gran influencia en las ventas.

- Descripción de la herramienta: Comúnmente, las herramientas de minería de datos, proveen una serie de algoritmos independientes, mientras que cleverpath proporciona un solo algoritmo llamado “Funcional Link Net”. Este algoritmo patentado se encuentra a un nivel muy bajo y no es expuesto a los usuarios, pero puede ser ejecutado de diferentes maneras (e.g. la clasificación y clusttering son una instanciacion del algoritmo). Cada instanciacion es conocida como un Neugent.
- Arquitectura: El mayor componente de esta herramienta es el Neugent IDE (integrated development inviroment).

  Arquitecturas Nuegent:
  - Clustering
  - Decisión tree
  - Predicción de eventos
  - Predicción de valores, el cual incluye un mecanismo para soportar análisis de sensibilidad.

Plataformas:

- Windows NT family
- Sun Solaris

Casos de aplicación:

- CRMs
- Cross-selling
- Mercadeo directo
- Calidad de productos
- Recomendación de productos
- Manejo de riesgos
- Detección de fraudes

- Diseño de torpedos
- Predicción de lesiones de jugadores
- Predicción de funcionamiento de jugadores

Representantes en Colombia:

Computer Associates de Colombia
Edificio grupo Santander Central Hispano carrera 7
No. 99- 53 torre 2 oficina 401
Bogota- Colombia
Teléfono: (571) 6326000
Fax: (571) 6326001
http://www.ca.com/offices/colombia/contact.htm

3.2.6. KnowledgeStudio 4.2 and Mining Manager 2.1

Proveedor:

Angoss Software

Características:

[Men00] Angoss ha desarrollado paquetes de aplicaciones que apuntan a la industria financiera.

- Niveles de exploración:
    - Exploración de datos básica a través de Visualización: Esta herramienta soporta exploración de datos básica a través de la perfilacion de valores y una plantilla que guía paso a paso al usuario que ofrece las mismas funcionalidades que se encuentran disponibles en una plantilla de Microsoft Excel. la plantilla puede ser editada, cortada y pegada en otras aplicaciones. Es fácil moverse de campo a campo dentro de la plantilla.

- Provee el descubrimiento de patrones en los datos: Este descubrimiento de patrones se hace a través de análisis de árboles de decisión, tres formas de redes neuronales y dos formas de análisis de clusters. Generalmente el descubrimiento automático de patrones se hace a través de una prueba de bondad de ajuste o a través de técnicas de reducción de varianza
- Provee soporte para la construcción de modelos: Modelos de árboles de clasificación y regresión se hacen internamente para ser aplicados a un set de datos. Otros modelos de representación, incluyendo las redes neuronales y los algoritmos de clustering son llevados a cabo internamente para la puntuación de datos.

- Técnicas de minería de datos:
  - Inducción supervisada
  - Descubrimiento de asociaciones
  - Descubrimiento de secuencias
  - Clustering
  - Visualización
- Funcionalidad para limpieza de datos → SI
- Herramientas para la transformación de datos → No
- Herramientas para manejar valores de datos que faltan o nulos → Si
- Herramientas para evaluar los modelos → Si

Plataformas:

- Unix
- Windows

No existen Representantes en Colombia

3.2.7. KXEN Analytics Framework 3.0

Proveedor:

KXEN

Metodología de referencia:

1. Acceso de datos
2. Manipulación de los datos
3. Preparación de los datos
4. Modelación de los datos
5. Ejecución del modelo
6. Producción

Características:

- Funciones: [Ang03] Procesar bases de datos con muchos campos para hacer matemáticamente predicciones y construir modelos, exportar modelos como códigos de programa en diferentes lenguajes, incluyendo: C, ANSI, SQL, SQL para SQL Server y XML.
- Técnicas de Modelado:
  - Regresión lineal
  - reglas de asociación
  - agrupamiento
  - series de tiempo

Plataformas:

- Windows NT 32
- Solaris
- AIX
- HP-UX
- Linux

Campos de Aplicación:

- Identificación de comportamiento de los clientes, especialmente de aquellos que tienen intención de irse, para así tomar medidas correctivas.
- Identificación de nuevos nichos de mercado**.**
- Cross-Selling
- Clasificación de clientes e identificación de los clientes importantes
- Segmentación del mercado para realizar un mejor mercadeo según las necesidades del segmento.
- Predicción del comportamiento de los clientes y productos que ellos usaran.
- Clasificación de mercados, obtener la campana de mercadeo correcta y el tiempo correcto.
- Manejo de canales de distribución[50]

No existe Representantes en Colombia

3.2.8. Insightful Miner 3.0

Proveedor:

Insightful Corporation

Características

- Acceso a Datos:
    - Archivos ASCII delimitados
    - Archivos planos provenientes de SAS, SPSS, Excel, entre otros.
    - Acceso a bases de datos ODBC
    - Accesos nativos Oracle, DB2, Microsoft, SQLserver y Sybase**.**

[50] http://www.kxen.com/industries/finance.php

- Manipulación de los Datos:
    - Métodos de estratificación para pruebas
    - Líneas: Filtros, particiones, muestras, pilas, etc.
    - Columnas: creación, filtros, modificaciones, normalizar, trasponer y reordenar.
    - Crear y modificar columnas y filtros de línea utilizando lenguaje de expresión.
- Limpieza de los Datos
    - Limpiar y arreglar valores perdidos con métodos de preservación de la varianza
    - Detección de duplicados
    - Manejo de datos perdidos.
    - Detección de outliers.
- Técnicas de Modelado y visualizadores
    - Modelos de predicción y clasificación con opciones básicas y avanzadas
    - Árboles de clasificación y regresión con árbol simple o técnicas de conjuntos
    - Regresión lineal y logística
    - Redes neuronales con perceptrones de múltiples capas
    - Métodos de entrenamiento de redes neuronales
    - Visualizador de redes neuronales
    - Clasificador de bayes Naive
    - Modelos de segmentación con K-Means clusterings [Ins06]

Plataformas:

- Windows
- Solaris

Campos de aplicación:

- Detección de fraudes
- Análisis de campañas de mercadeo**.**
- Análisis de comportamiento de los clientes
- Gestión de calidad de datos
- Reducir de fuga de clientes
- Análisis financieros
- Análisis en Biomedicina

No existe representante en Colombia

3.2.9. Quadstone System V.5

Proveedor

Portrait Software

Metodología

Metodología propietaria basada en la metodología CRISP-DM

Características

- Modelos de Predicción:
  - 5 variantes de algoritmos de árboles de decisión (incluyendo el ID3 y el C&RT entre otros)
  - 6 variedades de modelos de regresión
  - Clustering K-means indirecto.
- Exportación de los datos:

  Como imágenes, reportes, hojas de cálculo, listas, modelos, XML, Set de datos estadísticos o tablas de bases de datos.

No existe representante en Colombia

Hasta el momento se ha presentado la información y características técnicas más relevantes de las principales herramientas a nivel mundial según la evaluación antes mencionada. De las 13 herramientas analizadas se entrega la información de 10 de ellas, esto se debe a que algunos de los proveedores que se encuentran ubicados en los cuadrantes de retadores y seguidores, son proveedores relativamente pequeños en el mercado de tecnologías de Minería de Datos, haciendo un poco difícil la obtención de información de los mismos.

Dentro de este mismo análisis se buscaba encontrar los precios de estas herramientas, pero esta es información que los proveedores no presentan públicamente en la mayoría de los casos, es por esto que a continuación se presenta una estimación basada en precios entregados por usuarios que han adquirido estas herramientas y que fue presentada en una encuesta de Kdnuggets.

Tabla 42. Estimación de precios de Herramientas de Minería de Datos

Nivel	***Rango de precios***	***Proveedores***
Nivel empresarial	US$ 10.000 o mas	SAS
		SPSS
		Oracle
		KXEN
		Insightfull
		IBM
		Fair Isaac
Nivel Departamental	Desde US$ 1.000 hasta US$ 9.999	Angoss
		Cart/Mars/TreeNet
		Megaputer
		Microsoft SQL Server
		ThinkAnalytics
		Equbits
		GhostMiner
		Mineset

Nivel Personal	Hasta US$ 999	Excel
		See5

Fuente: http://www.kdnuggets.com

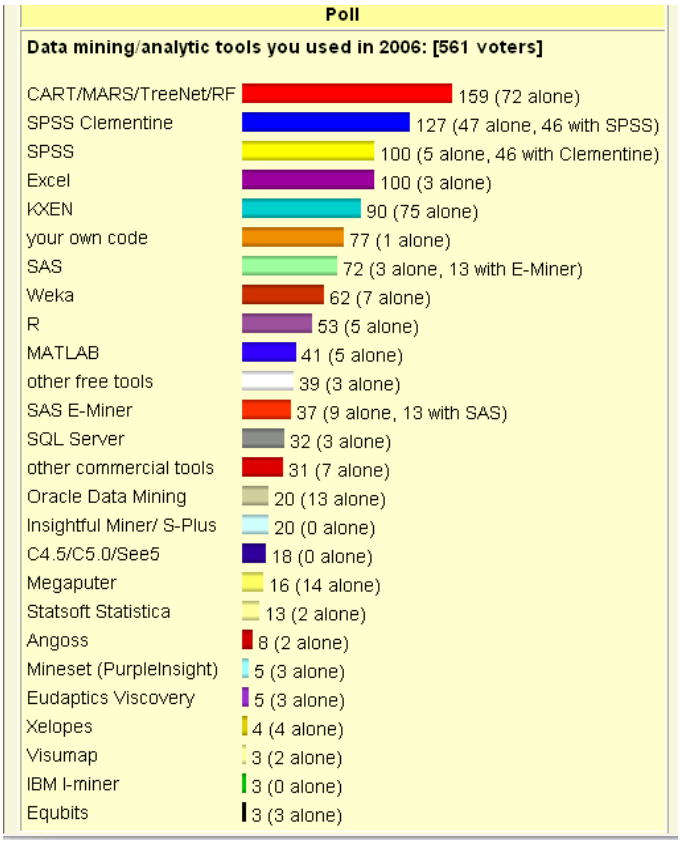
Estos precios son basados en datos entregados en las encuesta realizadas y presentan una idea general de cómo se encuentran los precios de las herramientas mas reconocidas en el mercado mundial que puede ser tomado solo como punto referente en el caso de Colombia.

3.3. BENCHMARKING DE HERRAMIENTAS DE MINERIA DE DATOS REALIZADA POR USUARIOS A NIVEL MUNDIAL

La evaluación realizada por el grupo METASpectrum [Met04], ahora GARTNER, se presento como una evaluación formal donde se tuvo como base la funcionalidad y la presencia de las herramientas en el mercado a nivel mundial. Pero además de esta evaluación y debido al propósito de este análisis, se decidió también presentar una evaluación un poco mas informal realizada por KDNUGGETS, donde se muestra cuales son las herramientas mas utilizadas a nivel mundial según los mismos usuarios a lo largo de este año. Esta evaluación es una recopilación de votos de usuarios por las diferentes herramientas de minería de datos que se tienen disponibles en el mercado, muchas de las cuales fueron mencionadas y descritas en el análisis anterior.

A diferencia de los resultados mostrados en el cuadrante anterior, esta ilustración presenta las herramientas más usadas y que han sido escogidas por su fácil adquisición y su fácil manejo independiente de su presencia en el mercado.

Figura 19 Herramientas de Minería de Datos mas utilizadas en el 2006



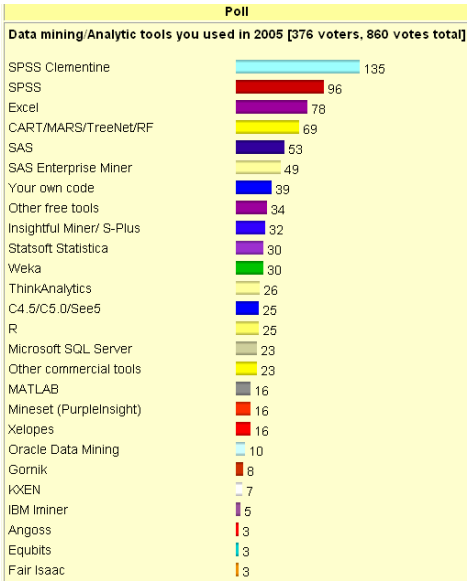
Fuente: http://www.kdnuggets.com/

Esta encuesta demuestra que las herramientas mas utilizadas por usuarios alrededor del mundo son Cart/Mars/TreeNet, SPSS Clementine, Excel, KXEN, SAS, Weka, Oracle Data Mining y Megaputer entre otras.

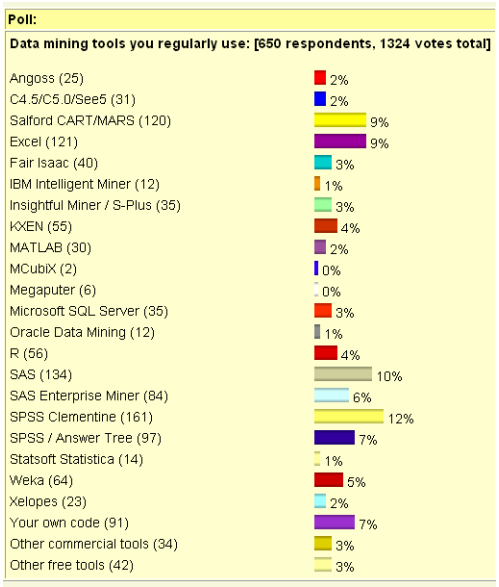
Estos resultados, si se comparan con los obtenidos por Kdnuggets en los 2 años anteriores, muestran como SPSS Clementine se ha mantenido en las primeras posiciones de la calificación durante los últimos tres años de encuestas y como Cart/Mars/TreeNet ha ido escalando posiciones a través de los años.

Figura 20

Herramientas de MD más utilizadas en el 2005



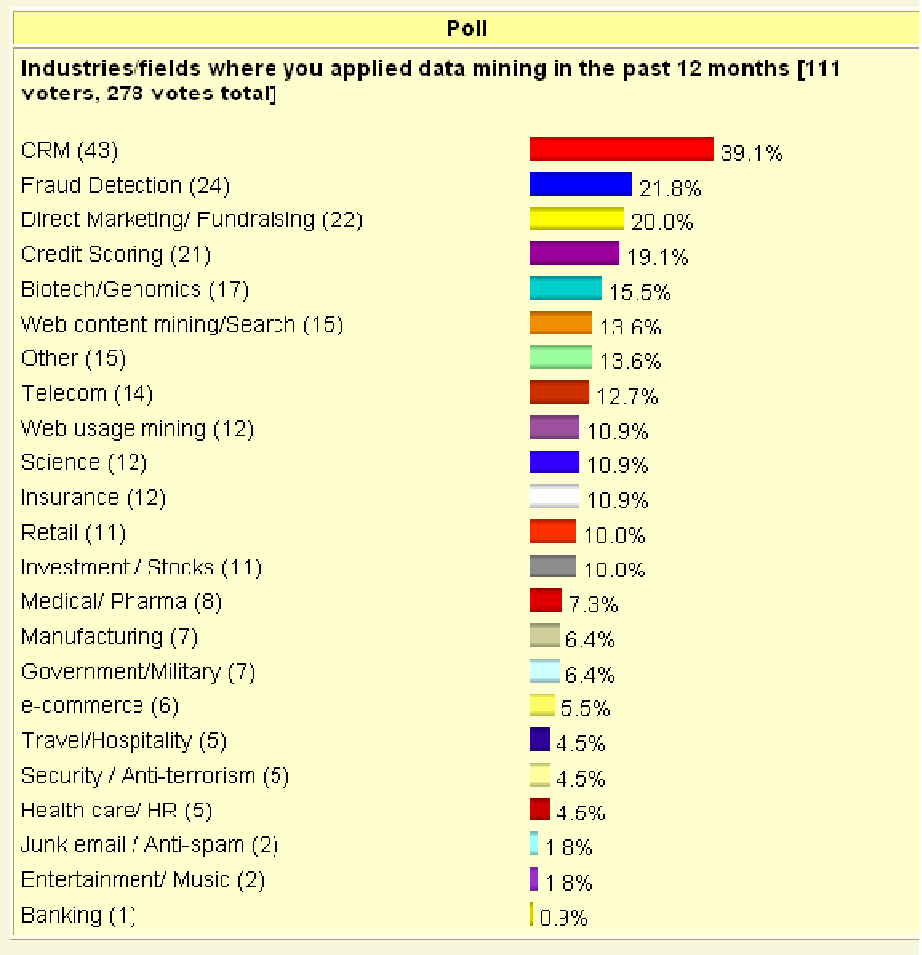
Herramientas de MD mas utilizadas en el 2004



Fuente: http://www.kdnuggets.com/

Por ultimo, en la ilustración 21 que se tiene a continuación, se muestra un grafico con los resultados obtenido en una encuesta que muestra cuales son los campos y sectores que más utilizan la minería de datos como herramienta de soporte al negocio.

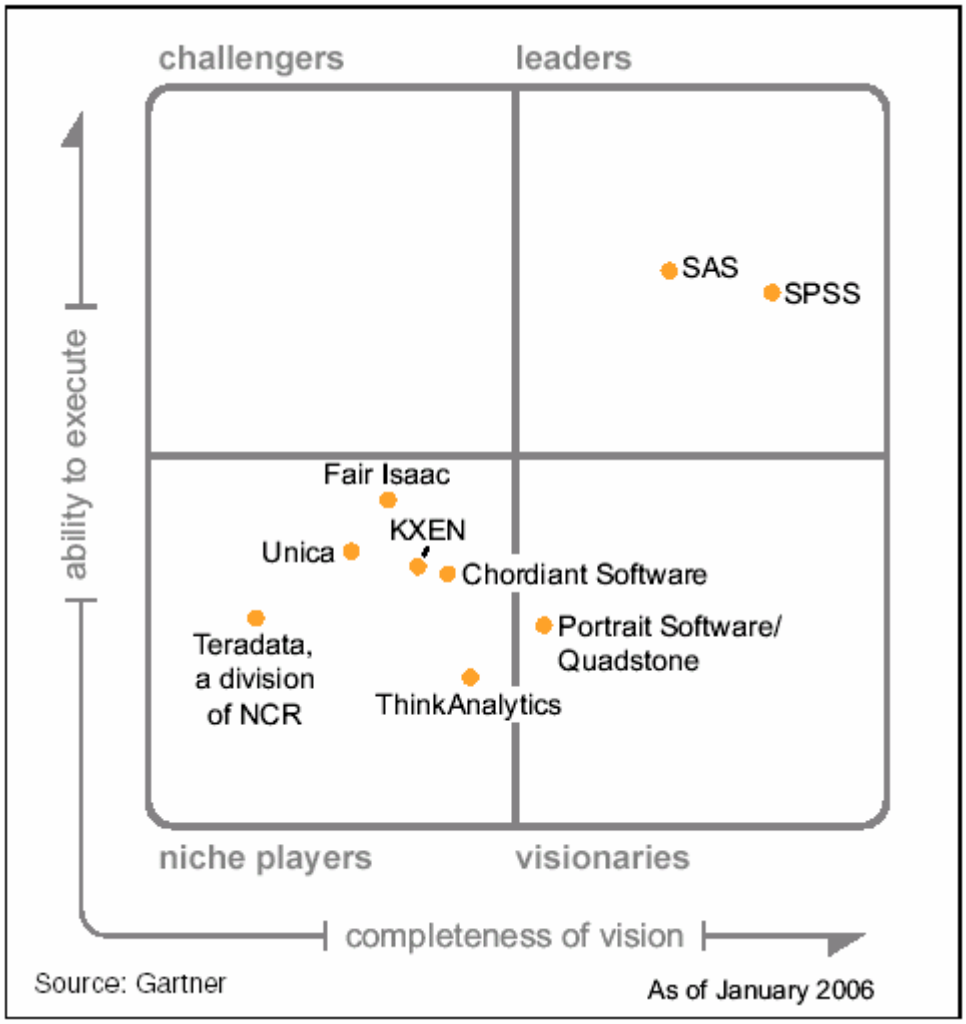
Figura 21. Campos donde más se aplica la Minería de Datos en el 2006



Fuente: http://www.kdnuggets.com/

Estos resultados muestran que uno de los campos donde más se ha explotado los beneficios de la minería de datos a nivel mundial, es en la administración de las relaciones comerciales con los clientes. Es por esto que para completar este análisis, se presenta el siguiente cuadrante realizado por GARTNER [Her06] que muestra las herramientas mas utilizadas a nivel mundial en el campo de CRM.

Figura 22. Herramientas más utilizadas a nivel mundial para CRM en el 2006



Fuente: [Her06]

Se puede observar como SAS con SAS Enterprise Miner y SPSS con Clementine son los líderes en herramientas de minería de datos que se encuentran enfocadas en la administración de relaciones comerciales con clientes.

La ilustración presenta a los proveedores en 4 cuadrantes diferentes, según el posicionamiento de sus herramientas, las funcionalidades que presentan y la visión que ellos tienen del mercado.

[Her06] Los vendedores líderes son los que sobresalen en el segmento de mercado de minado de datos de clientes y que además tienen un gran impacto en la dirección y el crecimiento del mercado. Mientras que los retadores son vendedores que han entrado al mercado del minado de datos de clientes ofreciendo sus herramientas de minería de datos como un complemento de sus aplicaciones de negocio establecidas. Normalmente ellos ofrecen una buena funcionalidad pero esta funcionalidad esta atada a sus propias aplicaciones. Los Visionarios son vendedores que tienen una fuerte visión de la evolución del minado de datos de clientes, es por esto que tienen sus objetivos centrados en esto. Por ultimo los actores de nicho, los cuales se centran en un segmento específico del mercado, están enfocados en una funcionalidad específica de CRM.

3.4. ANALISIS DE HERRAMIENTAS DE MINERIA DE DATOS EN COLOMBIA

Ahora, después de tener una visión clara de lo que esta ocurriendo a nivel mundial con respecto a la minería de datos y sus proveedores, se pasara a analizar como se encuentra este mercado en el ámbito nacional.

Dentro de las herramientas analizadas anteriormente y que se presentaron como las principales a nivel mundial, se tienen 5 de estos trece proveedores en Colombia: SPSS, SAS, Oracle, IBM y Computer Associates. Pero no solo estos proveedores con sus herramientas tienen presencia en Colombia en el campo de Minería de Datos, en la siguiente tabla se presentara el conjunto de proveedores y herramientas más representativas a nivel nacional. La información que se muestra fue basada en entrevistas realizadas a los diferentes proveedores y en el hecho de que se tienen una representación autorizada en Colombia.

Aunque algunos de los proveedores no presentan casos de éxito en Colombia hasta el momento, esto no significa que ellos no tengan la herramienta disponible

para el mercado Colombiano, es mas, muchos de ellos cuentan en las entrevistas que están creando campañas para el reconocimiento de estas herramientas como parte de los portafolios de soluciones que una compañía puede tener a su disposición.

Tabla 43. Principales proveedores de Herramientas de Minería de Datos en Colombia[51]

Proveedor	***Herramienta de Minería de Datos***	***Sector en el que se ha aplicado en Colombia***	***Proyectos Realizados***
SPSS	SPSS Clementine	Sector Financiero – Bancos	2
Oracle	ODMS	Sector de telecomunicaciones	1
SAS	SAS Enterprise Miner	Sector Financiero - Bancos	1
Computer Associates	Cleverpath Predictive Análisis	*No existen casos de éxito de implementación de procesos de Minería de Datos en Colombia hasta el momento.*	0
IBM	DB2 Intelligent Miner	Sector Publico Sector Industrial - comercial	2
Microsoft	SQL Server 2005 y SQL Server 2000	*No existen casos de éxito de implementación de procesos de Minería de Datos en Colombia hasta el momento.*	0
Teradata	Teradata Warehouse Miner	*No existen casos de éxito de implementación de procesos de Minería de Datos en Colombia hasta el momento.*	0

Aunque en el caso de Microsoft como proveedor de herramienta de Minería de datos no se presenta hasta el momento un caso de éxito en este campo, se decidió incluir este proveedor dentro de la gama de herramientas debido al posicionamiento y reconocimientos de sus herramientas SQL Server 2000 y SQL Server 2005 entre las industrias Colombianas.

[51] Construcción propia basada en información suministrada por los proveedores

A continuación se presentara información y características técnicas de las herramientas que no fueron consideradas al inicio de este capitulo en el análisis mundial y que son presentadas en la tabla anterior.

3.4.1. Teradata Warehouse Miner

[MiT02] Teradata Warehouse Miner es un conjunto de herramientas de minería de datos e interfaces de aplicación que provee un alto rendimiento y minado escalable de los datos. El predecesor de Teradata Warehouse Miner fue Teradata Miner Stats, el cual tenía solo las funciones básicas de la minería de datos.

Metodología

1. Definición los problemas del negocio
2. preparación de arquitectura y tecnología
3. Pre-procesamiento de los datos
    a. funciones de transformación
    b. funciones de visualización de datos
    c. funciones de exploración
    d. Matrices de reducción de datos
4. Creación del Modelo
    a. funciones estadísticas avanzadas
    b. algoritmos
    c. Visualización de datos
5. Ejecución del Modelo
    a. Integración del modelo
    b. Gestión del ciclo de vida

Características técnicas

- Técnicas de Modelado
  - Regresión Lineal
  - Regresión logística
  - Análisis de factores
  - Arboles de decisión
  - Clustering
  - reglas de asociación y análisis de secuencias
- Visualización de Datos
  - Estadísticas descriptivas
  - Regresión Lineal y logística
  - Análisis de clusters

Plataformas

- Windows XP, 2000, NT

Campos de aplicación

- Segmentación de clientes
- Identificación de clientes que puedan responder a una promoción especifica.
- Detección de Fraudes.
- Identificación de clientes que posiblemente se retiraran.
- Optimización de canales de comunicación con los clientes.

Representantes en Colombia

Contacto: Gustavo Correa

Carrera 37 # 30 – 20

Bogota, Colombia

TEL: 369 2000 Ext. 201

Email: Gustavo.correa@ncr.com

3.4.2. SQL Server 2005

Esta herramienta es considerada como una plataforma integral de inteligencia empresarial, que incluye las características, las herramientas y las funciones necesarias para crear aplicaciones analíticas de tipo tradicional e innovador[52].

Componentes del sistema de inteligencia empresarial

1. Extracción , transformación y carga
2. Almacenamiento de datos relacionales
3. bases de datos multidimencionales
4. Minería de datos
    a. Analysis Services para SQL Server 2000 y SQL Server 2005
5. Informes administrativos
6. Análisis y consultas ad hoc
7. Herramientas de desarrollo y administración de bases de datos

Características técnicas

- Acceso de datos: Se puede obtener acceso de los datos desde la base de datos relacional o desde los cubos de Analysis Service
- Técnicas de Modelado
    - Árboles de decisión
    - Naive Bayes
    - reglas de asociación
    - Análisis de secuencias
    - series de tiempo
    - redes neuronales
    - Segmentación [SQL05]

[52] http://www.microsoft.com/latam/technet/productos/servers/sql/2005/dwsqlsy.mspx

Plataformas

- Windows

Campos de aplicación

- Análisis de riesgos crediticios
- Análisis de rotación de clientes
- Retención de clientes
- Análisis de perfil
- Campanas de distribución de correos
- Análisis de cesta familiar
- Pronostico de ventas
- Predicción de cotización de acciones
- Cotización de tazas del seguro
- Detección de fraudes
- Análisis de infusión de red [SQL05]

Representantes en Colombia

Microsoft Colombia

Tel: 01 800 051 0595

http://support.microsoft.com/gp/contactenos/?ln=es-la

Una de las principales dificultades que se presenta cuando se plantea la idea de adquirir una solución de minería de datos por parte de una empresa, puede ser la de identificar cual seria el proveedor adecuado para la organización, cuales son los criterios y puntos principales que se deben identificar para seleccionar un buen proveedor según las necesidades y problemas del negocio que se desean resolver. Esta puede ser una pregunta difícil de resolver debido a la gran cantidad de opciones que se tienen en el mercado de la minería de datos. Pero además de lo cuestionamientos relacionados con los proveedores, surgen otras preguntas con

respecto a las herramientas que ellos presentan. Identificar cual será la herramienta que mejor se acomode a los objetivos planteado por la empresa, cual es la herramienta con el mejor abanico de características técnicas según las necesidades del usuario, entre otras preguntas, surgen debido al poco conocimiento que se tiene de tema de minería de datos.

Para una organización o empresa que inicie el proceso de selección de una herramienta de minería de datos por primera vez, puede ser difícil responder y cubrir todas las preguntas que se pueden generar en torno a la decisión de que proveedor y que herramienta escoger.

Es por esto que en este capitulo se presenta, primero, un conjunto de pasos o fases que se pueden utilizar como una referencia que ayudara y guiara el proceso de selección y un conjunto de criterios, que pueden dar una mejor visión de lo que la empresa busca en las herramientas y en los proveedores. Debido a que algunos de los criterios presentados pueden tener mayor peso con respecto a otros según los usuarios, los pesos que se asignan a cada uno se dejan a criterio del grupo de trabajo de la organización, pues son estos quienes pueden identificar cuales de los puntos mencionados son los mas importantes para el proyecto que deseen implementar.

Una vez creado un contexto claro para el usuario u organización con los puntos claves que se deben identificar en el análisis de proveedores y herramientas de minería de datos, se presenta un segunda parte en este capitulo, el cual da una visión clara de cual es la situación de los proveedores y las herramientas de minería de datos a nivel mundial, mostrando cuales son las mas importantes, tomadas desde dos puntos de vista claves en el mercado; el primero, por parte de empresas evaluadoras de tecnologías y la segunda por parte de los mismos usuarios de estas herramientas de minería de datos. El objetivo de este benchmarking es generar una base de conocimiento sobre las herramientas mejor

posicionadas a nivel mundial, presentando algunas de sus características técnicas y casos en los cuales se puede aplicar estas herramientas según los proveedores. Después de realizar este análisis, se encontró que los resultados arrojados por las empresas evaluadoras como Gartner y Metaspectrum y los resultados obtenidos en las encuestas libres realizadas a los usuarios de herramientas de minería de datos presentan unas pequeñas diferencias en cuanto a sus resultados, lo cual es de entenderse debido a las metodologías utilizadas en las evaluaciones. Aunque no son muy significativos, si se debe cuestionar por que una herramienta como Cart/Mars/TreeNet, la cual ha estado en los primero puestos de las encuestas informales en los últimos 3 años, no este presente en los resultados presentados por las empresas evaluadoras. Además de este caso, los resultados guardan cierta consistencia en ambos resultados.

Por ultimo, y después de identificar cual es la situación mundial de los proveedores y las herramientas de minería de datos, se entra a analizar cual es la situación actual en Colombia. Basándonos en los resultados obtenidos en la evaluación del punto anterior, se busca, cual de las herramientas y proveedores que se consideran como principales a nivel mundial, se encuentran presentes en Colombia, encontrando 5 de estas en el ámbito nacional. Una vez identificadas los proveedores, se realizo una entrevista informal a los representantes de estos. Las conversaciones arrojaron que solo 4 de estos proveedores han implementado en algún momento proyectos de minería de datos en Colombia, SAS, SPSS, Oracle e IBM. En el caso de Oracle, su primer y único proyecto se encuentra en proceso de implementación en estos momentos. Para los casos de IBM, SAS y SPSS, son proyectos que se han implementado con gran éxito. También se puede identificar en las respuestas de los proveedores que los sectores en los cuales se esta haciendo uso de estas herramientas son el sector financiero ubicándose solo en bancos, el sector de las telecomunicaciones, el sector publico y el sector industrial-comercial. Por ultimo se les pregunto a estos proveedores cuales eran los sectores en los que se espera tener una mayor demanda de estas herramientas

en los próximos 5 años y la respuesta fueron en el sector financiero principalmente.

Mientras en Colombia se identifica que las soluciones de minería de datos son y serán principalmente utilizadas (en los próximos 5 años) por el sector financiero, más precisamente bancos para detectar fraudes y clasificar clientes, según la información entregada por los proveedores de estas herramientas. A nivel mundial el panorama es diferente, puesto que el campo en el cual se demandan más estas herramientas es el CRM (administración de relaciones con los clientes).

4. CONCLUSIONES

Este proyecto de grado muestra nuestro interés de acercar a las empresas colombianas y mas específicamente de la ciudad de Medellín con una serie de tecnologías, conocimientos y metodologías para el tratamiento, manejo y exploración de grandes cantidades de datos; que comercialmente se conoce con el nombre de minería de datos. Esto porque se vio en trabajos desarrollados anteriormente que el estado del arte de la minería de datos en Colombia estaba prácticamente en pañales y que solo habían pilotos de pruebas en algunas grandes empresas.

Una forma de entender mejor los casos de estudio, las mejores practicas, y las herramientas de minería de datos presentadas; es teniendo la información básica de las principales metodologías a seguir en un proyecto de extracción de conocimiento, de las tareas principales que se persiguen, y de los algoritmos que son utilizados para ejecutar esas tareas. Sin esta información le seria muy difícil entender al lector detalles importantes de los casos, de las herramientas y de la misma metodología.

Establecer los principales tipos de tareas de la minería de datos no fue tarea fácil, esto porque no hay un consenso de autores o de los mismos proveedores de las herramientas en cuanto a este tema. Existen muchas diferencias respecto al nombramiento y al concepto general que se tiene de la tarea. La forma como se trato de superar esta dificultad fue estableciendo un total de 4 tareas pero desde un punto de vista más práctico, estas tareas se puede decir que encierran cada uno de los objetivos principales de la minería de datos, por lo cual son muy generales y pueden agrupar otras tareas.

Identificar cuales eran los algoritmos mas usados de acuerdo al tipo de tarea, también se convirtió en una dificultad en su momento. Esto por la gran cantidad de algoritmos existentes, porque además cada algoritmo puede ser utilizado para diferentes tareas, y así mismo como hay unos que son muy generales, hay otros que son muy específicos y solo se usan dependiendo de la tarea, de los tipos de datos, etc. Otro factor que influyo en esto fue que la mayoría de los algoritmos tenía variaciones que lo particularizaban, por lo cual la lista cada vez era mayor y también había que tener en cuenta aquellos algoritmos que eran propiedad de alguno de los proveedores. La forma como se trato de subsanar esto fue tomando los algoritmos libres o estándares y con la ayuda de un experto tratar de clasificarlos según la tarea.

Conocer cuales son las principales tareas de minería de datos es importante para poder estudiar las metodologías, sobre todo con la CRISP-DM pues el desarrollo de un buen proyecto de minería de datos utilizando esta metodología, tiene que ver con que tan bien se defina un problema de negocio en términos de una tarea de minería.

La ventaja de seguir una metodología radica en que establece una secuencia lógica a seguir para desarrollar proyectos de minería de datos de la mejor forma posible. Establecer cual de las existentes es la mejor seria difícil porque cada una tiene sus ventajas y falencias; por lo cual no se puede dar una clasificación muy certera de cual es la metodología principal. Aunque cabe resaltar que en esta tesis se estudia más a fondo y se utiliza como referencia y guía la que se considera como el estándar mundial para esta clase de proyectos, y no esta sujeta a una herramienta en particular, contrario a lo que pasa con otras.

Los proyectos de minería de datos no se desarrollan siguiendo etapas que van una detrás de la otra, como si se siguiera una línea recta la cual nos llevara a obtener ciertos resultados. En realidad es más un ciclo el que se sigue, pero este

ciclo se puede ver roto algunas veces, cuando por los mismos motivos del proyecto se tiene que retroceder a etapas anteriores y replantear lo que ya se ha hecho.

Dentro de esas etapas a seguir es muy aconsejable documentar todo lo que se hizo en el proyecto, de esta forma se esta controlando la fuga de conocimiento y se tienen bases para comenzar otros futuros proyectos de este tipo, donde se van a dejar de cometer muchos errores y el rendimiento de estos será muy superior al anterior o anteriores.

Conocer las diferentes herramientas que existen de minería de datos también es de vital importancia a la hora de comenzar a estudiar la posibilidad de desarrollar un proyecto de extracción de conocimiento, ya que escoger la herramienta apropiada permitiría obtener mejores resultados.

Construir un portafolio con todas las soluciones ICT de minería de datos existentes es algo complicado, ya que encontrar la información necesaria para hacerlo, no es una tarea fácil. Información que incluye especificaciones técnicas, soporte, precios, etc. Por lo cual esta tesis muestra un resumen de las principales herramientas, de acuerdo a ciertos parámetros.

La principal dificultad para construir el portafolio fue conseguir la información referente a cada herramienta y proveedor; ya que algunos de estos proveedores consideran que compartir este tipo de información al público en general, podría ser algo que aprovecharía la competencia para conocer las debilidades que se tienen. La información más difícil de conseguir fue la referente a los precios y solo se pudieron establecer rangos de precios en los cuales se encontraban las herramientas.

Una forma de apoyar el portafolio pudo haber sido probando algunas de las herramientas principales y presentar una valoración de los resultados obtenidos, de forma que el lector tenga puntos vistas teóricos y prácticos.

La existencia de una gran cantidad de proveedores y herramientas de minería de datos, puede conllevar al equipo encargado de conseguir este material, a cometer errores y adquirir la herramienta equivocada, y según algunas de las fuentes consultadas esto traería como consecuencias: perdida de tiempo y dinero de la organización; y el incremento del riesgo de no cumplir con el objetivo propuesto.

Se deben tomar ciertas medidas que guíen a los grupos o personas encargadas de contar con un método para escoger la mejor solución ICT de minería de datos de acuerdo a la evaluación de ciertos criterios. De esta forma se minimiza el riesgo de perder tiempo, dinero y no cumplir con las metas propuestas.

Dentro de los resultados encontrados sobre las herramientas analizadas, se puede observar como el mercado de esta tecnología a nivel mundial se encuentra apuntando principalmente al sector comercial, enfocándose con gran empeño en el manejo de las relaciones con el cliente (CRM), mientras que a nivel local, las expectativas de los proveedores de estas tecnologías, por lo menos en los próximos 5 años es enfocarse en el sector financiero, pues es uno de los sectores que presenta una gran aceptación y reconocimiento de estas herramientas para las labores y procesos cotidianos de sus organizaciones. Mostrando gran interés en ayudas para la clasificación de sus clientes y para la detección de fraudes o procesos sospechosos.

Aunque en investigaciones realizadas anteriormente sobre el estado del arte de la Minería de Datos mostró que este tema se encontraba en una fase de inicial de reconocimiento de las herramientas en Colombia, en esta investigación se puede observar como algunas empresas y organizaciones colombianas ahora se

encuentran utilizando estas herramientas dentro de sus procesos regulares. En total se pueden contar 5 procesos de implementación de una herramienta de minería de datos de manera exitosa y 1 caso que se encuentra en estos momentos en proceso de implementación. Lo cual demuestra un avance considerable en este tema dentro de las empresas colombianas.

Este portafolio puede tener una vida útil de por lo menos un año y medio, debido a que en la campo tecnológico es muy difícil mantenerse por los avances que salen a diario y mas en este campo. Por lo que se aconseja a quien quiera seguir ampliando este trabajo sobre las herramientas, que utilice información muy actualizada para que las empresas tengan un mayor conocimiento de este tipo de herramientas y escojan la que mejor se adapte a sus necesidades.

Los casos presentados de aplicación de la minería de datos en diferentes campos de la industria siguen una misma estructura, que va de acuerdo a la metodología CRISP-DM aunque no con todo el detalle que esta exige.

Se puede ver una cierta similitud en el desarrollo de cada una de las fases en los distintos casos. Se siguen más o menos los mismos procedimientos para desarrollarlas. Esto se puede ver mas fácilmente con las fases de comprensión y preparación de los datos, en las cuales se enfatiza en el uso de la estadística descriptiva para conocer los datos y como pueden estos influir en el desarrollo de las etapas posteriores.

Como ya hemos mencionado anteriormente la principal dificultad para armar los casos, fue contar con la información necesaria para hacerlo. Al parecer es información que las empresas guardan con recelo y no están muy dispuestas a compartirla fácilmente por lo que esta representa. Pero se trataron de construir de la mejor forma posible estos casos que se mostraron, teniendo como base fuentes de gran credibilidad.

El primer caso presenta un proyecto desarrollado por un hospital de Israel en donde se buscaba sacar perfiles de pacientes con trombosis. El resultado de este estudio fue el descubrimiento de dos segmentos en los cuales se descubrieron otras enfermedades presentes que podían ser tratadas para evitar un mayor grado de enfermedad en los pacientes. Este proyecto permitió aplicaciones con otras enfermedades.

En el segundo caso se hace una clasificación de los clientes de una cadena de ropa para predecir la respuesta que tendría una campaña de mercadeo utilizando correo directo. El resultado fue un modelo que incrementaba en aproximadamente el 21% las ganancias de la empresa, además se pudieron mejorar las relaciones con los clientes y el sistema encargado de esto.

El siguiente caso se dio en un banco canadiense donde se quería mejorar el esquema de segmentación de los clientes de la entidad. El resultado fue un nuevo conjunto de grupos de clientes que reflejan mejor el orden natural del mercado, lo que permitió enfilar campañas específicas de mercadeo y ventas a cada grupo, tratando de conseguir el mayor beneficio en términos económicos.

El cuarto caso ilustra una situación bastante particular en las cadenas de supermercados. Se trata del análisis de las cestas de mercado de los clientes de estas cadenas. Este estudio permitió descubrir la relaciones de consumo de los principales productos de estas tiendas, y con esta información desarrollar estrategias que incrementaran las ganancias. La principal estrategia es desarrollar formas de dar recomendaciones a los clientes sobre productos que pueden llevar de acuerdo a su comportamiento a nivel de compras.

El último caso estudia el comportamiento de compras de los clientes de una compañía que vende mercancía por correo, con el fin de saber si es un cliente que solo hace un pedido o un cliente que realiza mas pedidos. El modelo resultante

permitió conocer los clientes más fieles y mejorar el sistema de manejo de las relaciones con el cliente (CRM)
Se puede ver que la mayoría de los casos son aplicados en el campo comercial, aunque también se mostró una aplicación en el campo de la salud. La misma tendencia se pudo observar a nivel mundial, donde los principales casos de aplicación se dan en el sector financiero y en el comercial.

Si bien, los casos de aplicación de minería de datos que estamos documentando; no se ajustaron por completo a la realidad (los nombres de las empresas son ficticios; no se supo a ciencia cierta cual fue la manera como se construyeron los casos en las empresas que se utilizaron como referencia para la construcción y definición del caso, y además la etapa de despliegue fue una iniciativa propia y no corresponde a la realidad de lo que sucedió en la empresa tomada como base). Se puede considerar que estos casos representan y proyectan de una forma correcta la documentación y descripción de un caso de este tipo a través de todas sus etapas. Ya que se siguieron estándares definidos y adoptados por grandes empresas para el desarrollo de proyectos de minería de datos; y porque se hizo un arduo trabajo de investigación que permitirá al lector comprender y darse una idea de las tareas a realizarse en cada una de las etapas de un proyecto de minería, desde una perspectiva no tan teórica sino mas bien desde la practica.

El hecho de no contar con casos bien documentados con los que se pudiera dar fe de que realmente fueron implementados en las empresas, limito al momento de dar una valoración del estado de éxito del proyecto como tal. No se tenía la información necesaria para evaluar estos resultados. No se sabia por ejemplo cuales eran los costo del proyecto, las ganancias que dejo el proyecto, en que áreas o que sucursales se aplicaron o se utilizaron los resultados obtenidos, como afecto la realización del proyecto la empresa, las relaciones con el cliente, la forma de trabajar, etc. Por esto se considera que esta fue una de las falencias o limitaciones que se tuvieron en el desarrollo de esta parte del proyecto.

Un complemento a este trabajo, seria desarrollar un proyecto de minería de datos de prueba, pasando por cada una de las etapas de la metodología CRISP-DM con sus respectivas tareas y resultados. De forma que el trabajo teórico que estamos presentando acá, sea complementado con la práctica. Logrando obtener una mayor credibilidad ante el lector y ante las empresas interesadas en este tipo de proyectos y que antes quieren empaparse un poco sobre el tema.

La minería de datos es un campo que muestra grandes avances cada día, donde el numero de empresas que están iniciando proyectos de este tipo se incrementa cada vez mas, buscando obtener ventajas competitivas frente a sus mas cercanos competidores, y apoyando el sistema de apoyo a las decisiones.

Es importante para las empresas colombianas tratar de comenzar a implementar proyectos de extracción de conocimiento, que les permitan explotar las ventajas que dan estos y competir internacionalmente al mismo nivel de otras grandes empresas. Esto es de vital importancia sobre todo cuando se avecinan tratados de libre comercio con otros países que cuentan con empresas de gran renombre en el mundo.

5. BIBLIOGRAFIA

[Met04] MetaSpectrum MetaGroup. Data Mining Tools Evaluation. 2004. Consultado el 28 de Julio de 2006.

[Cle05] Clemetine, Descubra el Poder Predictivo de sus Datos. White paper, Enero 2005.

[Her04] Hernández, José. Introducción a la Minería de Datos. 2004. Prentice Hall

[Sas05] SAS, SAS Enterprise Miner 5.2 Specifications. 2005. Consultado el 21 de Septiembre de 2006. Disponible en: http://www.sas.com/technologies/analytics/datamining/miner/factsheet.pdf

[Ber05] Berger Charlie, Oracle Data Mining White Paper. 2005. Consultado el 21 de septiembre de 2006. Disponible en: http://www.oracle.com/technology/products/bi/odm/pdf/bwp_db_odm_10gr2_0905.pdf.

[Blo04] Bloor Reserch. Clever Path Portal from Computer associates. 2004. Consultado el 30 de Septiembre de 2006. Disponible en: http://www.bloorresearch.com/research/product_evaluation

[Men00] Mendoca, Manoel. University of Maryland. Mining Software. Engineering Data: a Survey. Consultado el 30 de Septiembre de 2006. Disponible en: http://www.thedacs.com/techs/datamining/

[Ang03] Angus, Jeff. InfoWorld. KXEN Analytic Framework 3.0.2 analysis. Consultado el 2 de octubre de 2006. Disponible en: http://www.infoworld.com/KXEN_Analytic_Framework_3.0.2/product_48211.html?view=0&curNodeId=0

[Her06] Hershel, Gareth. GARTNER, Magic Quadrant for Customers Data Mining 1Q06. Enero de 2006. Consultado el 10 de Septiembre de 2006. Disponible en: http://www.spss.com/pdfs/spss_magicquadrant.pdf#search=%22Magic%20Quadrant%20for%20Customer%22

[Par01] Parr Rud Olivia. Data Mining Cookbook Modeling Data for Marketing, Risk, and Customer Relationship Management. 2001.

[IBM04] IBM, Manual de Referencia de Administración y Programación de DB2 Intelligent Miner v 8.2". Consultado el 15 de octubre de 2006.

[Gon04] "Metodología CRISP-DM vs. SEMMA". Jose Emilio Gondar Nores. Data Mining Institute S.L. Consultado el 10 de septiembre de 2006. Disponible en: http://www.estadistico.com/arts.html?20040426

[Ins06] Insightful Miner Features". Insightful Corporation. 2006. Consultado 26 de Septiembre de 2006. Disponible en: http://www.insightful.com/products/iminer/features.asp

[MiT02] Miller Tim, Tate Brian. Discovering Teradata Warehouse Miner. Diciembre 2002. Disponible en: http://www.teradata.com/t/pdf.aspx?a=83673&b=86841

[BAB+01] Baragoin Corinne, Andersen Christian M, Bayer Stephan, Bent Graham, Lee Jieum, Schommer Christoph. Mining Your Own Business in Health Care Using DB2 Intelligent Miner for Data. 2001.

[BrM05] Bruera Maria del Rosario, Martínez Néstor. Business Intelligence. Consultada el 17 de agosto de 2006, Disponible en: http://web2.ufasta.edu.ar/extensionfastafi/050609%20MRBruera%20NMartinez%20TSIN/BI_FASTA_3.ppt

[Lar06] Larose Daniel T. Data Mining Methods and Models. Published by john Wiley and sons, Inc.

[SPS06] Data mining with Clementine for smarter retailing. Consultado el 21 de marzo de 2006. Disponible en: http://www.spss.com/downloads/Papers.cfm?List=all&Name=all

[ESP03] Elliott Kenneth, Scionti Richard, Page Mike. The Confluence of Data Mining and Market Research for Smarter CRM. 2003. Consultado el 21 de marzo de 2006. Disponible en: http://www.spss.com/downloads/Papers.cfm?prod_familyID=00007&Name=Market_Research

[SLB+01] Schommer Christoph, Lee Jieum, Bent Graham, Bayer Stephan, Andersen Christian M, Baragoin Corinne. Mining Your Own Business in Banking Using DB2 Intelligent Miner for Data. 2001.

[Giu03] Giudici Paolo. Applied Data Mining Statistical Methods for Business and Industry.2003

[BeL04] Berry Michael J.A, Linoff Gordon S. Data Mining Techniques For Marketing, Sales, and Customer Relationship Management Second Edition.

[Lar05] Larose Daniel T. Discovering Knowledge in Data, An Introduction to Data Mining. 2005.

[SBA+01] Mining Your Own Business in Retail Using DB2 Intelligent Miner for Data. Schommer Christoph, Baragoin Corinne, Andersen Christian M, Bayer Stephan, Bent Graham, Lee Jieum. 2001.

[Mon02] Montano Moreno, Juan José. Redes Neuronales Artificiales aplicadas al Análisis de Datos. Montaño Moreno Juan José.2002. Consultado el 23 de septiembre de 2006, disponible en: http://www.tdx.cesca.es/TESIS_UIB/AVAILABLE/TDX-0713104-100204//tjjmm1de1.pdf

[BrG04] Britos Paola, García Martínez Ramón. SELECCIÓN DE HERRAMIENTAS DE EXPLOTACION DE DATOS. UNA PROPUESTA METODOLÓGICA. Consultado el 09 de octubre de 2006. Disponible en: http://www.itba.edu.ar/capis/rtis/rtis-6-2/seleccion-de-herramientas.pdf.

[Luc06] De Luca Venegas Mauricio Pascual. Plan para enfocar las campañas bancarias utilizando data mining. 2006.

[DEA02] Daedalux, Data Mining White paper. 2002. Consultado el 17 de Marzo de 2006. Disponible en: http://www.daedalus.es/AreasMD-E.php

[DEA06] Daedalux, Algoritmos de Extracción de Conocimiento. 2006. consultado el 23 de marzo de 2006. disponible en: http://www.daedalus.es/AreasMDFase3-E.php

[CCK+00] Chapman Pete (NCR), Clinton Julian (SPSS), Kerber Randy (NCR). Crisp-DM 1.0 Step by Step Data mining Guide. 2000. consultado el 20 de Marzo de 2006.

[Ede00] Edelstein Herb. Building profitable customer relationships with data mining. Consultado el 17 de Marzo de 2006. Disponible en: http://www.spss.com/registration/premium/consol056.cfm?WP_ID=60

[KRM03] Elliott Kenneth, Scionti Richard, Page Mike. The Confluence of Data Mining and Market Research for Smarter CRM. Consultado el 21 de Marzo de 2006. Disponible en: http://www.spss.com/registration/premium/consol056.cfm?WP_ID=133

[BOM01] Bank of Montreal becomes master of its destiny with IBM scoring tool. Consultado 6 de marzo de 2006. Disponible en: www.adastracorp.com/our_customers/BMO_Award2001_Success_story.pdf

[SQL05] Microsoft SQL Server 2005 data mining helps Clalit preserve health and save lives. Consultado el 10 de marzo de 2006, Disponible en: http://members.microsoft.com/CustomerEvidence/Search/EvidenceDetails.aspx

[Her05] Hernandez Orallo, Jose. Mineria de Datos. 2005. consultado el 10 de Julio de 2006. Disponible en: http://www.dsic.upv.es/~jorallo/cursoDWDM/dwdm-III-1.pdf

[CRM05] CRM Customer Relationship Management con SPSS. consultado el 17 de Noviembre de 2006. disponible en: http://www.spss.com/la/soluciones/crm.htm