28

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES "ACATLÁN"

APLICACIÓN DE UN MÉTODO DE MINERÍA DE DATOS

SEMINARIO TALLER EXTRACURRICULAR

QUE PARA OBTENER EL TITULO DE

LICENCIADO EN MATEMÁTICAS APLICADAS Y COMPUTACIÓN

PRESENTA

ELIZABETH SUSANA MIRANDA HERNÁNDEZ

Asesor: **JUAN TORRES LOVERA**

Fecha: Octubre, 2002

CONTENIDO

Introducción

Históricamente, la mayoría de los datos operacionales han sido generados o coleccionados para propuestas de investigación, actualmente, los negocios empresariales tienen una cantidad masiva de datos, sin embargo, no han sido generados con propósito de análisis. El problema hoy en día de la mayoría de las empresas es que a pesar de que cuentan con una cantidad de datos que excede a terabytes (1,024 Gigabytes) de información, no es explotada adecuadamente enfocándose en los objetivos de negocio.

La minería de datos, que para fines de la presente investigación se entiende como el proceso de seleccionar, explorar y modelar grandes volúmenes de datos para descubrir patrones de conducta antes desconocidos, es una herramienta de explotación de información que ayuda a las empresas a generar conocimiento de los datos, haciendo uso de técnicas sofisticadas de estadísticas e inteligencia artificial, como son: las regresiones logísticas, árboles de decisión y redes neuronales, para crear un modelo del mundo real, basado en datos recolectados, por ejemplo: referencias personales, nivel de ingresos de los clientes, información crediticia, etc., los cuales son obtenidos de una gran variedad de fuentes como son departamentos de crédito, INEGI y bancos de datos. Este modelo genera patrones en la información que pueden apoyar a la toma de decisiones –proceso de pensamiento que ocupa toda la actividad que tiene por fin solucionar problemas– y predecir nuevas oportunidades de negocio.

Un giro importante de empresas dentro de la actividad diaria de las personas, que se encarga de la venta de bienes y servicios, son las dedicadas a las ventas a detalle, mejor conocidas a lo largo del mercado con el nombre de *"retail"*, las cuales ofrecen sus productos a consumidores, siempre a través de un mostrador, o un lugar físico específico o también, por medio de catálogos, puerta a puerta, por teléfono, máquinas (de refrescos, dulces, galletas, etc.) y a través de Internet.

El deseo principal de toda empresa enfocada a la venta de bienes y servicios es conocer a sus clientes potenciales, y así, mejorar sus estrategias de mercado y tener información de las necesidades y preferencias de estos, por anticipado, y así, ofrecer el mejor servicio posible para mantener una relación de excelencia con el cliente que proporcione un beneficio mutuo. Por un lado la satisfacción de las necesidades del cliente y por el otro incrementar los ingresos de la empresa, ya que el cliente estará conforme con la compra realizada y los proveedores por los ingresos obtenidos.

De acuerdo con lo anterior, este trabajo esta enfocado a responder a la pregunta de negocio de las empresas dedicadas a las ventas a detalle (EDVD), que es: ¿Quiénes son mis clientes potenciales?, y tiene el objetivo de establecer una guía que detecte a dichos clientes, al aplicar un método de minería de datos (MD) administrada por un sistema para la toma de decisiones (SAS), con el fin, de generar mayores ingresos a dichas empresas.

La minería de datos (MD) es el resultado de un largo proceso de investigación de muchos autores y desarrollo de herramientas. El inicio de esta evolución comienza en el momento de depositar datos en computadoras y acumular información para su posterior explotación, más adelante, se mejoró el acceso a datos al apoyar la extracción, transformación y almacenamiento de la información en diferentes bases de datos, y actualmente la creación de tecnologías permite al usuario navegar en sus datos en línea a través de Internet.

La minería de datos (MD) esta lista para su aplicación en los negocios empresariales porque esta soportada por tres tecnologías suficientemente maduras:

- Colección masiva de datos
- Poder de multiprocesamiento de la información
- Algoritmos de minería de datos

En el transcurso de la presente investigación se define y amplia el concepto de minería de datos (MD), y se hace mención de la herramienta del sistema SAS que provee de las técnicas avanzadas de la MD.

Esta investigación fue impulsada por la necesidad de ayudar al consultor de MD a ejercer su trabajo de manera eficiente, guiándolo para la aplicación de un proyecto enfocado a las EDVD, y de esta forma no pierda tiempo en consultas a múltiples fuentes de información de MD.

El capítulo I de esta investigación, inicia con una descripción general de las EDVD para que el lector se familiarice con el giro de la empresa, continua con la descripción del sistema SAS el cual ayuda a la toma de decisiones, posteriormente se explica el concepto y las técnicas de MD, la herramienta del sistema SAS que se utilizará en el transcurso de este trabajo y aunque se utiliza un sistema específico, también se presenta un comparativo de las diferentes herramientas de MD en el mercado, que pueden servir para este fin.

En el capitulo II se detallan los pasos a seguir en la planeación y ejecución de un proyecto de MD, que guiara al consultor desde la definición del problema de negocio hasta la validación y revisión de los beneficios del proyecto, también se presenta un caso práctico dirigido a las EDVD que sirve para ilustrar la aplicación del método de MD.

Finalmente en el último capítulo se concluye con la exposición de los resultados obtenidos y se explica la relación de estos con la toma de decisiones, se puntualizan los alcances y limitaciones del mismo, las tendencias de MD y se presentan los escenarios posibles en un futuro próximo.

Es indudable, que el conocimiento que generan los datos almacenados es vital para cualquier empresa, las cuales compiten en el mercado muchas veces por ofrecer la mejor oferta al mejor cliente, en el tiempo y canal indicado. De acuerdo con esto, la pregunta obligada es ¿cómo hacerlo? los cuestionamientos automáticos que se deben conocer son:

- ¿Quiénes son mis clientes potenciales a quien se deben dirigir mis ofertas?
- ¿Qué características tienen mis clientes potenciales?, que ayudarán a definir segmentos de poblaciones
- ¿Qué preferencias tanto de productos y copras tienen mis clientes?

Este trabajo da respuesta exclusivamente a la primera pregunta y deja asentadas las bases para que el lector pueda con cierta facilidad contestar las interrogantes faltantes, ya que dar repuesta a las demás preguntas podría ser desarrollado por investigaciones posteriores.

Así pues, la presente investigación, da respuesta a uno de los principales problemas que atañen a los negocios empresariales al presentar un caso práctico que se puede poner en producción en las EDVD si es que el negocio cuenta con los elementos aquí expuestos y orientará al consultor para la aplicación de cualquier proyecto de MD; además de presentar al interesado una de las principales herramientas de MD que ayuda a la toma de decisiones de diferentes áreas de negocio.

Capitulo I

Conceptos generales del sistema para la toma de decisiones (SAS) y la minería de datos.

Objetivo

Introducir una herramienta que permita el análisis del perfil del consumidor que parte de describir y definir a las empresas dedicadas a las ventas a detalle, el sistema para la toma de decisiones (SAS) y la minería de datos.

1.1. Entorno y descripción de las empresas dedicadas a las ventas a detalle.

En este tema se presentará una perspectiva histórica de las empresas dedicadas a las ventas a detalle (EDVD), y se describirá el giro de estas empresas (ya que el presente trabajo esta enfocado a estas y se pretende explicar a las EDVD) para que el lector se familiarice con este término ya que "las ventas a detalle es una parte vital de nuestro modo de vida"[1], pues cualquier de nosotros ha estado involucrado con éstas la mayor parte de su existencia. Es por eso que se decidió enfocar el trabajo a una empresa de este giro, ya que la mayoría de productos y servicios llegan a nuestras manos por medio de empresas dedicadas a las ventas a detalle (EDVD).

Las ventas a detalle(VD), menciona Rondal W. Hasty son un mundo de nuevos y útiles productos y servicios que mejoran la vida de las personas haciéndola mas confortable, feliz o rica y algunas veces pobre. Es posible menciona, que una de las primeras adquisiciones que todo mundo haya hecho, fuera en una EDVD, al comprar dulces, juguetes o regalos.

"En nuestra cambiante sociedad las necesidades y deseos de los consumidores no permanecen inmutables. Tecnología y estilos de vida, moda y muchos otros factores afectan las necesidades y deseos de los consumidores"[2] , por esto, las EDVD siguen siendo parte imprescindible en nuestra sociedad. Actualmente, es difícil que pase un día sin que alguna persona entre a una EDVD de cualquier clase.

A continuación se iniciará con la perspectiva histórica de las EDVD, para entender el surgimiento de estas y analizar porque hoy en día son vitales en nuestro mundo cotidiano.

Perspectiva histórica

La mayoría de las EDVD están asociadas con la venta de bienes en tiendas de almacenamiento, estos almacenamientos pueden ser de diferentes tipos: Tiendas de productos específicos, Tiendas departamentales, Supermercados, Tiendas de descuento, etc.
Estas tiendas ofrecen bienes a consumidores siempre a través de un mostrador, o un lugar físico específico. Sin embargo, existen otros tipos de ventas que son las ventas de bienes a través catálogos, puerta a puerta, ventas por teléfono y venta por medio de máquinas (maquinas de refrescos, dulces, galletas, etc.), los bienes que son adquiridos a través de estas formas también son considerados según Ronald W.Hasty ventas a detalle, aunque no exista un contacto físico con el consumidor.

No importa la visualización que tengamos de las ventas a detalle (VD), el mundo del comercio es dinámico, y es un mundo de productos e ideas que contribuyen a la belleza y satisfacción del cliente. Sin embargo, las VD nunca fueron consideradas algo honorable, en algunas sociedades, las VD fueron toleradas como un mal necesario.
Históricamente en los imperios Helénicos y Romanos, la sociedad despreciaba las ocupaciones enfocadas al comercio, la sociedad se enfocaba en aspectos intelectuales más que en cuestiones materiales. Solamente esclavos o miembros de grupos minoritarios se desempeñaban en funciones de VD.

---

[1] Ronald W. Hasty, "Retailing", Harper & Row, Publishers, Tercera Edición, New York, 1983.
[2] Warren G Meyer, James R Stone III, Donald P. Kohonx, E.Edward Harris, "Marketing, Ventas al por menor, para empleados, gerentes y empresarios", Edit. McGraw-Hill, Octava Edición, 1991.

El punto de partida de las VD, considera Ronald W. Hasty es el hecho de que las personas dejaran el estilo de vida de autosuficiencia que llevaban, esto es, satisfacer sus necesidades por ellos mismos a través del cultivo, criaderos de especies, etc. y emigraran hacia la ciudad, por lo que las personas se volvieron más dependientes de otras para conseguir comida, ropa y demás. La especialización de las personas en diferentes labores se acelero y esto influencio el crecimiento de las VD.

"Las VD es una institución que se encuentra mas cerca y en contacto con el consumidor, por lo que es el mercado más capaz que puede interpretar sus necesidades".[3] Las personas encargadas de las VD son muy hábiles para interpretar las necesidades del consumidor, por lo que proporcionan variedad en el mercado sobre estilos, materiales, colores, precios, tallas, etc. y los presentan de manera efectiva, para que el consumidor pueda adquirirlo de forma sencilla y atractiva.

Empresas dedicadas a las ventas a detalle

En esta era en la que vivimos, rica en productos y servicios (Productos que son bienes cultivados o manufacturados y disponibles para la venta; y servicios que son beneficios o satisfacciones que mejoran la apariencia, salud, comodidad o paz mental de los usuarios), la mayoría de estos productos y servicios son adquiridos a través de vendedores o comerciantes minoristas.

Warren G. Mekyer y James R. Stone afirman que los negocios tienen dos funciones básicas: la producción de bienes y servicios y la fijación y distribución de bienes y servicios. La producción definen, se refiere a la creación, cultivo, procesamiento o manufactura. Después de que un producto ha sido manufacturado o un servicio se ha hecho disponible, debe ser comprado o vendido en el mercado. Un mercado es cualquier punto en el cual la propiedad cambia de manos. *Marketing* según la Asociación Norteamericana de *Marketing* es el proceso de planear y ejecutar la concepción, la fijación de precios, la promoción y la distribución de ideas, bienes y servicios para crear intercambios que satisfagan los objetivos individuales y organizacionales. El proceso de *marketing* se inicia con la idea de un producto o servicio en la mente de un comerciante y termina con los clientes satisfechos.

Las EDVD o el *marketing* minorista o también conocido en el ambiente de negocio por su nombre en inglés de *retailing* es solamente según Warren G. Mekyer y James R. Stone un tipo de mercado que empieza con el productor y termina con el consumidor. Las EDVD incluyen todas las funciones involucradas en vender (o alquilar) bienes y servicios a usuarios finales, incluyendo hogares, individuos, y otros que compran bienes y servicios de consumo final. Las actividades de VD incluyen compra de bienes y servicios para la venta, almacenaje exhibición, marcada de precios, publicidad, venta, financiación, servicio y otras actividades necesarias para completar la venta con los compradores. Las EDVD según Warren G. Mekyer y James R. Stone, ocupan el lugar en el canal de distribución que conecta con el consumidor. La metas de las EDVD son lograr una utilidad al servir las necesidades y deseos que tienen los consumidores de productos y servicios. Por tanto, la clave para lograr utilidades en las ventas al por menor reside en la habilidad del minorista para servir a grupos o consumidores objetivo.

El presente trabajo muestra una forma basada en modelos matemáticos para encontrar clientes potenciales o población "objetivo" al utilizar herramientas computacionales que ayuden a las EDVD a generar mayores utilidades y sirvan para satisfacer las necesidades del consumidor o cliente.

[1] Ronald W. Hasty, "Retailing", Harper & Row, Publishers, Tercera Edición, New York, 1983.

"Las EDVD exitosas seleccionan la clase de clientes a los cuales desean servir, y estos consumidores llegan a ser la población objetivo, las personas para las cuales compran mercancía o diseñan servicios. Identificar una población objetivo para servirla es crítico ya que ninguna EDVD, aún compañías gigantes, pueden servir efectivamente a toda clase de consumidores y obtener una utilidad".[2]

La diferencia entre las empresas dedicadas al mercado mayorista y las EDVD es que el mayorista se efectúa por medio de negocios que compran productos, generalmente en grandes cantidades, y los venden usualmente en pequeñas cantidades a EDVD o usuarios industriales y otros negocios en lugar de consumidores. Sin embargo, muchos mayoristas también venden algunos de sus productos a consumidores. La clasificación de mayoristas o minoristas depende de cuál tipo de comprador cuenta con más de la mitad de las ventas del vendedor.

La perspectiva histórica de las EDVD y la definición de esta nos dieron un panorama general de las empresas de este giro, solo falta definir el proceso que las EDVD llevan a cabo dentro del entorno mercado-consumo, que nos ayudará a definir en cuales de estos entra el presente trabajo.

Las VD pueden considerarse en un ciclo que empieza con la identificación de las necesidades y deseos de los consumidores y termina asegurándose de que los consumidores estén satisfechos, a continuación se presentará una breve descripción del ciclo de las EDVD, que ayudará a entender cómo operan estas .

Ciclo de EDVD

Cada uno de los cinco elementos del ciclo de las EDVD definidos por Warren G. Mekyer y James R. Stone, sirven para conocer el proceso que se lleva a cabo para la venta de algún bien o servicio.

1. Identificación de las necesidades y deseos de la población objetivo. ¿Qué clases de productos y servicios se desean y requieren?, ¿Qué tamaños, qué formas y cuántos de cada uno?, ¿Cuándo y dónde se necesitan?

2. Determinación del precio que se va a cobrar. ¿Están los clientes deseosos y capaces de pagar lo suficiente para cubrir los costos hechos por la persona de la EDVD más una utilidad justa?

3. Información a los consumidores acerca de los productos y servicios. ¿Dónde deben anunciarse?, ¿Se necesita la venta personal?, ¿Cómo podría exhibirse el producto o servicio?

4. Entrega de los bienes del productor al consumidor. ¿Qué medios de transporte se deberían usar?, ¿Cómo se manejaría la parte financiera? Con respecto a servicios, ¿deberían venir los clientes al centro de servicio o debería visitarlos el proveedor?

5. Asegurarse de que los clientes estén satisfechos. ¿El producto o servicio realmente llena las necesidades o deseos propuestos?, ¿Indican las experiencias de los consumidores qué cambios se deberían hacer?

[2] Warren G Meyer, James R Stone III, Donald P. Kohonx, E.Edward Harris, "Marketing, Ventas al por menor, para empleados, gerentes y empresarios", Edit. McGraw-Hill, Octava Edición, 1991.

En la presente investigación solamente se identifica a la población objetivo sin entrar a detalle de los productos o servicios que estos deseen, ya que esto podrá ser tema de una investigación posterior.

En la sociedad en la que nos encontramos con cambios constantes en tecnología, estilos de vida, moda y muchos otros factores que afectan las necesidades y deseos de los consumidores, las ventas a detalle deben ser y son dinámicas, lo cual estimula a los comerciantes progresistas, siendo así, la clave del éxito la flexibilidad y el deseo de cambiar con el tiempo.

Es importante mencionar los cambios que se presentan en un futuro próximo, al referirnos a las EDVD, ya que existen factores variables que se deben considerar en todo negocio de las VD, pues, se conviene analizar la nueva información de entrada del proceso de venta, por tanto, se presenta a continuación las tendencias de las EDVD.

Tendencias de las EDVD

La revolución tecnológica ha traído muchos artículos electrónicos que ayudan a las EDVD, un ejemplo de estos son los sistemas computacionales que ayudan en las cajas registradoras que procesan transacciones de tarjetas de crédito, facturas, y mantienen registros de control de inventarios.
Estos sistemas se usan actualmente a lo ancho y largo del mercado, en restaurantes y tiendas de abarrotes del vecindario, hasta en las grandes tiendas de renombre. Estos sistemas hacen que las personas de las EDVD puedan identificar y servir segmentos específicos del mercado obteniendo su utilidad.
El crecimiento de las compras electrónicas, que han evolucionado casi paralelamente a la de las computadoras, traen profundos cambios a las EDVD, ya que deben encarar este avance tecnológico, al usar una combinación del concepto de almacenes, catálogos y servicios de compras electrónicas, actualmente conocido con el nombre de tiendas virtuales; que son las ventas de diferentes productos a través de Internet en las cuales se realizan transacciones de compraventa por medios electrónicos.
Así, lo menciona Warren G. Mekyer y James R. Stone, los consumidores podrán firmar en una computadora, indiferentemente desde sus casas o desde un quiosco especial en su almacén favorito. De igual forma que el teléfono revoluciono al mundo entero a principios de siglo, la transferencia electrónica parece ser una de las innovaciones que llegarán al uso común de los consumidores.

Es indudable que las EDVD deben actualizarse no tan solo en el proceso de compra-venta de producto, si no también en herramientas que faciliten y ayuden al proceso de identificación de clientes objetivo, por lo que en el siguiente tema se presentará una de las herramientas de utilidad para cualquier empresa dedicada no tan solo a las VD si no de cualquier enfoque de negocio que requiera utilizar modelos matemáticos de selección e identificación de población objetivo, al hacer uso del sistema SAS que es una herramienta robusta y sencilla de manejar.

1.2 Introducción al sistema SAS

Una vez que se describió a las empresas dedicadas a las ventas a detalle (EDVD), se debe tener conocimiento de la herramienta que se va a manejar para el análisis de datos, lo cual es caso de estudio del presente trabajo.
La herramienta que se maneja en este trabajo tiene por nombre SAS, y se encuentra distribuida por la empresa del mismo nombre, la cual se describe en los párrafos siguientes, posteriormente se continua con la descripción de la herramienta, además se describen los diferentes productos, lo cual permitirá al lector tener conocimiento sobre lo robusto que es el sistema que se usará a través del presente trabajo.

Historia de la Compañía

A continuación se presenta la historia de la compañía SAS, obtenida del su sitio en Internet.[3]

SAS es una de las 10 empresas privadas más grandes del mundo encargadas de venta de software, con tres millones de usuarios alrededor del mundo y productos de software instalados en mas de 33,000 sites en mas de 110 países, SAS invierte en investigación y desarrollo el 30% de sus ingresos.

SAS fue incorporada en Julio de 1976, por el desarrollo y la puesta en venta del modulo de software Base, SAS estaba originalmente localizada en Raleigh, Carolina del Norte. En 1980 la compañía fue transferida a Cary, Carolina del Norte, en donde reside actualmente su corporativo.

El producto que impulso a la compañía fue el software Base de SAS, que fue desarrollado originalmente para realizar un análisis de datos de agricultura, alojados en un mainframe de IBM, de la Universidad del estado de Carolina del Norte. Con el paso de los años el sistema SAS ha llegado a ser un sistema para la entrega de información, incluyendo mas de 25 módulos integrados, que permiten a las organizaciones un completo control de los datos, en cuanto a acceso, manejo y análisis de datos y presentación de estos, a todos los módulos del sistema SAS, se les antepone la palabra SAS, en el presente trabajo se mencionará el sistema SAS refiriéndose a los módulos en general que lo integran.

SAS con su Arquitectura Multiplataforma la cual fue desarrollada en 1988, permite implementar el sistema SAS en mainframes, mini computadoras, workstation, y PC's, además, permite tener el acceso a la mayoría de los manejadores de bases de datos conocidos que son: Oracle, Informix, Sybase, SQL server, Teradata, etc.

Hoy, SAS desarrolla un conjunto de módulos integrados para la entrega de información que ayudan a la toma de decisiones, el sistema SAS incluye capacidades y herramientas para *data warehouse* incluyendo manejo, organización y explotación del data *warehouse*, incluye herramientas para el análisis multidimensional de la información, herramientas de análisis estadístico, investigación de operaciones, control de calidad y MD.
SAS también desarrolla software para diferentes soluciones de negocio por ejemplo para la consolidación financiera, reporteo, manejo de servicios de tecnologías de información y recursos humanos.

[3] Información obtenida del sitio en Internet de SAS (http://www.sas.com)

El sistema SAS cuenta con diferentes módulos, a continuación se presenta una breve descripción de algunos de estos (información obtenida de la página principal del sistema SAS en Internet[3]):

Base SAS® - Este es el único producto obligado del sistema SAS, puesto que contiene las herramientas requeridas para crear y modificar archivos SAS. Base SAS provee el acceso a los datos ayudándose de otros módulos de SAS, la manipulación de estos, ya que provee un lenguaje de cuarta generación, y soporta a la vez el lenguaje de consulta estructurado (SQL), con el cual podemos realizar consultas y manipulación de la información, cuenta con procedimientos para llevar a cabo el análisis de la información, éstos incluyen técnicas para la sumarización de la información , además de contar con procedimientos para obtener estadística descriptiva e inferencial.

Algunos de los módulos que están enfocados a la visualización de los datos:

SAS/GRAPH® - Es un módulo que provee una variedad de gráficas a color, diagramas y mapas los cuales tienen una presentación, si se desea, con formato en html al utilizar tanto la tecnología java y activex para este fin.

AppDevStudio®.- Provee de una interfase de desarrollo para aplicaciones de negocio enfocadas a la presentación detallada de la información, permite crear estas aplicaciones con tecnología de applets, servlets, JSP, además de aplicaciones con CGI 's.

SAS/INSIGTH®.- Es una herramienta muy interactiva para análisis gráfico de datos que permite al usuario explorar los datos a través de una variedad de representaciones gráficas. Todos los campos o registros están ligados de manera que modificaciones en una gráfica se reflejan inmediatamente en todas las demás.

Módulo para la administración de Datawarehouse:

SAS/Warehouse Administrator® - Es la herramienta por la cual se integran todos los componentes existentes de SAS para la construcción, manejo y organización de Data Warehouses. Provee un único punto de control para implementadores y administradores de múltiples Data Warehouses ó Data Marts.

Herramienta y modulo de SAS para la limpieza de la información:

SAS Data Quality®. Es un modulo que permite analizar, homologar, estandarizar y limpiar la información, y evita la duplicidad de la información haciendo uso de diferentes algoritmos de lógica difusa.

Dataflux DfPower Studio and Match modules®. Es una herramienta que provee unà interfaz gráfica *point-and-clic* que consiste en normalizar datos inconsistentes, además, permite identificar datos duplicados y cercanos a la duplicidad.[4]

Módulos enfocados a estadística:

SAS STAT®. Es un modulo que provee una gran variedad de herramientas estadísticas enfocadas al análisis estadístico de la información.

SAS ETS®. Provee una gran variedad de métodos para el análisis de series de tiempo a través de una interfaz gráfica *point-and-clic*.

[3] Información obtenida del sitio en Internet de SAS (http://www.sas.com) y de la página principal de Dataflux
[4] Información obtenida del sistio en Internet de Dataflux (http://www.dataflux.com)

Herramienta de minería de datos:

SAS Enterprise Miner®. Es una herramienta con una interfaz gráfica que permite realizar MD, al hacer uso de métodos estadísticos y matemáticos como regresiones lineales, redes neuronales, árboles de decisión, entre otros.

Gráfica de los módulos de SAS en secuencia de utilización para la herramienta de Minería de datos

Modulos de limpieza, extracción, transformación y carga	Módulo para el almacenamiento de la información	Modulos para la explotación y análisis de la información
SAS Base	Data warehouse. Administrator	SAS Enterprise Miner
SAS Data Quality		Applets Studio
Dataflux		SAS/STAT
		SAS/ETS
		SAS/GRAPH
		SAS/INSIGHT

Uso del sistema SAS en la minería de datos

La herramienta que se utiliza en el transcurso de los siguientes temas, es SAS Enterprise Miner®, se decidió el uso de esta herramienta por las siguientes características: es capaz de trabajar con una cantidad masiva de datos, provee de un conjunto de herramientas de MD para la preparación y visualización de datos, modelos predictivos, descubrimientos de asociaciones y validación de los modelos. y maneja una solución de principio a fin en el procesamiento de datos y dado que SAS Enterprise Miner es una herramienta integrada con los módulos de SAS, los resultados obtenidos del análisis pueden ser incluidos en cualquier herramienta de explotación ya sea para almacenamiento de los datos o para el análisis multidimensional de estos.

En los posteriores temas se presentará una matriz comparativa de las diferentes herramientas de MD, en la cual se vera la potencialidad de la misma.[5]

[5] SAS Institute, "Finding the Solution to Data Mining, Exploring the features and Componentes of Enterprise Miner , Release 4.1 from SAS", 2001.

1.3 Minería de datos

"Minería de Datos es el proceso de seleccionar, explorar y modelar grandes volúmenes de datos para descubrir patrones de conducta antes desconocidos".[6]

Históricamente, la mayoría de los datos fueron generados o coleccionados solamente para investigación, actualmente, las empresas tienen un cúmulo de datos de las diferentes transacciones u operaciones realizadas los cuales deben ser aprovechados, los especialistas de negocio según David Hand, Heikki Mannila, and Padhraic Smyth buscan tener un acercamiento a la minería de datos(MD) que satisfaga sus necesidades, lo primero que se requiere para este hecho es entender la MD, segundo, la MD debe tener un buen desempeño al hablar de tecnología, esto es, debe contar con un sistema de computo robusto que permita entregar resultados en un tiempo óptimo, y lo más importante, el proceso de MD y los resultados obtenidos deben ser precisos.

La MD se lleva a cabo a través de un análisis, el cual puede ser supervisado y no supervisado, a continuación se presentará una definición de cada uno de estos.

Análisis supervisado

La mayoría de los casos en los que se aplica la MD incluye el uso de una o más variables objetivo. Para la presente investigación la variable objetivo se entiende como una variable binaria, la cual contiene 2 clases que representan presencia o ausencia de algún evento (usualmente se denota 1 para la presencia del evento y 0 para la ausencia. Este tipo de análisis es llamado supervisado o enfocado a un objetivo específico. Los métodos usados en este análisis son los de clasificación y predicción, los cuales utilizan usualmente datos históricos para el desarrollo del modelo.
Una vez que se revisan los resultados, el análisis supervisado propone el desarrollo de nuevos modelos con algunas variaciones o mejoras, se pueden restringir campos o registros y analizar los resultados obtenidos. Este proceso es llamado MD iterativa la cual se basa en realizar el proceso de MD de forma iterativa hasta obtener un satisfactorio, explicativo, robusto y óptimo resultado.

Análisis no supervisado

Las aplicaciones de la MD que no contienen una variable objetivo, son considerados dentro del análisis no supervisado o análisis de datos dirigidos, en este caso no existe una variable objetivo disponible con la cual medir la calidad de las respuestas propuestas. Un ejemplo del análisis no supervisado es cuando se pretende identificar las características de clientes que tienen teléfono celular, un análisis no supervisado produce una serie de grupos de clientes cada uno con diferentes características de clientes, los cuales pueden ser inspeccionados de forma manual para obtener resultados significantes, por ejemplo en algunos de los grupos realizados puede darse el caso de obtener características que sirvan para clasificar al cliente en el plan adecuado de telefonía que más le convenga, de acuerdo a los resultados generados.

Áreas de negocio de aplicación

Una vez que se presentaron los tipos de análisis posibles en la MD, se presentarán algunas aplicaciones de MD, en diferentes áreas de negocio, el objetivo que se persigue en cada caso puede cumplirse haciendo uso de la MD.

[6] David J. Hand, Heikki Mannila, and Padhraic Smyth, "Principles of Data Mining", Massachusetts Institute of Technology, 2001.

Otorgamiento de crédito.- El otorgamiento de crédito se enfoca en extender crédito al solicitante, el objetivo en este caso es anticipar y reducir el incumplimiento del crédito que va a ser otorgado.

Detección de fraude: El objetivo en esta área es descubrir los patrones característicos de fraude deliberado, los bancos utilizan estos patrones para prevenir transacciones fraudulentas de tarjetas de crédito y cheques falsificados, al igual las compañías de seguros utilizan este análisis para identificar abusos en reclamaciones o reclamaciones ficticias.

Telefonía.- En las compañías de telefonía, es importante conocer como minimizar la cancelación de servicio del cliente, a este tipo de análisis es comúnmente denominado Análisis de cancelación o *Churn Análisis*, el cual se basa en un proceso para identificar clientes los cuales tengan tendencia a la deserción, además, incluye la aplicación de conocimiento para identificar las características de estos.

Empresas dedicadas a las ventas a detalle.- En las EDVD uno de los objetivos más importantes es conocer cuáles son los clientes rentables o potenciales, que sirven para diseñar nuevas estrategias de mercadotecnia enfocadas a estos.

1.3.1 Técnicas para la minería de datos

Para llevar a cabo el proceso de MD es necesario conocer algunas de las técnicas que se utilizan a lo largo del proceso, ya que el presente trabajo hará mención de estas.

Visualización de la información

La visualización de datos, es una de las herramientas más versátiles de la MD, y es simplemente la representación gráfica de los datos. El proceso de representación de datos gráficamente es usado, hoy en día, en la mayoría de las herramientas de consulta, ya que nos ayuda a conocer información que no podríamos apreciar con simples números.

El punto de la visualización de datos es "dejar que el usuario entienda que es lo que esta pasando en su información." [7] La minería de datos, extrae información de bases de datos que el usuario no necesariamente conoce. Las relaciones entre las variables no son intuitivas y son de hecho lo que la MD desea encontrar. Una ventaja importante en el proceso de MD, es que el usuario no conoce de antemano qué es lo que el proceso descubrirá, por lo que puede tomar la salida resultante y trasladarla en una solución para el problema de negocio.

Existen muchas formas de representación de información, la visualización es una forma de maximizar el valor de los datos para quien los analiza, ya que el analista observa el comportamiento de la información desde otra perspectiva.

SAS cuenta con diferentes módulos de visualización de información, SAS/INSIGTH® y SAS/GRAPH® proveen de una variedad de gráficas que ayudan a percibir fácilmente la información que se maneja.

[7] Kurt Thearling, Bary Becker, Dennis DeCoste, Bill Mawby, Michel Pilote, and Dan Sommerfield, Artículo publicado "Visualling Data Mining Models"

Regresiones

Las técnicas de regresiones pueden ser usadas para realizar predicciones. Esta técnica es usada para encontrar, la relación entre las variables de entrada y la variable objetivo. La Regresión Logística es usada para variables objetivo binarias y ordinales; la lineal para variables objetivo de intervalo.

Por muchos años, las regresiones han sido una técnica estándar para la estimación de la variable objetivo, mucho antes de la llegada de las redes neuronales y los árboles de decisión. Las regresiones lineales intentan predecir el valor de una variable objetivo al continuar como una función lineal de una o mas variables independientes. Las regresiones logísticas intentan predecir la probabilidad de una variable binaria u ordinal que adquirirá el evento de interés como una función de una o más variables independientes de entrada.

Las Regresiones han sido particularmente exitosas dentro de la calificación de aplicación en dónde la variable objetivo puede ser usada para estimar las características de los clientes, así como, la probabilidad de respuesta en una campaña publicitaria. Al igual que con las redes neuronales, las relaciones no lineales que son particulares, pueden ser incorporadas en estos modelos.

Por lo tanto, es una buena idea comparar modelos de redes neuronales y árboles de decisión con las regresiones, si el problema de negocio permite el uso de los tres modelos.

SAS/STAT® cuenta con procedimientos para llevar a cabo los diferentes tipos de regresiones que se deseen hacer, al igual, que SAS/Enterprise Miner® .

Árboles de decisión

Los Árboles de Decisión son herramientas analíticas usadas para descubrir reglas y relaciones con respecto a la variable objetivo, dividiendo y subdividiendo la información contenida en el conjunto de datos.
La herramienta que cubre la mayoría de los algoritmos de árboles de decisión, a través de una interfase gráfica es SAS/Enterprise Miner®.

La representación de un árbol, es una segmentación de los datos que es creado aplicando una serie de reglas simples. Cada regla asigna una observación a un segmento basado en el valor de entrada. Una regla es aplicada después de otra, resultando una organización de segmentos dentro de segmentos. La organización o jerarquía es llamada árbol, y cada segmento es llamado nodo. El segmento original contiene el conjunto de datos entero y es llamado el nodo raíz del árbol. Un nodo con todos sus sucesores forman una rama del nodo, los nodos finales son llamados niveles. Para cada hoja, una decisión es hecha y aplicada a todas las observaciones en la hoja. El tipo de decisión depende en el contexto. En el modelado predictivo, la decisión es simplemente el valor predictivo.

Cuando la variable objetivo es categórica, el Árbol de Decisión recibe el nombre de Clasificación. Y si es continuo recibe el nombre de Árbol de Regresión.

El Árbol se ajusta a los datos mediante un particionamiento recursivo. El particionamiento significa segmentar los datos en subgrupos que sean lo más homogéneos posible respecto a la variable objetivo.

El método es recursivo porque cada subgrupo es el resultado de la división de un grupo procedente de otra división anterior.

Criterios para la construcción de Árboles de Decisión

- **Búsqueda de divisiones**
  ¿Qué divisiones deben considerarse?
- **Criterio para la división**
  ¿Qué división es la mejor?
- **Regla de paro**
  ¿Cuándo debo parar la división?
- **Regla de poda**
  ¿Deberían cortarse algunas ramas?

Búsqueda de divisiones

El número de divisiones posibles que se pueden realizar es muy grande, en todos menos en los caso más sencillos. Ningún algoritmo de búsqueda de partición examina detalladamente todas las particiones posibles. En su lugar, se imponen varias restricciones para limitar las divisiones posibles que se pueden realizar. La restricción más común es tener en cuenta sólo las divisiones binarias. Otras restricciones implican discretizar las entrada continuas, algoritmos de búsqueda stepwise y el muestreo.

Criterio para la división

En algunas situaciones el valor de una partición es evidente. Si el objetivo esperado es el mismo en los nodos hijos que en los nodos padres, no realiza ninguna mejora y la división carece de valor. En contraste, si la división da como resultado nodos hijos puros la división es indiscutiblemente mejor. En los Árboles de Clasificación, las técnicas de construcción de Árboles más utilizadas como criterio de división está basado en el test chi-cuadrado de Perrazo, en el Indice de Gini y en la Entropía. Los tres miden la diferencia que existe en las distribuciones de las clases en los nodo hijo, y los tres métodos suelen dar resultados similares. Cabe decir, que no todas las herramientas contienen estas técnicas pero el resto de las técnicas para la construcción de Árboles no varían mucho.

Regla de poda

En los Árboles de Decisión, la complejidad del método se mide por el número de hojas. Un Árbol se puede dividir continuamente hasta que todas las hojas sean puras o contengan un solo caso.

Probablemente este Árbol ajuste perfectamente con los datos de entrenamiento, pero puede que no dé buenas predicciones en datos nuevos. En otro extremo, puede que el árbol tenga sólo una hoja (el nodo raíz). Puede que cada caso tenga el mismo valor de predicción (ninguna regla de datos). Existen dos métodos para determinar el árbol que tenga el tamaño adecuado:

1. **Utilizar reglas que impidan el crecimiento de un Árbol (pre-poda)**
   Una regla universal aceptada de pre-poda es parar el crecimiento si el nodo es puro. Otras dos reglas aceptadas son parar el crecimiento si el número de casos en un nodo se encuentra por debajo de un límite especificado o pararlo si la división no es significativa en estadística en un nivel especifico.

2. **Crear un Árbol grande y eliminar algunas ramas después (post - poda)**
   La post-poda crea una sucesión de Árboles de gran complejidad. Se necesita un criterio de valoración para decidir el mejor (sub) Árbol.

La pre-poda requiere menos cálculo, pero tiene el riesgo de podar futuras divisiones válidas bajo las divisiones débiles.
Se han propuesto cientos de algoritmos de Árboles de Decisión en el aprendizaje de máquinas, el análisis estadístico, y en la literatura del reconocimiento del modelo. Los más comerciales y aceptados son CART, CHAID, o C4.5 (C5.0)

Existen muchas variaciones del algoritmo CART (Árboles de Clasificación y de Regresión Breiman 1984). El método estándar de CART está restringido a divisiones binarias y utiliza la post-poda. En él se consideran todas las combinaciones binarias posibles. Si hay muchos datos puede que se utilice el muestreo dentro del nodo. El criterio de división estándar se basa en el índice de Gini para los Árboles de Clasificación y la reducción de la varianza para los Árboles de Regresión. También se pueden usar otros criterios para los problemas de clase múltiples (criterio twoing) y los Árboles de Regresión (mínima desviación absoluta). Utilizando la validación cruzada v-fold se puede construir un gran Árbol y luego podarlo. Si hay datos suficientes se puede utilizar un fichero de validación única.

CHAID (detección de interacción automática de chi-cuadrado) es una modificación del Algoritmo AID desarrollado primeramente en 1963 (Morgan y Sonquist 1963, Kass 1980). CHAID utiliza particiones múltiples y la pre-poda para crear los Árboles de Clasificación. Encuentra la mejor división múltiple usando el algoritmo stepwise. El algoritmo de búsqueda de división se diseña para entradas categóricas, por lo que las entradas continuas se deben discretizar. Los criterios de división y de paro se basan en la significancia estadística (test de chi-cuadrado).

La familia de Árboles de Clasificación ID3 se desarrolló en la literatura de aprendizaje de máquinas (Quilan 1993) C4.5 sólo tiene en cuenta las divisiones tamaño –L para las entradas categóricas nivel –L y divisiones binarias para entradas continuas. El criterio de división se basa en la mejora de la entropía. La post-poda se realiza utilizando pesimistas en la tasa de error del conjunto de entrenamiento.

Redes neuronales

Las redes neuronales son modelos que intentan reproducir el comportamiento del cerebro. Este modelo realiza una simplificación, al averiguar cuáles son los elementos relevantes del sistema, ya sea porque la cantidad de información que se dispone es excesiva o porque es redundante. Una elección adecuada de sus características, más una estructura conveniente, es el procedimiento convencional utilizado para construir redes capaces de realizar una determinada tarea.

Las Redes Neuronales son un método computacional comúnmente usado para identificación y clasificación de patrones, en particular las redes neuronales se utilizan en el ambiente financiero para modelar fraude en tarjetas de crédito.

SAS/Enterprise Miner® provee diferentes arquitecturas de redes neuronales que permite al usuario modificar el modelo hasta encontrar uno satisfactorio, si es que existe.

Análisis de Agrupamiento

El análisis de agrupamiento es una técnica de clasificación de patrones. El análisis de agrupación o *Clustering*, ejecuta agrupación de observaciones, de esta forma, segmenta conjunto de datos o bases de datos completas. El análisis de agrupamiento coloca la información dentro de las agrupaciones sugeridas para los datos. La información en cada grupo tiende a ser similar a cada otra en alguna forma, y la información en grupos diferentes tienden a ser distintas.

El análisis de agrupamiento puede ser aplicado en el caso en que las compañías deseen conocer las características de sus clientes rentables, poco rentables y nada rentables, y de esta forma crear estrategias de mercado para conservar a los clientes importantes y buscar nuevos con estas características.

Algoritmos Genéticos

"Los Algoritmos Genéticos son métodos adaptativos que pueden usarse para resolver problemas de búsqueda y optimización." [8]

Los algoritmos Genéticos usan una analogía directa con el comportamiento natural. Trabajan con una población de individuos, cada uno de los cuales representan una solución factible a un problema dado. A cada individuo se le asigna un valor ó puntuación, relacionado con la bondad de dicha solución. En la naturaleza esto equivaldría al grado de efectividad de un organismo para competir por unos determinados recursos. En cuanto mayor sea la adaptación de un individuo al problema, mayor será la probabilidad de que él mismo sea seleccionado para reproducirse, al cruzar su material genético con otro individuo seleccionado de igual forma. Este cruce producirá nuevos individuos – descendientes de los anteriores – los cuales comparten algunas de las características de sus padres. Cuando menor sea la adaptación de un individuo, menor será la probabilidad de que dicho individuo sea seleccionado para la reproducción, y por tanto de que su material genético se propague en sucesivas generaciones.

Análisis de liga o relación

El análisis de liga o relación (Link análisis), es un acercamiento descriptivo que permite explorar datos que pueden ayudar a identificar las relaciones a través de los diferentes valores de las bases de datos. Los dos análisis comúnmente conocidos son: descubrimiento de asociación y descubrimiento secuencial. El primero encuentra reglas de asociación entre los artículos en un evento, como puede ser el pago de una transacción. El análisis de descubrimiento secuencial, es muy similar, solamente que en este se toma en cuenta una asociación con respecto al tiempo.

[8] Olivia Parr Rud, "Data Mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management", Edit. John Wiley & Sons, 2000.

1.3.2 Herramientas de minería de datos

En esta sección se pretende hacer mención de las diferentes herramientas que existen en el mercado sobre MD, a lo largo del presente trabajo se habla solamente de la herramienta de MD del sistema SAS, sin embargo, se debe tomar en cuenta que existen otras y que dependiendo de las necesidades del área de negocio o investigación estas pueden satisfacer los requerimientos.

A continuación se presenta una matriz de comparación con las características más relevantes de las herramientas de MD, las matrices tendrán un asterisco por cada característica que se contenga, para que el lector aprecie la diferencia entre estas. No siempre la herramienta que contenga la más alta tecnología tiene que ser la mejor, tenemos que elegir la que satisfaga los requerimientos necesarios para el proyecto que estamos o vamos a llevar a cabo.

Las herramientas que se analizarán son:

- Enterprise Miner (SAS)
- ntelligent Miner (IBM)
- KnowledgeSEEKER (Angoss)
- CA Neugents (Computer Associates International, Inc.)
- Data Mining Suite (Information Discovery, Inc.)
- Insightful Miner (Insightful Corporation)
- CHAID (SPSS)
- Splus (Mathsoft)
- Discovery Server (Pilot)
- Knowledge Discovery Workbench (NCR)
- Clementine (Integral Solutions)

Matriz Comparativa

			Características	
Empresa Vendedora	Producto	Dirección de Internet	Cliente/ Servidor	Tipos de Datos a los que tiene acceso
SAS	Enterprise Miner	http://www.sas.com	•	A la mayoría
IBM	Intelligent Miner	http://www-3.ibm.com/software/data/iminer/	•	DB2
Angoss	KnowledgeSEEKER	http://www.angoss.com	•	ASCII, ODBC, EXCEL, LOTUS, DBASE, SAS, S-PLUS, SQL*NET
Computer Associates International, Inc.	CA Neugents	http://www.computerassociates.com	•	A la mayoría, excepto SAP y Adabas
Information Discovery, Inc.	Data Mining Suite	www.datamining.com	•	Grandes bases de datos en SQL
Insightful Corporation	Insightful Miner	http://www.insightful.com/		ASCII, Excel, OBC.
SPSS	CHAID	http://www.spss.com/datamine/		ASCII
Mathsoft	Splus	http://www.mathsoft.com/		
Pilot	Discovery Server	http://www.pilotsw.com/	•	SQLServer, Oracle7.2
NCR	NCR TeraMiner	http://www.ncr.com		
Integral Solutions	Clementine	http://www.spss.com/spssbi/clementine/		Oracle, Ing, Sybase,Inf

Las características de las herramientas que se muestran en las matrices comparativas anteriores, pueden ser verificadas en la dirección en Internet antes mencionada.

Matriz Comparativa

	Características						
Empresa Vendedora	Muestreo	Escalabilidad	Selección de variables	Visualización	Redes Neuronales	Árboles de decisión	Series de Tiempo
SAS	mas de 3 tipos	•	•	•	•	•	•
IBM	aleatorio	•	•	•	•	•	•
Angoss	aleatorio			•		•	
Computer Associates International, Inc.		•	•		•	•	
Information Discovery, Inc.		•	•	•			
Insightful Corporation			•	•	•	•	
SPSS	propio		•	•		•	•
Mathsoft							
Pilot	aleatorio				•	•	
NCR							
Integral Solutions		•		•	•	•	

Matriz Comparativa

	Características						
Empresa Vendedora	Regresiones	Agrupamiento	Validación	Manejo de modelos	Permitir presentación en web con JAVA	Publicación en WEB	Procesamiento en paralelo
SAS	•	•	•	•		•	
IBM		•		•			•
Angoss				Solamente árboles			
Computer Associates International, Inc.				Solamente redes neuronales			
Information Discovery, Inc.				Algoritmos reservados	•	•	
Insightful Corporation	•			•	•		
DataMind		•					
SPSS	•	•					
Mathsoft							
Pilot		•					
NCR							
Integral Solutions	•			•			•

Al analizar las matrices anteriores se aprecia claramente que la herramienta que cubre con el mayor número de características y técnicas de MD, es SAS Enterprise Miner®, además, de abarcar completamente con todos los pasos del proceso de minería de datos —que se describirán en el siguiente tema al comenzar con una muestra de los datos—, el análisis de la información y se termina con el modelado y análisis de los resultados. La funcionalidad de esta herramienta ayuda a usuarios de diferentes grados de experiencia estadística, a través, de una interfaz gráfica de usuario (GUI) a planear, implementar y refinar sus proyectos de MD.

Una vez presentadas las diferentes herramientas de MD existentes, se continuará con la descripción del proceso de MD y de la herramienta del sistema SAS que apoya de manera total el trabajo que se presenta.

1.4 El sistema SAS y la minería de datos

Con el fin, de llevar a cabo una exitosa aplicación del proceso de MD, SAS ha desarrollado un método denominado SEMMA, acrónimo de Muestra (*Sample)*, Exploración (*Explore)*, Modificación *(Modify.)*, Modelado (*Model) y* Validación (*Assess)* por sus siglas en inglés, el cual se tomará para la aplicación del caso práctico que se desarrolla en el capítulo II, cabe mencionar, que el método que se presenta a continuación es parte del proceso de planeación para la aplicación del método de MD, que se describirá también en el capítulo II. Una vez que se recolecta la información suficiente, el método de MD puede ser aplicado.

A continuación se detallará en que consiste cada uno de los componentes del método de la MD.

Muestreo. (Sample) Se refiere a la extracción de datos de un extenso cúmulo de estos, para un manejo más rápido de los procesos.

SAS hace uso del muestreo, para un costo óptimo del manejo de la información en cuanto uso de equipo de cómputo. SAS obtiene una muestra confiable y representativa del total de los datos, lo cual, ayuda a reducir el tiempo de proceso requerido para la entrega de información de negocio.

Exploración. (Explore) Se buscan tendencias o anomalías, que ayudan a entender el tipo de datos que se manejan.

La exploración de la información, a través, de métodos gráficos nos ayuda a refinar los datos para tener un conocimiento de la información que se maneja y de los resultados obtenidos, después de la realización del muestreo.

Modificación. (Modify) En la etapa de modificación, podemos crear, seleccionar o transformar variables de acuerdo al conocimiento previo de la exploración de la información.

Dado que el proceso de minería es dinámico e iterativo, podemos cambiar la información para explorar nuevos resultados.

Modelado. (Model) El modelado es la búsqueda de forma automática, del resultado predicho esperado, a través, de una combinación de variables

Cada tipo de modelo debe ser aplicado dependiendo de los datos que se manejen, por ejemplo las redes neuronales son utilizadas cuando se presentan relaciones no lineales en las variables.

Assess. (Validación) En este paso, evaluamos la eficiencia y confiabilidad de los métodos utilizados, en el proceso de minería.

Descripción de la herramienta de minería de datos del sistema SAS

La herramienta de SAS para la MD, es una interfase gráfica para el usuario (GUI), utilizada para el descubrimiento del conocimiento y tiene el nombre de SAS Enterprise Miner®.

SAS Enterprise Miner® provee de una variedad de capacidades dentro de la MD, que puede satisfacer las necesidades de usuarios de diferentes áreas de negocio como son: tecnología de información, analistas de negocio y analistas con perfil estadístico. Con esta herramienta los usuarios pueden:

- Identificar los clientes más rentables
- Determinar cuáles son los clientes que van a ser absorbidos por la competencia
- Determinar la combinación de productos de los clientes para su compra
- Incrementar la lealtad del cliente
- Detectar y prevenir conductas fraudulentas en los sitios de comercio electrónico.

Enterprise Miner contiene una gamma de modelos y algoritmos incluyendo árboles de decisión, redes neuronales, regresiones, series de tiempo, clustering, asociaciones, entre otros, además de la generación de un reporte en html que permite documentar el proceso de MD utilizado.

La interfaz gráfica del usuario se encuentra organizada con el fin de que el usuario pueda seguir el método SEMMA descrito con anterioridad, de manera sencilla, ya que la interfaz se encuentra organizada de tal forma que se indican las técnicas incluidas en cada paso del proceso de MD.

Una vez descrita la interfaz que se utilizará a largo de ésta investigación, se presentará la relación existente entre la minería de datos y lo importante que es esta dentro de los negocios empresariales para que tengamos un panorama general de la MD y posteriormente aplicar este concepto a un caso práctico

Relevancia de la minería de datos en los negocios empresariales

La minería de datos es parte de una larga serie de pasos, que toman un lugar entre la compañía y sus clientes. La forma en la cual la MD impacta a los negocios depende de los procesos que estos tengan y no de la MD. Por ejemplo, en el proceso de sacar al mercado un nuevo producto, el trabajo del director de mercadotecnia es entender su mercado, con este entendimiento viene la habilidad de interactuar con los clientes de su mercado al hacer uso de diferentes canales, además, de interactuar con diferentes áreas del negocio, incluyendo mercadeo directo, impresión de anuncios, telemercadeo, anuncios a través de la radio y televisión, y muchos otros.

La MD, extrae la información de las bases de datos la cual es desconocida por los clientes . Las relaciones entre las diferentes variables y la conducta del cliente , no es algo intuitivo y es precisamente lo que se espera encontrar a través del proceso de MD; y basándose en que el cliente no conoce lo que se descubrió con el proceso, la información obtenida es de tal importancia que puede traducirse en la solución del problema de negocio.

Los consultores de MD, necesitan entender cuales son los resultados de esta, antes de que apliquen este proceso, ya que usualmente esta relacionada con la extracción de patrones desconocidos en la conducta del cliente y esto puede llegar a ser un proceso un poco complicado.

Si la información necesaria existe en la base de datos, el proceso de MD puede modelar cualquier actividad del cliente, ya que el objetivo principal es encontrar patrones relevantes para el actual proceso de negocio

Lo importante dentro de un proceso de MD es que las personas que vayan a aplicar este proceso cuenten con una clase de *software* la cual automatice el proceso de búsqueda en un cantidad extensa de datos, que encuentre patrones no conocidos y que estos sean buenos predictores de la conducta de los clientes.

La MD ayuda a los negocios empresariales a enfocar sus campañas de mercadotecnia más acertadamente, permite alinear estas campañas más estrechamente a las necesidades, deseos y actitudes de los clientes y prospectos. Por tanto si deseamos aplicar la MD a un problema de negocio, debemos contar con una herramienta que facilite este proceso, con expertos que tengan conocimiento de las técnicas que se van a aplicar y analistas de negocio con experiencia. En el posterior capítulo se detallará los requerimientos para ala aplicación de la MD a un proyecto en producción.

Hasta el momento se ha descrito el ambiente de las EDVD, SAS la empresa desarrolladora de un sistema modular del mismo nombre el cual se encarga de la entrega de información que ayuda a la toma de decisiones y la MD, un proceso para la búsqueda de patrones de comportamiento desconocidos enfocada a diferentes áreas de negocio, además de que integra a las diferentes herramientas que pueden servir para el mismo fin.

Se ha mencionado el método que se debe seguir para llevar a cabo el proceso de MD, lo que falta por mencionar es el proceso previo y posterior a la aplicación del método esto es el proceso de planeación para la aplicación del método, y dado que "La planeación es un proceso para ganar conocimiento y así apoyar la toma de decisiones para guiar la acción conforme a ciertos objetivos"[9], lo que pretende el siguiente capítulo es presentar una guía que sirva para la obtención de conocimiento del área de negocio a la cual se le aplicará el proceso de MD, se continua con la aplicación del método a un caso práctico y se analizan en el tercer capítulo los resultados obtenidos que permitirán apoyar a la toma de decisiones en las EDVD.

Dentro de la planeación de un proyecto de MD se describirán los pasos a seguir para la ejecución del método de MD, para este fin, el cliente debe contar obviamente con un cúmulo de información la cual quiera analizar, y con un histórico mínimo de 6 meses para que el pronóstico generado sea acertado. En el proceso de planeación, se describirán los pasos a seguir desde la obtención de esta información hasta la aplicación y revisión del proceso de MD.

A continuación se presenta un cuadro sinóptico del proceso de MD , que describirá los pasos que se ejecutaran en el siguiente capítulo en el caso práctico, para que el lector tenga una concepción general de lo que se pretende hacer en los restantes capítulos, una vez que se han descrito los conceptos generales de los temas a tratar.

[9] Arturo Fuentes Zenón, "Un sistema de metodologías de planeación", Agosto 1999.

Cuadro sinóptico del proceso de minería de datos

Proceso de minería de datos

1. Definir el problema de negocio que se quiere resolver
2. Evaluar el ambiente de datos
3. Disponibilidad de datos
4. Aplicación del método de minería de datos elegido(en este caso se elegirá el método propuesto por el sistema (SAS) que incluye los siguientes pasos:
    - Muestreo(Se obtiene una muestra para la obtención de resultados en un tiempo optimo)
    - Exploración (Se visualiza la información de manera gráfica, para una comprensión de la misma)
    - Modificación (Se modifica la información de acuerdo al análisis anterior)
    - Modelado(Se aplica un modelo matemático para el pronóstico)
    - Validación(Se validan los modelos para la elección del mejor)
4. Revisión de los resultados.Se analizan los resultados y se les entrega a las personas de marketing encargadas de la toma de decisiones, para generar mejores estrategias de negocio.

El anterior cuadro sinóptico muestra el proceso a desarrollar en los próximos dos capítulos, que describirán todos los pasos del método de planeación y se presentarán los resultados del proceso con los alcances y limitaciones del mismo.

Conclusiones

El proceso de minería de datos, que se encarga de descubrir patrones de conducta antes desconocidos, es un concepto que se ha reforzado a lo largo de los años con herramientas que facilitan el análisis de los datos, un ejemplo de estas, es la del sistema SAS llamada SAS/Enterprise Miner®. SAS ha desarrollado una herramienta que proporciona una cantidad de modelos suficientes para la búsqueda de patrones de conducta, ejemplo de estos son: las regresiones lineales, los árboles de decisión y redes neuronales, entre otros, además de apoyarse en una técnica de planeación y aplicación de MD para la realización de un proyecto.

Como se menciono en la sección 1.3, un proyecto de minería de datos es aplicable a diferentes áreas de negocio ya sea Telefonía Banca, EDVD, etc. La investigación presentada se enfoca en una de las empresas dedicadas a las ventas a detalle, y se considera relevante el análisis ya que como se menciona en el apartado 1.1 todo sujeto es consumidor de bienes y servicios lo largo de su existencia, también se considera que el principal objetivo de las empresas es generar mayores ingresos año con año, por lo que este trabajo guiara al consultor para la aplicación de un método de minería de datos de manera eficiente , y se usará la herramienta del sistema SAS, para obtener una base de datos de clientes potenciales que ayuden a las EDVD a la toma de decisiones para generar mayores ingresos.

No se debe dejar de lado que existe una gran variedad de herramientas de minería de datos en el mercado, las más importante están mencionadas en el subtema 1.3.2, y lo que hace la diferencia entre estas, son los tipos de modelos que manejan, las herramientas de visualización, los tipos de datos a los que tienen acceso y la forma en que procesan la información, por lo que el presente trabajo muestra una forma en la cual se lleva a cabo la aplicación del método de MD y aunque esta enfocada al sistema SAS, dicho trabajo dará una visión general a cualquier persona que se encuentre en un proyecto de MD, trabajando con alguna otra herramienta.

Fuentes de consulta

1. Ronald W. Hasty, "Retailing", Harper & Row, Publishers, Tercera Edición, New York, 1983.
2. Warren G Meyer, James R Stone III, Donald P. Kohonx, E.Edward Harris, "Marketing, Ventas al por menor, para empleados, gerentes y empresarios", Edit. McGraw-Hill, Octava Edición, 1991
3. Página principal de SAS Institute (http//www.sas.com)
4. Página principal de Dataflux (http://www.dataflux.com)
5. SAS Institute, "Finding the Solution to Data Mining, Exploring the features and Componentes of Enterprise Miner , Release 4.1 from SAS", Ed. 2001.
6. David J. Hand, Heikki Mannila, and Padhraic Smyth, "Principles of Data Mining", Massachusetts Institute of Technology, 2001.
7. Kurt Thearling, "Data mining and Custumer Relationships management", Folleto extraído del libro "Building Data mining application for CRM" por Alex Berson, Stephen Smith y Kurt Thearling, Edit. McGraw Hill, 2000.
8. Olivia Parr Rud, "Data Mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management", Edit. John Wiley & Sons, 2000.
9. Arturo Fuentes Zenón, "Un sistema de metodologías de planeación, COMISIÓN NACIONAL DEL AGUA, Subdirección general de programación, Agosto 1999.

Otras fuentes de consulta

1. J. Barry Mason, Morris L. Mayer, Hazel f. Ezell, "Fundamentos de Comercio al Menudeo", Edit. CIA. Continental, S.A. de C.V., México, 1989.
2. William J.E. Potts ,"Data Miner Primer, Overview and Applications and Methods", SAS Institute Inc., 1998
3. Knowledge Discovery Nuggets (http://www.kdnuggets.com/)

Capitulo II

Aplicación de un método de minería de datos, para las empresas dedicadas a las ventas a detalle.

Objetivo

Analizar una técnica de planeación para un proyecto y aplicar el método de minería de datos propuesto por el sistema SAS a un caso práctico.

2.1 Planeación para la aplicación del método de minería de datos

"Hacer planes no es el fin de la empresa, pero ésta no puede sobrevivir sin una planeación adecuada"[10], por lo que, para cualquier proyecto en el cual estén involucrados recursos humanos, tecnología, costos y beneficios, se requiere que se lleve a cabo una planeación previa; para las empresas como lo señala Ackoff. El planear significa: "ir recopilando información sobre los proyectos de la compañía, de abajo hacia arriba"[11], es decir, se debe pasar por los diferentes niveles de la empresa, esto se verá en el método de planeación propuesto por SAS para llevar a cabo la planeación de un proyecto de MD, ya que en las diferentes etapas de planeación se recolecta información de las personas del área de negocio, tecnología, usuarios finales, etc., de tal forma que todas las áreas se encuentren involucradas en el proyecto que se pretende realizar, y así formar un solo sistema enfocado a un fin común, el cual, menciona Ackoff "analizándolo de forma estructural parezca un todo divisible, pero que visto funcionalmente parezca un todo indivisible en el sentido que alguna de sus propiedades esenciales se pierden cuando se desmiembra."[12]

La planeación que se llevará a cabo en este proyecto es interactiva ya que para cualquier proceso de MD es necesario analizar el pasado, el presente y de esta forma dar soluciones para un futuro próximo basándose en modelos matemáticos para este fin.

Las cinco etapas de la planeación interactiva que plantea Ackoff son:
1. - Formulación de la problemática, en donde se identifican las amenazas y oportunidades que enfrenta el sistema.
2. -Planeación de los fines, es decir, la especificación de los logros que se van a perseguir, mediante el diseño del futuro más deseable.
3. -Planeación de los medios, en donde se selecciona y/o se crean los caminos a seguir para aproximarse al futuro deseado.
4. - Planeación de los recursos, en la cual se determinan tipos, cantidades, tiempos y fuentes de origen de las entradas que se requerirán para lograr los fines planteados.
5. -Diseño de la implementación y el control en donde se determina quién, qué, cuando y dónde se van a llevar a cabo las acciones y operaciones, aunado con las medidas de desempeño y la forma de moni torearlas para poder controlar el desarrollo de la implantación.

En este trabajo a partir del método de planeación propuesto por el sistema SAS, se confirmará que este método incluye las etapas de la planeación interactiva, ya que parte de la formulación de la problemática, incluida en la fase de "Definición del problema de negocio", se continua con la definición de fines la cual se realiza en conjunto con el área de negocio, lo que en este trabajo se plantea también dentro de la "Definición del problema de negocio", la planeación de los medios y de recursos se refleja en la etapa de "Evaluación del ambiente" y finalmente el diseño de la implementación la encontramos en la etapa de "Implementación en la producción".
La planeación es de suma importancia en un proyecto porque todo proyecto implica realizar algo que no se ha hecho antes. Por lo que, en este tema se describen las tareas que se deben llevar a cabo en un proceso de planeación de proyectos propuesto por el sistema SAS.

Un método de planeación de proyectos es cualquier enfoque estructurado para guiar al equipo durante el desarrollo del plan del mismo. Puede ser algo tan sencillo y utilizar formas y planillas estándar(sean de papel o electrónicas, formales o informales) o tan complejas y hacer uso de una serie de simulaciones necesarias.

---

[10] C.P. y M.A: Mária Luisa Saavedra Gracia, "Un esquema de Planeación desde el enfoque sistémico"
[11] Idem
[12] Idem

El plan del proyecto es un documento formal y aprobado, utilizado para administrar y controlar la ejecución del proyecto. Se deberá distribuir tal como se haya definido el plan de administración del proyecto (en este caso se definirá los pasos a seguir de acuerdo al método de SAS)

Un plan de proyecto se emplea para:

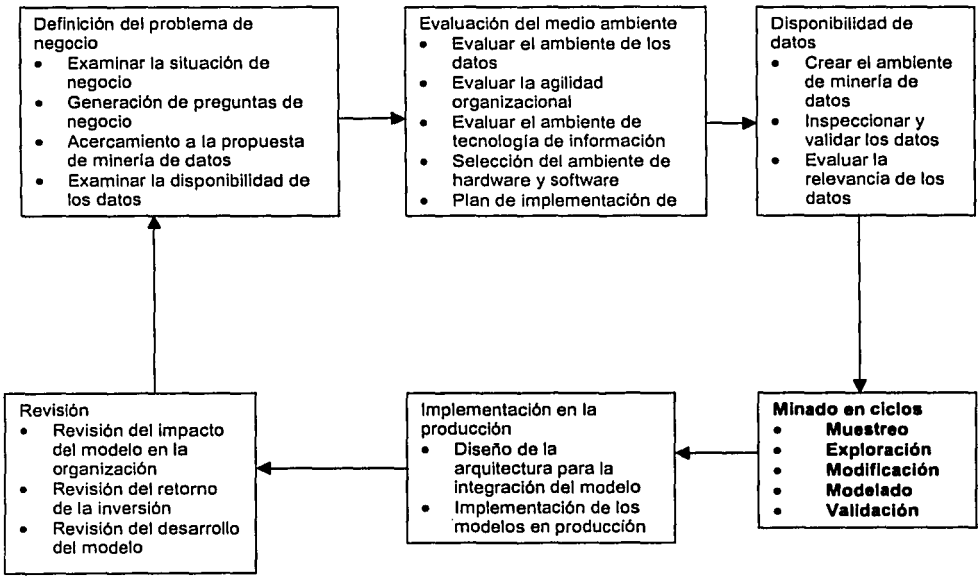
- Guiar la ejecución del proyecto.
- Documentar suposiciones relacionadas con la planeación del proyecto.
- Documentar decisiones relacionadas con la planeación del proyecto relativas a las alternativas elegidas.
- Facilitar la comunicación entre los responsables de las diversas tareas.
- Definir importantes revisiones administrativas en cuanto a contenido.
- Proporcionar una línea de base para la medición del progreso y el control del proyecto.

A continuación se presentará el método de planeación de proyectos de MD propuesto por el sistema SAS.[13]

2.1.1 Visión general del método de planeación

Proceso de planeación de un proyecto de minería de datos

La siguiente figura muestra el proceso iterativo que se debe llevar a cabo para un proyecto de minería de datos, y resalta el método que se utilizará en el caso práctico.



[13] SAS Institute Inc., "Manual of Data Mining Project Methodology", Cary, N.C.,Mayo 2000.

El método propuesto por SAS para ejecutar un proyecto de minería de datos, consiste en seis actividades, que incluyen:

- Definición del problema de negocio: el objetivo es definir los problemas de los clientes desde el punto de vista de minería, plantear una revisión exhaustiva para saber si el cliente tiene los datos apropiados para llevar a cabo el análisis requerido.

- Evaluación del medio ambiente: la evaluación del medio ambiente nos va a proporcionar un estudio de los clientes de Tecnologías de información, de negocio y de la organización y preparar el ambiente para la colección de información, preparar la propuesta de arquitectura y el plan que se debe llevar a cabo para el análisis de la información.

- Disponibilidad de datos: una vez que el problema de negocio ha estado definido con detalle, es necesario encontrar y preparar los datos relevantes para el proyecto.

- Ciclo de minería: Ejecución del método de minería de datos.

- Implementación en la producción: una vez que han sido probados y verificados los modelos a utilizar en minería de datos y que han sido exitosos al manejar un ambiente en particular, se debe de implementar estos modelos en algunos casos directamente en los sistemas operacionales o en un ambiente de data warehouse

- Revisión: se lleva a cabo una revisión para validar los beneficios obtenidos a través de la minería de datos en la organización.

Los anteriores elementos del método de planeación, serán descritos a continuación para tener un enfoque general de las actividades que se deben realizar en cada una de las etapas mencionadas.

2.1.2 Definición del problema de negocio

Antes de la ejecución de esta fase, se deberá tener un entendimiento de las necesidades, organizadas de forma prioritaria y de las decisiones de la organización, al concluir esta fase se deberá contar con dos elementos importantes que son: 1. La definición del problema de negocio. 2. El plan de negocio.
Las actividades a realizar dentro de esta fase son las siguientes:

- Examinar la situación de negocio, esta fase incluye lo siguiente:

  Entender el ambiente de negocio del cliente (Negocio del cliente, terminología del cliente, operaciones de negocio, estructura organizacional, incluyendo políticas internas de la organización)

  Identificar la situación del cliente: En esta fase se deben utilizar técnicas de cuestionarios y entrevistas para tener un entendimiento del negocio, por lo que se debe investigar a fondo a la compañía y el proyecto el cual quiere realizar, ya que la mayoría de las veces el cliente no tiene bien definido el problema.

- Generación de preguntas de negocio

  Identificación de preguntas específicas de negocio: Una vez que se tenga un entendimiento del negocio y el ambiente, se deben realizar preguntas específicas de negocio, que puedan ser respondidas a través técnicas de minería de datos.
  Algunos ejemplos son los siguientes:
  ¿Existen datos disponibles que pueden ser usados para una mejor planeación?
  ¿Existen preguntas que no han sido resueltas y que afectan los costos, utilidades, ventas en el mercado, etc.?

- Acercamiento a la propuesta de minería de datos

  Con las preguntas específicas realizadas, es posible realizar una identificación preliminar del problema que se debe atacar en el proceso de minería de datos.

- Examinar la disponibilidad de los datos

  Identificar las fuentes de datos: Se deben identificar las fuentes de datos relevantes del problema de negocio, ya sean sistemas operacionales, un datawarehouse existente o datamart, o datos externos.

  Examinar los datos: De hecho para la definición del problema de negocio, es necesario analizar y discutir la estructura, accesibilidad de los datos. La identificación de los datos que se van a utilizar es relevante para la definición del problema de negocio, por lo que se incluye para la identificación del problema.

  Asegurarse de contar con la información suficiente: Es importante asegurarse de que la información que se ha recabado es la suficiente y esta disponible para analizar el problema de negocio, además de verificar la legalidad de la información.

- Desarrollar el plan de negocio

  Cuantificar y definir el problema de negocio desde el punto de vista del ambiente de negocio: La cuantificación de problemas de negocio requiere el establecimiento de medidas que proporcionen el grado de éxito del proyecto, típicamente va relacionado con el retorno de la inversión que se obtenga una vez concluido el proyecto.

Se debe definir el problema de negocio desde el punto de vista de la medición del éxito del mismo, y considerar qué efectos tiene el proyecto en las estrategias, ventas, o *marketing*.

Estimar la oportunidad inicial negocio: La estimación de la oportunidad de negocio es necesaria para que el proyecto se lleve a cabo, y tener un soporte del proyecto dentro de la organización, es necesario también realizar un análisis del técnico para asegurarse de que el proyecto se haga con las tecnologías actuales de minería de datos.

Determinar las necesidades para la prueba piloto: En caso de que se tenga experiencia a priori con un proyecto similar, la prueba piloto no es necesaria. De otra forma, si se es novato en el proceso de minería, la prueba piloto puede servir para llevar a cabo de forma satisfactoria el proyecto, además de:

Crear e incrementar el interés del cliente en corto tiempo.
Dar tiempo al cliente para que obtenga confianza en el equipo de desarrollo del proyecto.
Enfocarnos en las mayores dificultades y cambios que puedan darse en el futuro.

➢ Acercamiento del proceso de minería de datos

La identificación del acercamiento del proceso de minería de datos puede ser usada para analizar los problemas de negocio seleccionados de los clientes.

El acercamiento del proceso de minería trae las siguientes implicaciones:
Tipos de datos que se van a utilizar
Elegir las técnicas y métodos de minería de datos a utilizar.
Requerimientos de una variable objetivo

Idealmente se debe realizar una tabla que enumere los posibles problemas de negocio y un acercamiento relevante con una solución asociada debe ser disponible. Sin embargo la lista de los problemas potenciales debe ser diversa, y exhaustiva.

Un acercamiento relevante incluye los siguientes puntos:

- Encontrar uno o más problemas similares al problema actual.
- Inspeccionar los acercamientos de minería de datos realizados con anterioridad, utilizados para resolver problemas de negocio.
- Determinar si el acercamiento revisado es aplicable al actual problema de negocio.

Identificar si es necesario contar con una variable objetivo

En este punto se debe revisar si es necesario definir una variable objetivo.
Si el problema es supervisado se debe definir la variable objetivo, si es no supervisado no se definen variables objetivo, en el capítulo I se explica estos dos tipos de análisis. Un ejemplo de análisis no supervisado es cuando se quiere entender el comportamiento de los grupos de clientes que existen en nuestro negocio, es decir, se desea encontrar características que distingan los diferentes tipos de clientes, que pueden ser: estatus económico, sexo, edad, facturación mensual, zona geográfica de procedencia, entre otros.

La presencia o ausencia de la variable objetivo impacta en los posibles análisis que se deseen realizar.

Identificar la variable objetivo (si es requerido)

Se debe identificar la variable objetivo de acuerdo al objetivo del negocio, muchas veces la definición de una variable objetivo es motivo de debate entre grupos de clientes que están incluidos en el proyecto, ya que por ejemplo, en un ambiente de ordenación de productos a través de Internet, pueden existir diferentes tipos de intereses que pueden ser: Cuáles de mis clientes van a responder a la propaganda enviada, que tan frecuentes son las adquisiciones de productos, de que tipo y cuantos son los productos que son adquiridos, etc. Es altamente recomendable que la variable objetivo sea escogida cuidadosamente no perdiendo de vista los objetivos de la empresa de negocio, institución, etc.

2.1.3 Evaluación del ambiente

Para llevar a cabo la evaluación del ambiente es necesario contar con:
La definición del problema de negocio
La definición de la variable objetivo(si es requerida)
Las fases que se incluyen en esta etapa son las siguientes:

- Evaluar el ambiente de los datos

Evaluar si las fuentes de datos son las adecuadas para las actividades de minería. Deben considerarse tanto el espacio de los equipos disponibles, tiempos de desarrollo. Si es necesario extraer información de sistemas transaccionales, debe considerarse las políticas o procedimientos que deben seguirse, además de un esfuerzo por parte de los programadores involucrados en el proyecto.

Las consideraciones del negocio deben incluirse no solamente en la decisión de que campos deben ser utilizados, sino también que archivos deberán ser utilizados o no, y que tan rápido pueden disponer de estos.

Otras consideraciones que deben realizarse son:

- Frecuencia de actualización de los datos, por ejemplo, que tan volátiles u obsoletos son.
- Contenido y nivel de detalle de los datos, por ejemplo, que tan factible o difícil es la extracción de los datos que se requieren y que tanto esfuerzo se necesita para su obtención.
- Periodo de retención de los datos, por ejemplo, ¿están disponibles los datos, de los periodos relevantes, para su utilización en el problema de negocio?

- Evaluar la agilidad organizacional

Identificar el ambiente organizacional: La identificación de los recursos externos y organizacionales que van a ser utilizados para el soporte del proyecto incluye:

- Patrocinador ejecutivo
- Líder de proyecto
- Minero de datos
- Especialista de negocio
- Supervisor de tecnologías de información
- Equipo de operaciones y de tecnologías de información

Identificación de la madurez organizacional y disponibilidad de participación

Una medida que necesita el equipo de trabajo para completar el proyecto es la madurez de la compañía para el proyecto y la disponibilidad de participación de la compañía, esto debe incluir la disponibilidad de la persona del nivel mas alto de supervisión del proyecto.

Se debe considerar para esta etapa el entendimiento que tiene la compañía sobre el proceso que se va a realizar, además de la motivación inculcada en los ejecutivos de la compañía, lo cual sirve para el soporte y ejecución de tareas en el proyecto, se deben considerar también las siguientes preguntas:

¿Qué experiencia en minería de datos tiene la compañía, y si tiene, desde qué fecha?
¿Pretende la compañía utilizar una solución en paralelo a este proyecto como propósito de prueba?
¿Es prioridad para el supervisor de más alto rango en el proyecto, proveer los recursos necesarios para concluir el proyecto de manera satisfactoria?
¿El establecimiento de recursos humanos proveídos por la compañía ha sido asignado para llevar a cabo un horario de tiempo completo o de medio tiempo en el proyecto?

Identificación de requerimientos de entrenamiento y planes de acción: En este punto se debe identificar que nivel de entrenamiento necesita el equipo de trabajo u orientación, para llevar a cabo el proyecto de manera satisfactoria, además, se tendrá que definir un plan de acción para completar las necesidades de entrenamiento.

- Evaluar el ambiente de tecnología de información

  La identificación de las plataformas de los clientes, la infraestructura de red, los sistemas operacionales, y el personal de soporte de tecnología de información, además de:

  - La infraestructura actual y la infraestructura planeada de tecnología de información para el ambiente de minería de datos del proyecto.
  - La infraestructura actual y la infraestructura planeada para los sistemas operacionales que van a ser requeridos.
  - La disponibilidad del personal de tecnología de información.

- Selección del ambiente de hardware y software

  Seleccionar la configuración de software: Basándose en el ambiente organizacional y de tecnología de información del cliente, se debe seleccionar la configuración del software apropiado para resolver los problemas de negocio.

  Seleccionar la configuración del ambiente de hardware: Basándose en el ambiente organizacional y de tecnología de información del cliente, se debe seleccionar el ambiente de hardware apropiado para resolver los problemas de negocio.

- Plan de implementación de la arquitectura

  Finalmente, los resultados del proyecto de minería de datos deben ser integrados dentro del ambiente de producción. Dependiendo de la solución del problema, la actividad de implementación puede ser el principal esfuerzo para el departamento de tecnología de información. Planear la implementación en esta fase ayuda a asegurarse que los recursos sean los apropiados y se encuentren disponibles cuando la implementación real sea requerida.

El procesamiento de grandes tablas, con ciclos frecuentes de actualización puede requerir que el diseño de la arquitectura para la implementación de los modelos utilizados en el proyecto sea completamente diferente a lo realizado de manera individual en la prueba piloto.

Los factores que son determinantes para la elección de la plataforma de implementación son los siguientes:

- Consideraciones de capacidad y desarrollo en la plataforma elegida
- Uso anticipado de resultados del proyecto
- Audiencia anticipada
- Periodo de duración anticipado del modelo
- Frecuencia de actualizaciones de fuentes de datos en los modelos desarrollados

Adicionalmente, se deben considerar los resultados de los modelos ya que pueden ser implementados dentro del un proceso de producción para que se actualicen cada vez que cambien las fuentes de información.

2.1.4 Disponibilidad de los datos (para el procesamiento)

Una vez que el problema de negocio ha sido detallado claramente, es necesario encontrar y preparar los datos que son relevantes para el proyecto.

➢ Crear el ambiente de minería de datos

Localizar las fuentes de datos internas: A menos que el *data warehouse* (almacén de datos) este disponible, se debe realizar un esfuerzo substancial para localizar y conectar las diferentes fuentes de datos que son necesarias para el problema de negocio. La mayoría de las veces los datos se encuentran en múltiples bases de datos a través de diferentes departamentos y es necesario que sean extraídas para el proceso de minería de datos. Se puede construir un nuevo ambiente de datos (*datamart*) para este fin.

Localizar fuentes de datos externas: Basados en el problema de negocio, es necesario extender el ambiente de datos para incluir la información externa.

Algunos ejemplos de fuentes de datos externos incluyen:

- Sistema de registro de información
- Datos socio-demográficos
- Datos económicos
- Entrevistas y censos públicos
- Reportes de crédito

Acceso a las fuentes de datos

Los datos pueden residir en sistemas operacionales, en un almacén de datos, en varios archivos, bases de datos, y data marts a través de diferentes departamentos. Debe tomarse una decisión sobre si el almacén de datos se debe hacer de forma separada o se mantienen los actuales repositorios y se realizan los accesos necesarios. Si datos externos son requeridos, se deben de combinar algunas de las anteriores estrategias para tener acceso a los datos. Sin embargo, se debe enfatizar que el almacén de datos de forma separada es el método más recomendable para tener acceso a las fuentes de información.

Algunas razones importantes por las que es recomendable tener un almacén de datos de forma separada son las siguientes:

- Si el espacio los permite, los datos pueden estar almacenados en una sola tabla, desnormalizada. Este procedimiento elimina la posibilidad de la posterior construcción de tablas enormes en tiempo de ejecución.

- Idealmente, el modelo de datos contiene un registro por unidad de análisis, por ejemplo, un renglón por cliente.

- Si los datos se encuentran en sistemas operacionales, estos necesitarán ser extraídos para separarlos del anterior almacenamiento, no interfiriendo en las operaciones normales realizadas. Si el espacio es un inconveniente, una muestra de la información necesita ser generada y extraída.

- Si los datos residen en un almacén de datos, es deseable que se encuentren desnormalizados todos o algunos de los datos, incluyendo el almacenamiento de datos duplicados. Si no es el caso, los datos deben ser modificados, transformados, o creados separadamente.

- Si los datos residen en varias fuentes de información a través de diferentes departamentos es recomendable la desnormalización y la agregación de estos en una base de datos o *data mart* si el espacio para el almacenamiento de la información es permitido.

- El área de trabajo es necesaria para manipular los datos que van a ser considerados. Los procedimientos usados para desnormalizar los datos pueden involucrar una cantidad enorme de datos, consultas complejas, y unión de múltiples tablas, por lo que es necesario un espacio de trabajo considerable para lo mencionado con anterioridad.

➢ Inspeccionar y validar los datos

El especialista de negocio y el minero de datos necesitan evaluar la integridad, redundancia, y forma correcta de los datos para el problema de negocio.

Garantizar la Integridad

La integridad se refiere a todos los requerimientos necesarios para que los datos sean disponibles y entendidos. Esto requiere un entendimiento del negocio con el fin de llevar a cabo este requerimiento. Cualquier deficiencia podría tener un impacto negativo en la utilidad y éxito del proyecto.

Asegurarse de la forma apropiada para los datos

Los datos necesitan ser inspeccionados para confirmar si se encuentran en la forma apropiada para el proyecto. Los siguientes puntos deben ser considerados.

- Una reducción de la complejidad es necesaria.
- Proveer datos extras, así como datos sumarizados, si es necesario.
- Puede ser necesario combinar los datos dentro de una forma significante, por ejemplo la construcción de varios estados financieros de los campos de las bases de datos.

- Evaluar la relevancia de los datos

  Ejecutar una revisión para asegurarse de que los datos seleccionados para el análisis del problema de negocio sean relevantes. El resultado de este análisis debe ser otra vez comparado con la intuición del negocio, con el objetivo de entender y reparar las deficiencias. La importancia de tener los datos adecuados antes de que el análisis detallado inicie no debe ser sobrestimado. Un método para la ejecución de esta evaluación es usar herramientas automáticas para validar la importancia de las variables con respecto a las diferentes variables objetivo. Alternativamente, si no existen variables objetivo, el investigar la dependencia entre las variables o su correlación es sumamente importante en esta etapa.

- Preparación de los datos

  Se requiere inspeccionar los diferentes valores de los datos y los temas a los que pertenecen, para realizar las transformaciones convenientes para el proyecto. Este es el primer paso a realizar en los datos, y se utiliza para corregir y validar si la calidad de los datos es la adecuada para la propuesta de análisis. La última elección entre alternativas de preparación de los datos esta incluida en una de las fases del método SEMMA para la minería de datos, de la cual se comento con anterioridad y se menciona posteriormente.

  Limpieza de datos

  La primera tarea involucrada en este análisis es la limpieza de los datos y la detección de los valores que son incorrectos. Por ejemplo, una variable numérica que deba estar dentro de un cierto rango especifico. Identificar si los datos son correctos es una tarea difícil en donde la solución puede ser aproximada. Un método para verificar la validación de los datos es la inspección de los valores *outliers* (o valores fuera de rango) de forma individual, para la cual se debe juzgar la validez de estos (por ejemplo, los salarios que se encuentran a ciertas desviaciones estándar por arriba o debajo de la media) La validación de valores *outliers* pueden ser un grupo interesante para analizar, ya que puede ser un pequeño grupo de clientes que producen ganancias por arriba de los estándares, y por lo que es necesario no eliminar dicha información.

  Eliminación de la redundancia

  La redundancia se refiere a la duplicación de la misma información dentro de diferentes campos. La redundancia puede ser directa o indirecta.

  - La redundancia directa ocurre en el momento en que la misma información ha sido extraída de diferentes fuentes de información al hacer usos de diferentes nombres de campos.
  - La redundancia indirecta ocurre cuando los mismo campos se encuentran en diferentes niveles de generalidad, esto es, en el momento que se tiene la información tanto total como disgregada, se tiene que decidir de acuerdo a la perspectiva de negocio que información tiene mayor sentido y cual debe ser ignorada.

  Identificación y manejo de valores faltantes

  Los valores faltantes o *missing* necesitan ser representados de forma clara, su permanencia necesita ser validada al igual que las implicaciones de estos en el análisis. Dependiendo del tipo de análisis las siguientes afirmaciones pueden ser aceptadas:

  - Ignorar simplemente esto valores cuando ocurran
  - Ignorar el registro completo que contenga este valor
  - Reemplazar estos valores por una medida estadística (por ejemplo la media)
  - Tener un sistema para tratar de predecir un valor apropiado

La entrada de valores faltantes o *missing* es factible solamente cuando la cantidad de valores no faltantes dentro de un campo no es muy alta. Si los valores faltantes en diferentes campos son relativamente frecuentes, el borrar completamente el registro puede no ser una opción factible, desde el hecho que se reduce el tamaño de datos en la tabla. La opción debe ser cuidadosa en medida de la información disponible.

Asegurar la consistencia de los datos

Específicamente, se debe asegurar que las medidas de los datos se encuentren en la misma escala, por ejemplo kilogramos, dólares, etc., además de no tener significados inconsistentes por ejemplo, al tener registrado ganancias antes y después de los impuestos, etc.

2.1.5 Minado de datos en ciclos

Para esta etapa las preguntas de negocio han sido clarificadas, los datos recolectados y representados de acuerdo a las necesidades del proyecto, por lo que el proceso de minería de datos puede iniciar.
Lo siguiente descripción es una visión general del método de minería de datos, ya que este se explicó a detalle en el anterior capítulo.

Antes de iniciar con esta fase se debe de contar con:

- La definición del problema de negocio.
- Definición de la variable objetivo (si fue necesaria)
- Lista de variables a incluir en el proceso de minería de datos.
- Los datos deben estar listos para aplicar la minería de datos.

Las cinco fases del método de minería de datos propuesto por el sistema SAS, denominado SEMMA, por las siglas en inglés de Sample, Explore, Modify, Model y Assess.

- Muestreo (**S**ample): La extracción de una porción de datos, puede ser la suficiente para la manipulación rápida de la información.

- Exploración(**E**xplore): En la exploración de los datos, se puede descubrir tendencias, anomalías, que puedan servir para el conocimiento de la información.

- Modificar(**M**odify): La modificación de la información puede realizarse al crear, seleccionar, y transformar las variables que sirven para el proceso de minería que se encuentra en estudio.

- Modelo(**M**odel): Modelar los datos, en este punto se puede por ejemplo dejar que el software busque automáticamente o interactivamente una combinación de predicciones confiables para los datos obtener resultados.

- Evaluación(**A**ssess): Este punto se evalúa la utilidad y la confiabilidad de los hallazgos en el proceso de minería, para seleccionar el modelo más apropiado.

2.1.6 Implementación en la producción

Los modelos que han sido probados, fueron desarrollados en un ambiente particular, tal vez similar o diferente al ambiente de producción del cliente. Por lo general, los modelos son usados para calificar al cliente por ejemplo, la probabilidad de respuesta de un cliente hacia un producto o servicio, probabilidad de que pague un préstamo realizado, etc., la mayoría de las veces esto requiere que se implemente directamente el modelo en un sistema operacional o un almacén de datos.

Esta fase incluye las siguientes actividades:

- Diseño de la arquitectura para la integración del modelo

  Consideraciones para la planeación de la implementación

  Los modelos deben tener un ciclo de vida finito. Estos posiblemente necesiten ser ejecutados una vez más incluyendo campos adicionales.

  Los datos pueden cambiar. El personal de la compañía pudo haber movido de lugar la información por tratarse de datos personales o información financiera como es la de créditos, estado civil, tipo de empleo, etc.

  Puede haber cambios regulatorios, dependiendo del ambiente competitivo de la compañía, por ejemplo agregación de nuevos productos o servicios, lo cual puede hacer que el modelo sea obsoleto.

  El modelo debe ser aplicado a datos que tienen un formato diferente al utilizado en la creación del modelo.

  Por lo anterior, es importante anticipar en que medida podemos incorporar cambios drásticos en nuestro modelo y hacérselo saber al cliente de forma cordial.

  Desarrollo de la arquitectura

  Diseño

  El plan de implementación de los modelos, debe incluir la disponibilidad de los resultados de los modelos a las unidades de negocio, si no es así no representaría un beneficio para la compañía.

  Dependiendo del tipo de modelo la actividad de implementación puede ser un esfuerzo mayor para el área de tecnología de información, para que eso se lleve de forma satisfactoria, el departamento de tecnología de información debe estar involucrado en las actividades de preparación de los datos antes de que inicie la fase de modelado de datos.

  Algunas complicaciones pueden aparecer, si existe dependencia de los resultados de un modelo, por ejemplo si la variable objetivo de un modelo es entrada para otro modelo propuesto, por lo tanto cuando existe una dependencia de este tipo es necesario realizar un plan de implementación muy eficiente para este hecho.

Creación de reportes administrativos y de negocio

Los reportes administrativos y de negocio que son requeridos por el cliente, dependen del ambiente de negocio. Sin embargo, existen algunas preguntas que pueden ser contestadas para ayudar a proveer al cliente de los reportes solicitados.

¿Qué fue lo que se encontró en el análisis realizado?
¿Qué información ayudará al manejo y análisis del desarrollo de la compañía?
¿Qué reportes de los resultados de los modelos son necesarios para los usuarios finales?

Tener los resultados disponibles para utilizarlos

Los modelos pueden ser implementados y usar una variedad de opciones que incluye la aplicación de modelos directamente a los datos, a través del los algoritmos generados por los resultados de los modelos; y de esta forma incorporarlos en un almacén de datos, o sistema operacional.

Se presentarán a continuación las opciones existentes de implementación, propuestas por el método de SAS:

- Incorporación en una aplicación de SAS: Esto ocurre cuando los resultados de los modelos son exportados a una aplicación en el mismo lenguaje y ambiente de programación, en este caso los resultados obtenido de la herramienta de minería de datos Enterprise Miner ™ son exportados con facilidad, al igual que las reglas de limpieza obtenidas de los meta-datos del diccionario de datos de la herramienta.

- Implementación en un almacén de datos de SAS: si los datos se encuentran alojados en un almacén de datos, los modelos pueden ser implementados en dos pasos:

  El primer paso debe incluir una carga inicial en el almacén de datos para aplicar los resultados obtenidos.
  El segundo paso es exportar los resultados obtenidos al modelo de almacén de datos.

  Dependiendo de la frecuencia de los cambios en las fuentes de información involucradas y de las nuevas fuentes necesarias para obtener los resultados, nuevos procedimientos pueden ser agregados en el modelo almacén de datos.

Aplicación de resultados a ambientes externos

Por lo general, es necesario exportar los resultados del modelo a un ambiente el cual no es el software SAS, una solución común a este caso, es proveer de acceso al cliente de las bases de datos y los resultados del proyecto y exportar dichos resultados a la base de datos original que el cliente maneja.

Aplicación de los resultados en los sistemas de producción

En algunas ocasiones, los resultados del proyecto de minería de datos pueden ser reescritos en un ambiente externo o directamente en un sistema de producción, y agregar este desarrollo al proyecto, y se realiza al tomar en cuenta lo dinámica que puede ser la aplicación, y el tiempo de desarrollo requerido.

- Implementación de los modelos en producción

  Existen dos pasos adicionales necesarios para que la implementación correcta de los modelos:

  El desarrollo de un programa de mantenimiento y reglas para el uso continuo del modelo.
  Creación de reportes del proceso de minería de datos y de los resultados relevantes obtenidos para el patrocinador ejecutivo del proyecto.

2.1.7 Revisión

- Revisión del impacto del modelo en la organización

  Esta revisión debe ser continua, e incluye una revisión del impacto del modelo en la organización por lo que se debe tomar en cuenta las siguientes preguntas:

  - ¿Qué cambios han ocurrido en la organización durante el curso del proyecto?
  - ¿Qué tipo de impactos tiene el modelo dentro de la organización?
  - La revisión debe incluir una reexaminación de las implicaciones de los modelos y de su uso en la organización. Por lo que se debe investigar ¿qué tanto se relaciona el modelo con el progreso de la organización?, y se deben tomar en cuenta las metas corporativas

  La organización debe tener beneficios sólidos de diferentes formas, a continuación se mencionarán algunos de estos:

  - Clientes rentables
  - La organización es vista como una organización que esta adquiriendo conocimiento
  - Los empleados están orgullosos del control de su organización, realizada diariamente.

- Revisión del retorno de la inversión

  Algunos de los beneficios que se deben considerar son los siguientes:

  Se debe comparar la ganancia obtenida antes y después del uso del modelo.
  Cuando los beneficios no se encuentran relacionados directamente con las utilidades de la organización, se deben considerar ¿Cuáles son los beneficios intangibles producidos por el uso del modelo de minería de datos?, y ¿Qué tan importantes son los beneficios intangibles para la organización?

- Revisión del desarrollo del modelo

  Se requiere determinar la validez del modelo a través del tiempo, y se debe considerar lo siguiente:

  - El resultado y el éxito del proceso son conocidos a lo largo de la organización.
  - Una vez que el proceso fue exitoso, se debe repetir el proceso y actuar enfocados en el siguiente problema de negocio.

2.2 Caso práctico

Se debe considerar, que el caso práctico propuesto solamente se involucrará en la aplicación del método de minería de datos, para este fin se toma en cuenta solamente las generalidades del proceso de planeación previo, ya que para este fin se debe involucrar a diferentes personas de negocio de la organización, y dado que el caso práctico entra realmente dentro de una prueba de concepto, que servirá para analizar solamente parte de la información del cliente, lo que se pretende a futuro, es que el método se aplique continuamente y genere resultados actualizados y confiables. Se debe tomar en cuenta que la información proporcionada fue otorgada mediante un convenio de confidencialidad por lo que no se hará mención de las políticas, procedimientos ni estructura de negocio de la organización, sin embargo se realizará un análisis de Fortalezas, Oportunidades, Debilidades y Amenazas (FODA) del proyecto.

El presente trabajo guía al consultor para la aplicación del método de MD, y da por hecho que, para un proyecto en producción, se debe realizar la planeación completa del proyecto con anticipación, y que esta, debe ser concluida una vez terminada la aplicación del método, obténgase la información necesaria descrita en el anterior subtema y refiérase a ella para la obtención del proceso de planeación mencionado.

2.2.1 Visión general

Descripción del problema de negocio

La EDVD, encargada de la venta en el ámbito nacional de artículos comestibles, ropa, calzado, juguetes, regalos, muebles, ferretería, tlapalería, farmacia, papelería, libros, etc., desea obtener conocimiento sobre los clientes más rentables de su empresa, ya que desea aplicar estrategias de mercadotecnia, dirigidas a los clientes más rentables y de esta forma obtener un mayor beneficio en las ventas generadas.

Esta empresa cuenta con un sistema de membresías, el cliente es el que determina el tipo de membresía que desea, de acuerdo a la membresía que puede ser: oro, plata o bronce, se ofrecen promociones sobre productos.

Para este proyecto, se requiere obtener conocimiento sobre los clientes potenciales, esta necesidad es muy específica y se encuentra enfocada al análisis supervisado, el cual esta descrito en el capítulo I, para lo cual es necesario la identificación de una variable objetivo, dicha variable objetivo es creada de forma binaria con valores 1 si el cliente no es rentable y 0 si el cliente es rentable, "por lo regular es más sencillo obtener clientes no rentables y discriminarlos"[14], la definición de esta variable objetivo se basa en reglas de negocio definidas por la organización y se aplican a cada caso presentado, el objetivo a través del proceso de MD es analizar cuales son los clientes con mayor probabilidad de ser rentables, y presentarlos a las personas de mercadotecnia para que empleen estrategias, que ayuden a generar mayores ingresos.

Una vez que se describió la problemática de la empresa y su necesidad, se dará un panorama general para comprender un poco la situación actual del proyecto de MD en la empresa. Presentaremos a continuación el análisis FODA, el cual nos será útil "para evidenciar los puntos fuertes del proyecto y neutralizar los débiles, además nos ayuda a aprovechar eficazmente las oportunidades que el entorno brinda y esquivar hábilmente las amenazas que se presenten."[15]

FODA es una sigla que significa Fortalezas, Oportunidades, Debilidades y Amenazas. Es el análisis de variables controlables (las debilidades y fortalezas son internas del proyecto y por lo tanto se puede actuar sobre ellas con mayor facilidad), y de variables no controlables (las oportunidades y amenazas las presenta el contexto y la mayor acción que podemos tomar con respecto a ellas es preverlas y actuar a nuestra conveniencia).

---

[14] SAS Institute, "Enterprise Miner: Applying Data Mining Techniques Course Notes", 1999.
[15] Arturo Fuentes Zenón, "Un sistema de metodologías de planeación", 1999.

FODA para el proyecto de minería de datos de una EDVD

Fortalezas	Oportunidades
• Confianza en el equipo de trabajo. • Asignación de una persona de la empresa, que apoye al proyecto para la obtención del conocimiento de los procesos involucrados. • Contar con una herramienta robusta que facilite la extracción, manipulación y entrega de información, además de proveer un ambiente y una interfase para el minado de la información. • Asignación del equipo de cómputo necesario para el proyecto y la entrega rápida de información.	• Poco tiempo de haber creado el área de minería de datos en la empresa, por lo que se buscan proyectos relacionados con este tópico. • Necesidad de la empresa de contar con herramientas que le ayude a obtener conocimiento de sus clientes para mejorar las estrategias de mercado. • La empresa requiere de herramientas que se encuentran actualizadas y que vayan a la vanguardia en cuanto al avance tecnológico. • El éxito del proyecto puede ser medible, al tomar en cuenta indicadores de costo-beneficio como el retorno de la inversión.
Debilidades	**Amenazas**
• Falta de integración en las diferentes áreas involucradas en el proyecto, que pueden ser sistemas y mercadotecnia lo que provoca retardo en la entrega de la información. • Retraso del proyecto por falta de negociación en cuanto al precio. • Entrega de información incompleta o inconsistente que puede llevar al atraso del proyecto. • Infraestructura de sistemas o redes inadecuadas para la realización del proyecto.	• Desvío del presupuesto del proyecto debido a una situación emergente de la empresa. • Despido del personal asignado al proyecto. • Ofrecimiento de mejores precios por parte de la competencia, que podría causar la cancelación del proyecto actual. • Lucha de poderes de los líderes de proyectos para manejarlo, lo que conlleva al atraso del mismo.

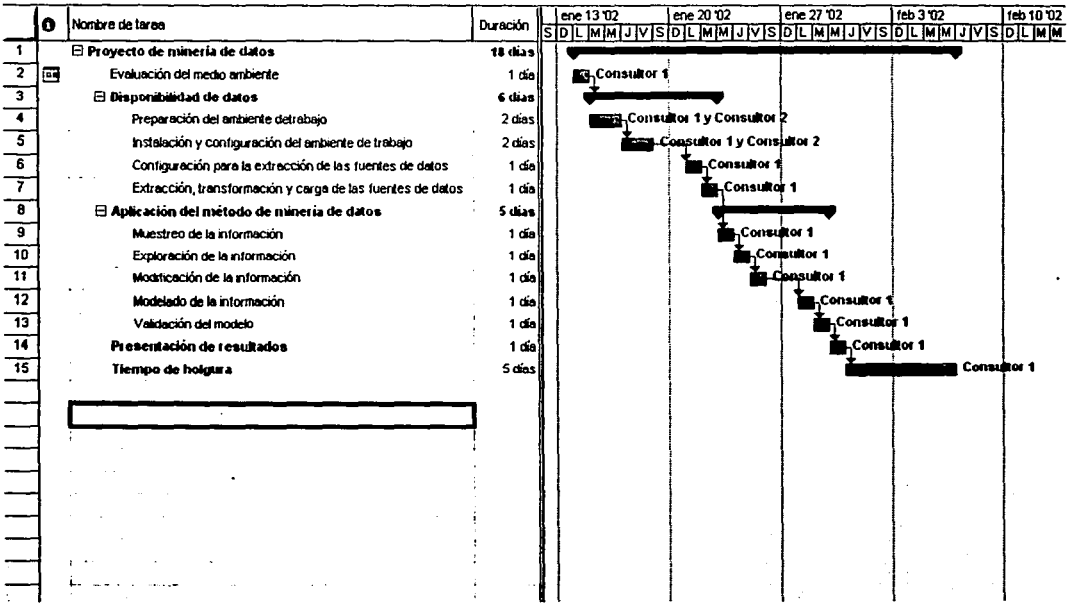
Una vez que se realizo el análisis FODA, se presentará un plan de acción, "que es una guía del desarrollo del proyecto paso a paso, con una calendarización precisa y en la que se consignan los requerimientos y responsables de cada etapa, para cuya elaboración conviene apoyarse en herramientas informáticas con que ya se cuenta y que están fundadas en las tradicionales técnicas de programación-presupuestación"[16], este plan de acción ira de la mano a la aplicación del método de planeación y del proceso de MD para la prueba de concepto antes descrita.

La siguiente tabla de actividades muestra el proceso que se llevarán a cabo para la realización del caso práctico y el control sobre las actividades que se realizan, para que en caso de tener un retraso en alguna actividad, se puedan obtener alternativas de solución en el tiempo que se requiera.

[16] Idem

Tareas a realizar

El siguiente cuadro muestra las tareas a realizar por cada consultor en el proyecto de minería de datos



Evaluación del medio ambiente

La información proporcionada para la prueba de concepto, se encontraba en formato de texto, por lo que se deben de llevar acabo técnicas de extracción y manipulación de la información para ponerla en formato propio de SAS.

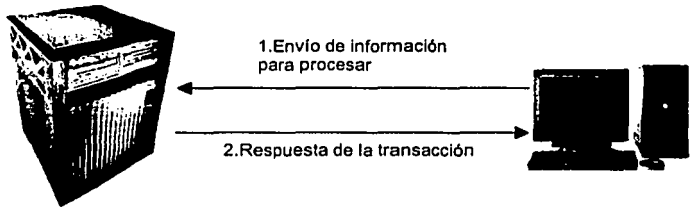
La información otorgada por la empresa es un historial de 6 meses en el cual incluye las variables que se presentan a continuación:

Descripción de La variable	Nombre de La variable	Tipo de la variable
identificador del cliente	(Id_cliente)	Carácter
Nombre del cliente	(nom_cliente)	Carácter
Iniciales del nombre	(iniciales)	Carácter
Dirección 1 del cliente	Direc_1	Carácter
Dirección 2 del cliente	Direc_2	Carácter
Dirección 3 del cliente	Direc_3	Carácter
Delegación	Delegación	Carácter
Código postal	Cod_postal	Carácter
Sexo	Sexo	Carácter
Fecha de extracción de la información	Fec_extract	Numérica
Facturación del cliente del mes 1	Fact1	Numérica
Facturación del cliente del mes 2	Fact2	Numérica
Facturación del cliente del mes 3	Fact3	Numérica
Facturación del cliente del mes 4	Fact4	Numérica
Facturación del cliente del mes 5	Fact5	Numérica
Facturación del cliente del mes 6	Fact6	Numérica
Forma de pago del cliente si fue en efectivo, tarjeta, cheque, mas usada del mes 1	Forma_pago1	Carácter
Forma de pago del cliente si fue en efectivo, tarjeta, cheque, mas usada del mes 2	Forma_pago2	Carácter
Forma de pago del cliente si fue en efectivo, tarjeta, cheque, mas usada del mes 3	Forma_pago3	Carácter
Forma de pago del cliente si fue en efectivo, tarjeta, cheque, mas usada del mes 4	Forma_pago4	Carácter
Forma de pago del cliente si fue en efectivo, tarjeta, cheque, mas usada del mes 5	Forma_pago5	Carácter
Forma de pago del cliente si fue en efectivo, tarjeta, cheque, mas usada del mes 6	Forma_pago6	Carácter
Edad del cliente	Edad	Numérica
Zona en donde vive el cliente (Zona Norte, Sur)	Region	Carácter
Tipo de membresía con la que cuenta el cliente, se divide en: oro, plata, cobre	Membresía	Carácter
Variable objetivo definida de acuerdo a las reglas del negocio del cliente, dependiendo de los lineamientos establecidos por la empresa.	Objetivo	Numérica

Disponibilidad de datos

Fue necesario crear una arquitectura cliente/servidor para optimizar el procesamiento de la información, para lo cual se instalo el sistema SAS en un servidor unix y en una PC, posteriormente se llevo a cabo la configuración necesaria para que interactuarán el servidor y la PC en una arquitectura cliente/servidor, una vez que estuvo preparada la arquitectura de trabajo, se realizo la extracción, transformación y carga de los datos, de formato texto a formato SAS.

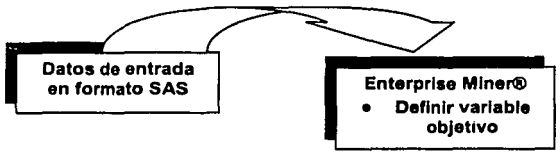
Arquitectura Cliente/Servidor



2.2.2 Aplicación del método de minería de datos.

Una vez que se tiene definido el problema de negocio, la variable objetivo (ya que es un análisis supervisado), la lista de variables a incluir en el proceso de minería de datos y el ambiente de trabajo, la aplicación del método de MD estará dirigido desde la herramienta SAS/Enterprise Miner ® descrita en el anterior capítulo.

El primer paso que se debe llevar a cabo en el proceso de MD, es la carga de la información a la herramienta de MD, y se continua con la indicación de la variable objetivo que se encuentra en el conjunto de variables introducidas.



Cuando se realiza la carga de la información y se define la variable objetivo, el sistema SAS realiza un análisis estadístico de las variables que se introdujeron al sistema.

Para las variables de clase

Nombre	Valor	Porcentaje de valores faltantes
(Id_cliente)	128	0%
(nom_cliente)	128	0%
(iniciales)	128	0%
Direc_1	128	0%
Direc_2	128	6%
Direc_3	127	54%
Delegación	16	8%
Cod_postal	128	0%
Sexo	3	11%
Fec_extract	1	0%
Forma_pago1	3	0%
Forma_pago2	3	0%
Forma_pago3	2	0%
Forma_pago4	3	0%
Forma_pago5	3	0%
Forma_pago6	2	0%
Region	5	2%
Membresía	4	0%
Objetivo	2	0%

Para la obtención de las estadísticas la herramienta SAS/ Enterprise Miner ® toma una muestra aleatoria del conjunto de información que se maneja y obtiene las estadísticas anteriores.

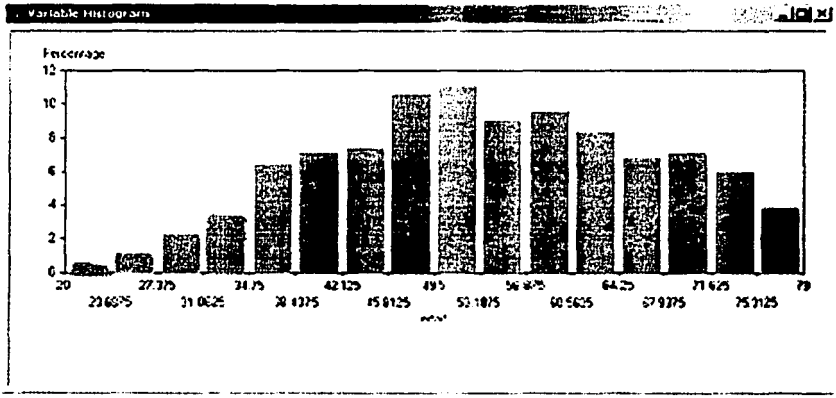
La figura anterior nos muestra el nombre de las variable de clase encontradas, sus estadísticas asociadas a cada variable que son: el número de valores encontrados en cada variable, el porcentaje de valores faltantes o missing, y el orden en el cual se encuentran las variables.

Para las variables de intervalo

Nombre	Min	Max	Mean	Desv. Standard	Valores faltantes	Sesgo	Curtosis
Edad	20	79	53.716	12.815	6%	-0.072	-.0739
Fact1	462.41	5.31E7	1.37E7	1.21E7	0%	0.9806	0.1744
Fact2	1944	5.17E7	1.38E7	1.19E7	0%	0.9145	0.0569
Fact3	999.56	5.39E7	1.36E7	1.21E7	0%	0.9774	0.1484
Fact4	648	5.41E7	1.38E7	1.23E7	0%	1	0.21
Fact5	2304	5.52E7	1.43E7	1.24E7	0%	0.9333	0.0645
Fact6	2423	5.34E7	1.41.E7	1.23E7	0%	0.9376	0.0095

En este caso se encontró siete variables de intervalo, por lo que las estadísticas asociadas por ejemplo para la edad son: el menor edad de los clientes es de 20, la mayor de 79, la media de la edad es de 53, con una desviación estándar de 12.815, el porcentaje de valores faltantes es del 6%, la distribución de la edad tiene un sesgo negativo de –0.072 por lo que la curva se encuentra sesgada un poco a la izquierda y una curtosis negativa del –0.739 por lo que la mayoría de los datos se encuentran distribuidos en el centro de la distribución y no en las colas de la curva, las últimas 2 medidas de forma se pueden comprobar viendo la distribución de la variable edad, que a continuación se presenta.

Distribución de la variable edad



Más adelante en la etapa de exploración se detallará un poco más el análisis de las variables.

2.2.2.1 Muestreo de la información

Una vez que realizamos la carga de la información y obtuvimos una primera vista del comportamiento de la información, se puede empezar con el proceso de mineria de datos y utilizar el método SEMMA, el inicio del método sugiere la realización de una muestra de la información para obtener resultados más rápidos en el procesamiento de la información, por lo que se realiza una muestra de la información y se continua posteriormente con la exploración de esta, para analizar el comportamiento de los datos involucrados en el procesamiento de MD.

Proceso de mineria de datos

La muestra que se realizará va a ser aleatoria, por lo que todos los registros tienen la misma probabilidad de ser elegidos.

Los distintos métodos de extracción que contiene SAS/Enterpise Miner® para la obtención de una muestra son los siguientes:

- Aleatoria: en el muestreo aleatorio cada observación en el conjunto de datos, tiene la misma probabilidad de ser seleccionado para la muestra.
- Cada N observaciones: genera una muestra dependiendo de un número aleatorio proporcionado por el usuario, si es el 3 toma de 3 en 3.
- Estratificada: Fuerza a la muestra para que tenga la misma proporción de información que la población, dependiendo de las variables elegidas. Por ejemplo: Si la variable seleccionada para la muestra estratificada tiene dos valores, cada uno con un porcentaje de 30% y 70% respectivamente en la población, la variable seleccionada tendrá el mismo porcentaje en la muestra.
- Muestreo por agrupación: en el muestreo de grupo, las muestras son formadas por registros que son similares en alguna forma, en la cual se especifica el número de grupos a formar.

Una vez que se ha realizado la muestra se realiza una partición de la información en tres estratos, que son:

- Entrenamiento: esta parte de la información se utiliza para realizar un ajuste preliminar del modelo que se aplicará.
- Validación: este estrato es usado para validar la adecuación del modelos, es usado para refinar el ajuste del modelo.
- Prueba: es usado para obtener la última estimación, y obtención del error mínimo para el modelo.

Los estratos anteriores son necesarios para la creación del ajuste y validación del modelo que se elegirá en este proceso de MD.

El método utilizado para la extracción de la muestra fue el aleatorio, y los resultados obtenidos son una tabla del estilo que se muestra a continuación con las mismas características que la tabla de entrada, solamente con menor número de registros.

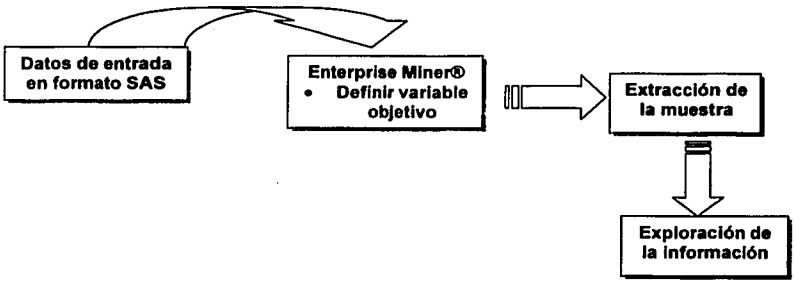
Muestra obtenida

nom_cliente	inicial	id_cliente	sex	eda	regio	membr	obj	fact_1	fact_2	fact_3	fact_4	fact_5	fact_6	forma	form	form	form	form	form
DUARTE MOROYOQUI REYES	C M	0000206032	M	75	Sur	Oro	0	66759.463	71871.3	258253	243986	1460354	480705	TC	TC	TC	TC	TC	EF
DUARTE MEJIA RAUL	E K	0000209732		55	Centr	Oro	0	205788.39	34337.4	35142.1	64808	256554.8	182322	EF	CH	TC	EF	TC	TC
	D B	0000213241	F	57		Oro	0	140391.57	63279.8	88226.8	193415	264449.4	137057	TC	TC	TC	TC	TC	EF
ROBERTO DURAN MOLINA	M E A	0000215554	F	52	Norte	Plata	1	37956.422	257816	81444.5	36729	143827.1	168924	TC	EF	TC	TC	TC	TC
ROBERTO DUERAS MICHEL	M	0000217158	F	74	Centr	Oro	0	98173.994	20427.4	105329	163680	93461.38	66296.4	TC	TC	TC	CH	TC	EF
DURAN MANCILLA RAMON EDUARDO	P J M	0000223182	F	79	Este	Plata	0	190716.92	230255	221424	255062	270819.6	138272	TC	EF	TC	TC	TC	TC
DUARTE MENDOZA RAMON	P A S	0000234540		48	Norte	Plata	0	220411.42	189470	96822.7	190766	165673.6	149040	TC	TC	TC	TC	TC	EF
DUBOIS MIRANDA ROGER A	J B	0000239844		60	Este	Bronce	0	163275.55	258369	224581	269460	179101.8	245184	TC	TC	TC	TC	TC	TC
DURAN MORENO JOSE REFUGIO	T H W	0000240051	F	38	Norte	Plata	1	109551.22	230207	119247	169708	262474.4	156225	TC	TC	TC	TC	CH	EF
ROBERTO CESAR DURAN MERAZ	J	0000248258	F	58	Norte	Plata	0	219074.8	448906	273757	247409	250034.2	62926.4	TC	EF	TC	TC	TC	TC
DURAN MENDOZA RAMONA LOURDES	A J	0000249335	M	52	Centr	Plata	0	281272.02	30026	82046.3	91238	176408.4	135261	TC	TC	TC	TC	TC	EF
DUARTE MEDINA RAMON	J T	0000252646	F	45	Norte	Plata	0	263650.33	115144	253666	185574	271232.3	217263	EF	CH	TC	EF	TC	TC
MARIA DEL ROSARIO DUARTE MADA	C T A	0000253939	F	42	Norte	Plata	0	269570.73	13823.2	98792.6	279618	85753.69	154533	TC	TC	TC	TC	TC	EF
DURAN MARTINEZ RAFAEL	N O	0000256366	U	40	Centr	Plata	0	189634.66	103663	23979.8	10908	92161.28	122124	TC	TC	TC	CH	TC	TC
	J L	0000259802		55	Centr	Oro	0	230026.35	205045	23521.2	44099	214830.9	199975	TC	TC	TC	TC	TC	EF
MARIA DEL REFUGIO DURA	H	0000262684		77	Centr	Oro	0	204969.52	196757	124730	224380	102129.4	210795	TC	EF	TC	TC	TC	TC
DURAN MARES MARIA DEL REFUGIO	J S	0000263419	F	27	Este	Bronce	1	294760.34	9610.96	67793.1	103133	126421.1	198749	TC	TC	TC	TC	TC	EF
ROSA MARIA DUCK MORENO	W	0000267236	F	60	Norte	Bronce	0	27543.75	13239.7	29636.1	196611	84105.68	128061	CH	TC	EF	TC	EF	EF
	T T	0000270962	U	58	Este	Plata	0	122477.53	40314.9	10441.4	85944	192598.1	189831	TC	TC	TC	TC	TC	EF
DUARTE MEJORADO RAMIRO	F	0000271101		54	Centr	Bronce	0	212071.45	87041.5	143980	304718	254441.9	144872	TC	TC	TC	TC	TC	TC

A continuación se analizarán los resultados obtenidos después del muestreo de la información, a través de histogramas y gráficas de cajas y brazos, para continuar con el siguiente paso del método llamado Exploración de la información.

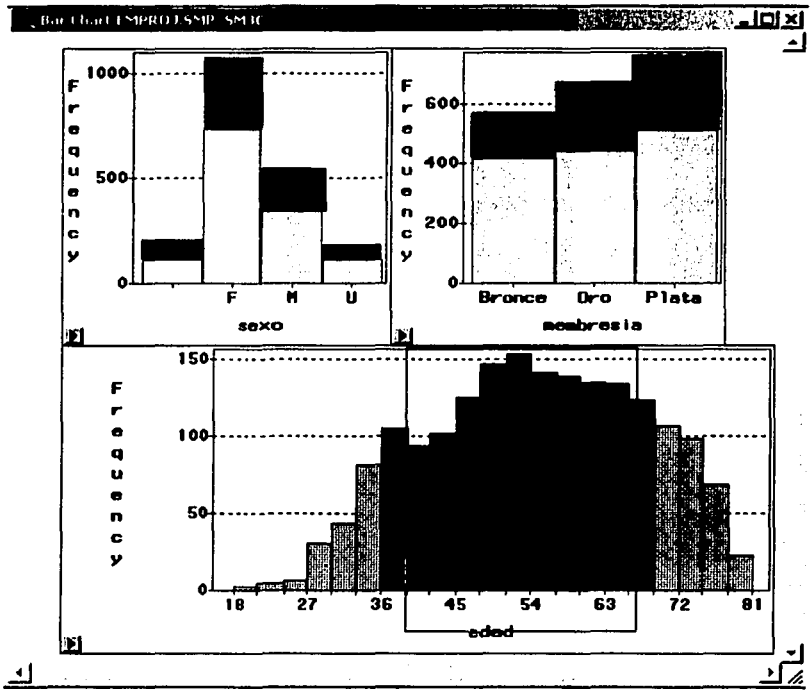
2.2.2.2 Exploración de la información

Proceso de minería de datos



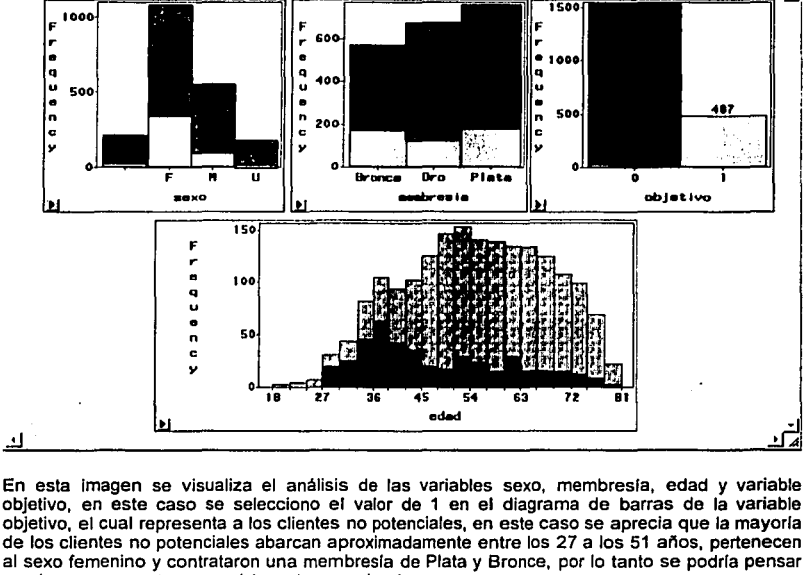
En este punto se considera el análisis de la información a través de una herramienta de visualización de la información conocida con el nombre de SAS/INSIGHT® descrita en el capítulo I, a continuación se mostrarán los resultados obtenidos en dicha herramienta.

Gráfica de distribución de las variables sexo, membresía y edad



La herramienta de exploración SAS/INSIGHT® incluida en la herramienta de MD, es interactiva, y se puede visualizar la información al tomar en cuenta diferentes variables en un solo análisis, como la gráfica que se muestra en la figura de arriba, en este análisis se seleccionó el grupo de edades más frecuentes que son aproximadamente de los 39 a los 69 años apreciados en el histograma que se encuentra en la parte inferior de la figura y en la parte superior se aprecia que predominan el sexo femenino y la membresía Plata.

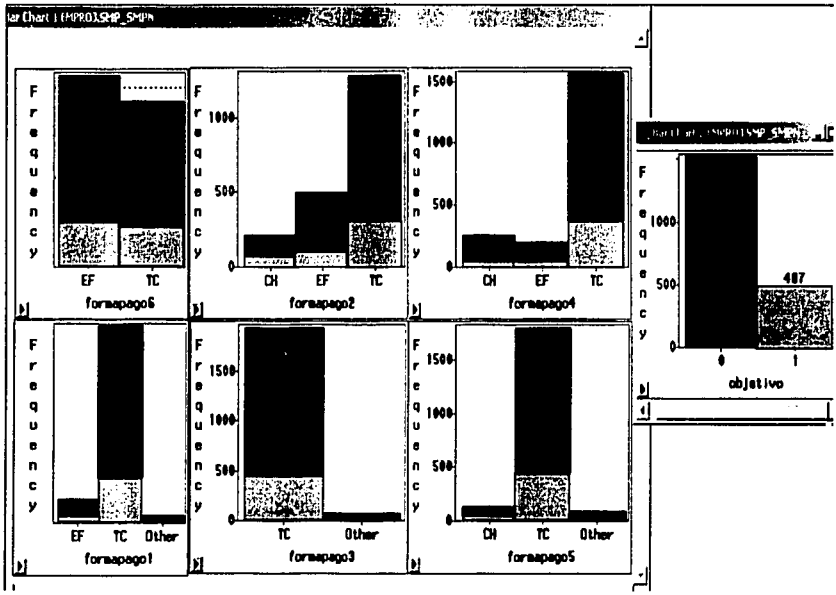
Gráficas de las variables sexo, membresía, edad y variable objetivo



En esta imagen se visualiza el análisis de las variables sexo, membresía, edad y variable objetivo, en este caso se selecciono el valor de 1 en el diagrama de barras de la variable objetivo, el cual representa a los clientes no potenciales, en este caso se aprecia que la mayoría de los clientes no potenciales abarcan aproximadamente entre los 27 a los 51 años, pertenecen al sexo femenino y contrataron una membresía de Plata y Bronce, por lo tanto se podría pensar en algunas preguntas para el área de negocio cómo son:

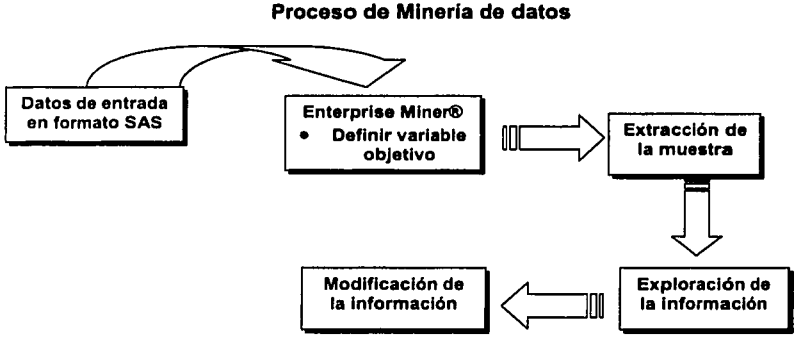
¿Qué estrategia se debe aplicar para atraer a los clientes no potenciales del tipo de membresía Bronce y Plata?
¿Qué tanto influyen las características de las membresías ofrecidas, para que sean elegidas por el sexo femenino?
¿Qué ofertas de productos son ofrecidos a los clientes entre el rango de edad elegidos como clientes no rentables?

Forma de pago de los 6 meses de historia y variable objetivo



En este caso se analizó la forma de pago de los clientes en los seis meses de historia, las variables correspondientes a los seis meses, son formapago1-formapago6 respectivamente por otro lado se selecciono el valor de 1 en la variable objetivo, por lo que se aprecia en los histogramas anteriores, que en la mayoría de los meses a excepción de 6to., predomina la forma de pago de tarjeta de crédito, y en el sexto mes el pago en efectivo supera por muy poco a la de tarjeta de crédito, por lo que el área de negocio podría analizar en este mes cuales fueron las promociones o acciones que pudieran haber influido en el que el cliente cambiará la forma de pago.

2.2.2.3 Modificación de la información



En la etapa de modificación de la información podemos realizar diferentes manipulaciones en la información las cuales pueden ser:

Transformación de variables

- Los valores de las variables pueden ser transformados y utilizar logaritmos, inversas, o simplemente aplicar una operación aritmética definida por el analista, a las diferentes variables.
- Pueden ser creadas nuevas variables
- Nuevas variables pueden ser creadas con un nuevo tipo de dato, esto es, podemos cambiar de variables de tipo carácter a numéricas.

Filtración de variables

- Pueden extraerse valores de rangos específicos y solamente tomar el rango de interés
- Los datos puedes ser eliminados basados en la frecuencia de ocurrencia
- Se pueden filtrar datos extremos que afecten a la distribución de la información

Manejo de valores faltantes

- Los valores faltantes pueden ser remplazados por algún valor estadístico, ya sea la media, la mediana o la moda.
- Se pueden reemplazar los valores faltantes al hacer uso de técnicas de árboles de decisión, los cuales analizan las características de todo el registro completo que incluya un valor faltante y busca uno similar que no tenga vacíos, una vez que es encontrado asigna el dato encontrado al faltante.

En el proceso de modificación de la información, lo que se va a realizar es la sustitución de valores faltantes en nuestros datos, en la primera etapa del proceso de minería de datos, analizamos que existe un bajo porcentaje de valores faltantes en la mayoría de las variables a

excepción de: DIREC_3 que tiene un 54% de los valores faltantes, SEXO con un 11%, las demás variables tienen un porcentaje menor al 10%.

Para esta prueba se van a sustituir los valores faltantes de la variable SEXO por el valor de NO DEFINIDO, y se van a excluir las variables que no generan mucha información cono son FEC_EXTRACT que hace referencia a la fecha de extracción de la información, y también la variable DIREC_3 ya que tiene un alto porcentaje de valores faltantes. A las variables numéricas restantes que tienen un porcentaje mínimo de valores faltantes, se reemplazará el valor faltante por la media de la variable.

El resultado de este proceso es la tabla utilizada para la muestra, con modificaciones en la variable SEXO, eliminación de las 2 variables mencionadas con anterioridad, además de la sustitución de las variables numéricas por la media. La tabla resultante es la que se utilizará en los posteriores análisis.

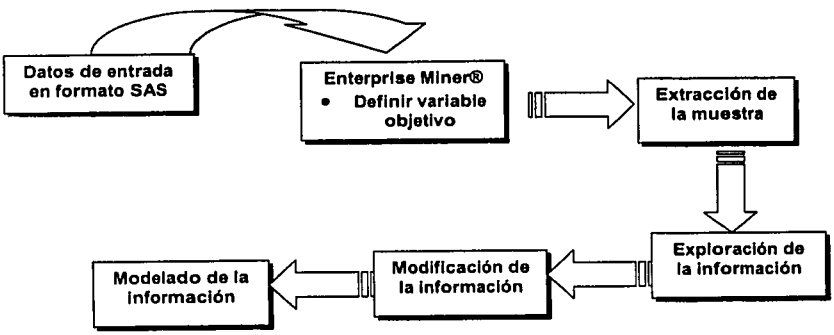
Tabla resultante

nom_cliente	iniciales	region	sexo	id_cliente	edad	membr	objetiv
OSCAR DURAN MEZA	M M	Norte	M	000361041	58	Bronce	0
DURAN MEZA RAUL	J R	Oeste	F	000361336	50	Oro	0
DUARTE MOINA ROBERTO	A G P	Norte	F	000363973	78	Oro	0
DURAN MELENDREZ ROBERTO	T J	Este	M	000365739	62	Plata	0
DURAN MARQUEZ REFUGIO	M D	Norte	F	000368120	33	Bronce	1
REFUGIO DURAN MARQUEZ	L B H	Centro	F	000368995	59	Plata	0
REFUGIO DURON MARQUEZ	J K	Norte	F	000372080	48	Plata	0
JOSE REFUGIO DURON MARQUEZ	T C J	Norte	NO DEFINIDO	000372250	60	Oro	0
DUARTE MALDONADO RUBEN DARIO	A R B	Este	U	000375148		Bronce	0
	A	Centro	NO DEFINIDO	000376966	41	Plata	1
DUENAS MALDONADO RAMON ALFONSO	M E	Centro	F	000377579	71	Plata	1
DUE-AS MALDONADO RAMON ALFONSO	L W	Centro	F	000378607	32	Plata	1
DURAN MARCIAL ROLANDO	W G	Centro	F	000379304	49	Bronce	0
DURAN MACHUCA ROSARIO	A W	Centro	M	000380294	55	Plata	0
DURAN MACHUCA ROSARIO	D L	Sur	NO DEFINIDO	000381995	61	Plata	0
DUARTE MORALES ROSALBA	B T	Centro	M	000383314	62	Oro	0
DURAN MENDOZA ROSA LINDA	N R	Norte	NO DEFINIDO	000384585	58	Oro	0

Tabla modificada

2.2.2.4 Modelado de la información

El modelado de la información se refiere a la construcción de un modelo que de respuesta a la pregunta de negocio. Cualquiera de las técnicas de MD proveen una respuesta satisfactoria la cual es llamada modelo. [17]



La elección de las técnicas dependen del problema que se plantea, al igual que del tipo de medida que tenga la variable objetivo ya que esta puede ser de intervalo, binaria, ordinal, nominal o se puede dar el caso que el problema de negocio no tenga una variable objetivo definida.

A continuación se presentarán los tipos de técnicas que se aplican para los diferentes problemas de negocio los cuales se clasifican por el tipo de análisis que se lleva a cabo ya sea supervisado o no supervisado, además también se debe considerar el tipo de técnica de MD que se debe aplicar de acuerdo a la variable objetivo que se maneje.

[17] SAS Institute, "Predictive Modeling Using Enterprise Miner Course Notes", 2001

Técnicas de MD dependiendo del análisis del problema de negocio

	Análisis supervisado	**Análisis no supervisado**
	Técnica de minería de datos	
Modelos predictivos	Árboles de decisión	No factible
	Redes Neuronales	
	Regresiones	
Segmentación	Árboles de decisión	Técnicas de agrupamiento
	Redes neuronales	
Análisis de relación	No factible	Técnicas de asociación

Técnicas de MD dependiendo del análisis de la variable objetivo definida

Para este caso de estudio la variable que se esta manejando es binaria, por lo que le corresponden los modelos de árboles de decisión, redes neuronales y regresiones logísticas.

	Tipo de variable	**Tipos de datos de entrada**	
		Cualitativos	**Cuantitativos**
OBJETIVO	Nominal	Árboles de decisión	Árboles de decisión
		Redes Neuronales	Redes Neuronales
	Binaria	**Árboles de decisión**	**Árboles de decisión**
		Redes Neuronales	**Redes Neuronales**
		Regresión Logística	**Regresión Logística**
	Ordinal	Árboles de decisión	Árboles de decisión
		Redes Neuronales	Redes Neuronales
		Regresión Logística	Regresión Logística
	Intervalo	Árboles de decisión	Árboles de decisión
		Redes Neuronales	Redes Neuronales
		Regresión Logística	Regresión Logística
	Variable objetivo no Definida	Técnicas de asociación	Técnicas de agrupamiento

De acuerdo a la información antes presentada y tomando en cuenta el análisis supervisado que se esta manejando y la variable objetivo de tipo binaria con la que se trabaja, las técnicas posibles y factibles que se deben aplicar en este caso de estudio son: regresiones logísticas, árboles de decisión y redes neuronales.

Para llevar a cabo este proceso se toma como entrada la tabla que se modifico anteriormente, los modelos que se van a aplicar en este caso son regresiones logísticas y árboles de decisión el cual tendrá el criterio de división del CHAID, el cual es el criterio mayor conocido por la desarrolladora del trabajo, las redes neuronales aunque son factibles para este caso de estudio, no se aplicarán ya que no se cuenta con la experiencia ni conocimiento sobre esta técnica de MD.

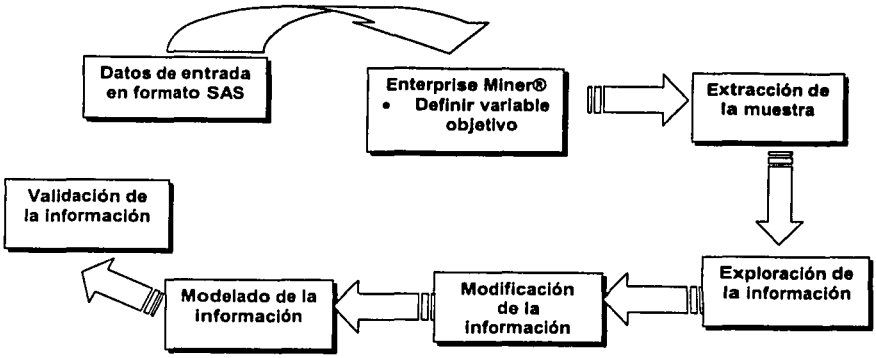
Una ventaja principal al utilizar una herramienta de *software* robusta, es que cuenta con modelos definidos para su aplicación y no hace falta crearlos a través de programación; una vez que tenemos conocimiento sobre los modelos que se van aplicar, estos se ejecutan sobre la tabla de entrada para obtener una tabla final por cada modelo aplicado, todos estos con una columna extra, que es la probabilidad de que el cliente sea y no sea potencial, para poder elegir cual de las respuestas de los modelos es la más acertada, se debe realizar una validación de los modelos para saber cual de estos dio un porcentaje de respuesta mayor, por lo que para este fin, se realizará la etapa de Validación.

Tabla con la probabilidad de que el cliente sea no potencia y potencial

nom_cliente	iniciales	region	sexo	objetivo	Predicted: EVENT for OBJETIVO	Predicted: NO EVENT for OBJETIVO
DIAZ CERDA ALEJANDRO	R M	Centro	F	1	0.0104365754	0.9895634246
DIAZ CERDA ALEJANDRO	P H	Centro	F	0	0.2295573893	0.7704426107
DOMINGUEZ ZABALETA ROSA	D H	Centro	F	1	0.4860004778	0.5139995222
	W M	Centro	F	1	0.2295573893	0.7704426107
MEDINA SOSA JORGE	F A J	Centro	F	1	0.1250916061	0.8749083939
DAVILA VENEGAS JOSE JARIN	D I	Norte	F	1	0.2295573893	0.7704426107
CONTRERAS JOSE ANTONIO MANDUJANO	G S	Centro	F	0	0.2295573893	0.7704426107
PADILLA RIVAS CONRADO	D	Norte	F	0	0.1188637366	0.8811362634
LOPEZ ESCALANTE DANIEL	D C	Centro	M	0	0.0343836428	0.9656163572
	J F		F	0	0.0775024005	0.9224975995
CASTILLO PEDRO LUIS GALLEGOS	S	Norte	NO DEFINIDO	0	0.2295573893	0.7704426107
SOBERANIS DE LA BARRERA DANIEL	J W	Este	M	1	0.0407342132	0.9592657868
MARTINEZ ABURTO HERON	A P	Centro	F	0	0.758002678	0.241997322
MENDOZA CABRAL DAVID	B K	Este	F	1	0.2295573893	0.7704426107
HERNANDEZ ORNELAS DULCE MARIA	H J H	Este	M	0	0.2295573893	0.7704426107
DE LA O DIAZ EDSON CARLOS	A A	Centro	F	1	0.6918799595	0.3081200405
GONZALEZ BENITO VILLEGAS	H E	Centro	F	0	0.5626905991	0.4373094009
	C D	Centro	U	0	0.2295573893	0.7704426107
MELLADO GONZALEZ JOSE	J G	Sur	M	0	0.0509374703	0.9490625297
MATA TOMAS PEREZ	N J	Este	M	0	0.2295573893	0.7704426107
IGLESIAS COLIN PEDRO	H	Centro	F	0	0.2295573893	0.7704426107
JORGE ANTONIO DIAZ GARCIA	S	Centro	F	1	0.0222556382	0.9777443618

Valor predicho

2.2.2.5 Validación



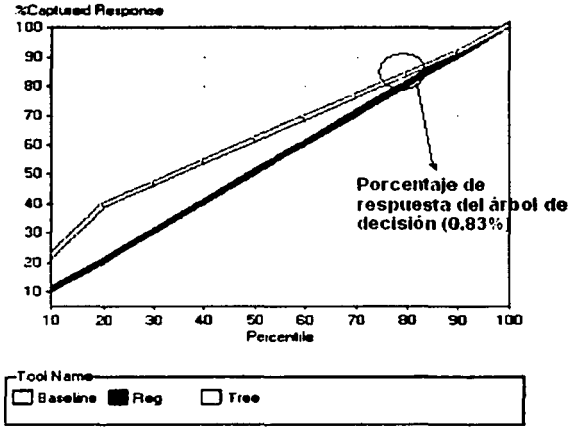
La etapa de validación es la parte final del proceso de MD, más no del proyecto, ya que una vez que se elige el mejor modelo, se debe preparar el ambiente que requiere la empresa para poner este proceso en producción, la etapa de validación contenida por la herramienta SAS/Enterprise Miner® provee de diagramas que permiten observar de manera gráfica las respuestas obtenidas por los diferentes modelos aplicados.

En los diagramas comparativos se muestra el porcentaje de respuesta de cada modelo, los porcentajes de respuesta se obtienen al ordenar en forma ascendente la probabilidad generada por el modelo, para cada registro. Una vez que se tiene ordenada la probabilidad, se divide la información en 10 percentiles y se grafican dichas probabilidades por cada percentil.

La gráfica resultante contiene dos líneas que representan a cada uno de los modelos, esta se interpreta de la siguiente manera: el modelo que produzca mayor probabilidad, es el que se ajusta mejor a la información, por lo que es el modelo a elegir.

Para este caso el mayor porcentaje de respuesta obtenido fue el del árbol de decisión, la gráfica que a continuación se presenta muestra que si se elige el 80% de la información o de los clientes, se tiene una probabilidad del 0.83% de que el cliente calificado como cliente no rentable lo sea verdaderamente.

Gráfica de validación de modelos



Los resultados obtenidos al utilizar la técnica de árboles de decisión son los que tienen un porcentaje de respuesta mayor a los de la regresión, por lo que el modelo generado por el árbol es que se debe aplicar para la solución a la pregunta de negocio de ¿Quiénes son mis clientes potenciales?, ya que se había mencionado con anterioridad, que se genera una tabla con la probabilidad de que el cliente sea o no potencial, por lo que dependiendo de las calificaciones obtenidas para cada cliente, estos resultados pueden ayudar al tomador de decisiones a seleccionar estrategias de *marketing* enfocadas a los clientes más rentables que ayuden a generar mayores ingresos.

Este proyecto no pretende que se tome en cuenta para una implementación en la producción, ya que faltaron modelos por aplicar por lo que no se podría decir que los resultados son definitivos, y que los árboles de decisión son la mejor opción para este caso de estudio, sin embargo se deben mencionar algunas características importantes del siguiente paso que es la puesta en producción de los modelos, a continuación se mencionarán algunos puntos a considerar en dicho proceso.

- Una parte importante dentro de la implementación de los resultados del modelo es tenerlos disponibles para las diferentes áreas de negocio involucradas, ya que de otra forma no se podría considerar el beneficio para la empresa.

- Dependiendo del tipo de modelo, la actividad de implementación puede ser manejada solamente con el soporte del departamento de Tecnologías de Información(TI), para que este proceso sea exitoso, el departamento de TI debe estar involucrado en el proceso de MD, tanto en la preparación de actividades a lo largo del proyecto, y en los tiempos estipulados para cada etapa del mismo.

El proyecto termina cuando todo el proceso de MD es puesto en producción y los miembros de la empresa toman el control sobre el mismo, además de llevar a cabo la revisión de los beneficios de los modelos en la organización y se lleva a cabo pocos meses después de la implementación en la producción, refierase a la sección 2.1 para la información sobre las medidas de implementación en producción y revisión del proyecto.

En el siguiente capítulo se analizarán los resultados obtenidos por el modelo de árboles de decisión, y se describirán los alcances y limitaciones del método aplicado a dicho caso de estudio, y se describirá la relación que tiene la MD en el proceso de toma de decisiones.

Conclusiones

No es difícil inferir que cualquier persona que se encuentre en un proyecto le gustaría obtener resultados exitosos, por lo que se debe llevar un control de: las actividades que se van a realizar, los recursos involucrados, y los gastos asociados al proyecto, para tal fin, se debe considerar un proceso de planeación de proyecto y un proceso de mantenimiento o soporte si fuese requerido.

El trabajo de investigación presentado hace uso del método de planeación descrito en la sección 2.1 y la aplicación del método de MD del sistema SAS, incluyendo algunos conceptos de planeación estratégica necesarios como lo es el análisis FODA, detallado en el apartado 2.2. A través de la planeación de un proyecto podemos tener una estimación de tiempo y esfuerzo necesario, además de poder detectar si existe algún desvío con respecto a lo que se quiere o espera.

La aplicación del método de MD a un caso práctico es una guía que permite analizar un proyecto en producción, estudiando las fortalezas, oportunidades, debilidades y amenazas que se presenten, además, permite al lector interesado en minería de datos o al consultor sistema SAS involucrase en los conceptos de MD, en el tipo de planeación que se debe aplicar en un proyecto y ofrece una visión general del manejo de la herramienta SAS/Enterprise Miner®. A través de todo el capítulo II se describió el método de planeación para la aplicación de un proyecto de MD, y se aplico el método de MD enfocado a una de las empresas dedicadas a las ventas a detalle (EDVD), con el fin de obtener a los clientes potenciales de la empresa, cabe señalar que no se aplicaron en este caso práctico todos los métodos estadísticos y matemáticos de MD conocidos y mencionados en la sección 2.2.2.4, por falta de experiencia en ellas, por lo que no se deben considerar los resultados obtenidos como los finales, ya que se estarían excluyendo métodos que podrían mejorar los resultados. Sin embargo este trabajo especifica de forma detallada el proceso a seguir en un proyecto de MD el cual se aplicará de forma eficiente mientras se tenga mayor conocimiento sobre los métodos matemáticos y estadísticos que se deben aplicar.

A pesar de que en el proceso de MD pueden ocurrir eventos no previstos en las diferentes áreas de negocio involucradas, la continuidad de la planeación nos dará pautas para la aplicación de planes de contingencia y de planteamientos de alternativas de solución.

Fuentes de consulta

10. C.P. y M.A: Mária Luisa Saavedra Gracía, "Un esquema de Planeación desde el enfoque sistémico"

11. Ídem

12. Ïdem

13. SAS Institute Inc., "Manual of Data Mining Project Methodology", Cary, N.C.,Mayo 2000.

14. SAS Institute, "Enterprise Miner: Applying Data Mining Techniques Course Notes", 1999.

15. Arturo Fuentes Zenón, "Un sistema de metodologías de planeación, COMISIÓN NACIONAL DEL AGUA, Subdirección general de programación, Agosto 1999.

16. Ídem

17. SAS Institute, "Predictive Modeling Using Enterprise Miner Course Notes", 2001.

Otras fuentes de consultas

1. Olivia Parr Rud, "Data Mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management", Edit. John Wiley & Sons, 2000.

2. Alex Berson, Stephen Smith, Kurt Thearling, "Building Data Mining Applications for CRM", Edit. MacGraw-Hill, 2000.

Capitulo III

Análisis de los resultados

Objetivo

Examinar los resultados obtenidos al aplicar un método de minería de datos, y definir los alcances y limitaciones del mismo.

En los capítulos anteriores se ha descrito uno de los negocios empresariales más importantes en nuestra vida cotidiana, que son las EDVD, también se ha hablado del concepto de la MD y del proceso de planeación y ejecución que tiene, además se presento un caso práctico involucrado en el proceso de MD el cual siguió el objetivo del presente trabajo que es el obtener clientes potenciales, esta tarea conduce la obtención de mejores prospectos y el manejo exitoso de las relaciones con el cliente, ya que una vez que se tiene el conocimiento sobre quiénes son mis clientes más rentables y las características de su comportamiento, esto se puede usar para realizar una prospección orientada a los intereses de la empresa. Y uno de los problemas a resolver para la realización de una prospección efectiva es el conocer cómo encontrar prospectos parecidos a los clientes potenciales que se encuentran en el negocio. Para esto primero debemos tener el conocimiento de qué clientes son más rentables para posteriormente elegir un perfil de cliente a seguir. La meta en todos los casos de las EDVD es identificar qué es lo que guía al cliente a una mayor rentabilidad.

Así, una vez que se obtenga la mayor rentabilidad de los clientes que equivale a una mayor lealtad, la empresa podrá generar mayores ingresos y ser competitiva en el mercado.

En el anterior capítulo se concluyo con la obtención de la tabla la cual contiene la probabilidad de que el cliente sea rentable o no, por lo que el principal objetivo del presente es mostrar los resultados obtenidos y ver los alcances y límites del proyecto presentado.

3.1 Presentación de resultados

En el proceso realizado de MD, se llevaron a cabo las etapas de selección, exploración, modificación, modelado, y validación, los datos de entrada de este proceso fueron las características de los clientes de las EDVD, y la salida fue una tabla con las mismas características que la de entrada, más dos variables que hacen referencia a la probabilidad de que el cliente sea potencial o no lo sea, esta calificación es asignada a cada cliente o registro de entrada.

La predicción obtenida usualmente es llamado *score*, el cual típicamente es un valor numérico que es obtenido al ejecutar un modelo sobre una tabla de datos.

En la etapa de validación del capítulo II, el modelo que genero mayor porcentaje de respuesta fue el árbol de decisión, el cual crea divisiones en los datos al aplicar una serie de reglas simples. Cada regla asigna una observación o registro a cada uno de los segmentos que se generan. Una regla es aplicada después de la otra, para la obtención de una jerarquía de segmentos, dicha jerarquía es conocida con el nombre de árbol, y a cada segmento generado es llamado nodo. El primer segmento contiene el conjunto total de datos y es llamado nodo raíz del árbol, a partir de este se desprenden nodos sucesores llamados ramas del árbol, finalmente los nodos finales son llamados hojas.

La predicción es el principal objetivo del proyecto, por o que la creación del modelo predictivo no fue una tarea sencilla, sino que se llevo a cabo a través de diferentes pasos como fueron:

La selección de variables

Hay veces que la información que se esta manejando incluye mucha basura, la cual no es interesante para el análisis, ya que el análisis debe encontrar relaciones interesantes y desconocidas entre las variables.

Existen variables que pueden ser redundantes y deben ser excluidas del caso de estudio, por ejemplo las variables que tienen alguna correlación con otras.

En este caso de estudio se excluyeron las variables que contenían un alto porcentaje de valores faltantes y variables basura como lo fue la fecha de carga de la información, la cual repetía el mismo valor para cada uno de los registros.

Manejo de los valores faltantes

La mayoría de los análisis contienen valores faltantes a lo largo de las variables de entrada. La sustitución de estos valores es necesaria en algunos modelos de MD, ya que en los modelos de regresión, los registros que contienen valores faltantes son eliminados del análisis, lo que provoca una gran alteración en los resultados. El árbol de decisión no es afectado por la anterior regla, ya que toma como un grupo diferente a los registros que tienen valores faltantes y trata de ir segmentando es grupo.

Para este caso de estudio se llevo a cabo el reemplazo de valores faltantes de las variables numéricas por la media y en la variable SEXO por la constante NO DEFINIDO.

Selección del modelo

Los modelos que se deben ejecutar en cualquier caso de estudio, debe llevar un análisis previo de la información y obviamente los resultados del modelo deben responder a la pregunta de negocio planteada.

El primer punto que debe ser conocido es el tipo de análisis que se esta llevando a cabo, ya sea supervisado o no supervisado, después se debe analizar el tipo de variable objetivo que se esta manejando (si es requerida) que puede ser binaria, de intervalo, nominal u ordinal, y una vez que se tengan estos datos listos, se puede elegir el modelo a aplicar.

En este caso se esta manejando un análisis supervisado, ya que se cuenta con una variable objetivo de tipo binaria, por lo que para este fin se eligieron regresiones logísticas y árboles de decisión, refierase al capítulo I y II, para una descripción detallada de los modelos.

Interpretación del modelo

La interpretación de los modelos, requiere a una persona que tenga conocimiento en los modelos aplicados y en la herramienta que arroja estos resultados. Uno de los modelos más sencillos de interpretar es sin duda el árbol de decisión, ya que este método arroja un diagrama de un árbol de acuerdo a la relación de las variables obtenidas.

Los árboles son populares porque por que además de ser sencillos de entender, son fáciles de aplicar, ya que no es necesaria una preparación de los datos ya que acepta variables de intervalo, ordinal y nominal y valores faltantes. Son fáciles de interpretar porque cada regla de partición se enfoca a una sola variable.

Algunos beneficios de los árboles de decisión son:

- **Interpretabilidad**
  Presentación en estructura de Árbol.
- **Escalas de medida mixtas**
  Nominal, ordinal, intervalo.
  Árboles de Regresión.
- **Robustez**
  Valores ausentes o missing.

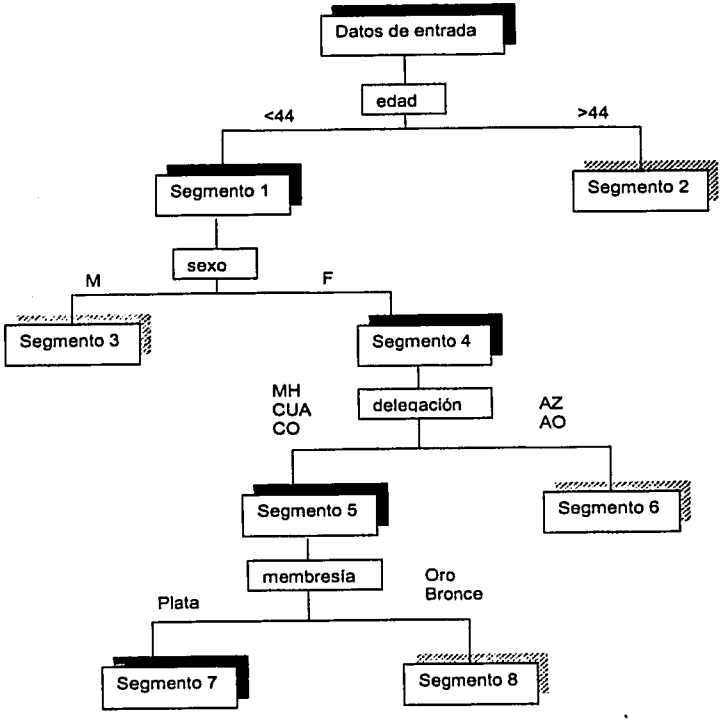
El diagrama de árbol es útil para valorar qué variables son importantes y cómo interactúan la una con la otra. A veces los resultados pueden aparecer escritos como reglas sencillas.

Las divisiones basadas en variables de entrada numéricas dependen sólo del orden de los valores. Al igual que muchos otros métodos, los árboles son métodos robustos para talar los valores extremos en el espacio de entrada.

El particionamiento recursivo tiene formas especiales de tratar los valores ausentes. Una manera, es tratar los valores ausentes como un nivel separado de variables de entrada. Los valores ausentes se pueden agrupar con otros valores en un nodo o bien tener su propio nodo.

A continuación se presentan los resultados obtenidos por el árbol de decisión de la herramienta SAS/Enterprise Miner® y se analizarán los segmentos de cada nodo que representan la divisiones realizadas de acuerdo a las características de los clientes de las EDVD.

Arbol de decisión generado por SAS/Enterprise Miner®



El anterior árbol nos hace referencia a las reglas utilizadas para la generación de la probabilidad asignada a cada cliente, los recuadros sombreados más oscuros son aquellos que se toman en cuenta para la determinación de los clientes potenciales.

Las reglas anteriores indican la siguiente secuencia en las decisiones : si el cliente es menor a 44 años, del sexo femenino, de la delegación Miguel Hidalgo, Cuauhtemoc o Coyoacan y tiene una membresía de plata, se tiene una probabilidad alta de que el cliente sea potencial, aquellos que no entran dentro de las reglas de decisión antes mencionadas tendrán una probabilidad alta de ser clientes no potenciales.

La probabilidad de ser un cliente potencial va a ser mayor mientras las características del cliente sean las mencionadas con anterioridad, y va ir disminuyendo cuando haya menores coincidencias a dichas características.

El resultado final del proceso de MD es una tabla como se muestra a continuación con todas las variables de entrada, en las que se encuentran agregadas dos columnas, que contienen la probabilidad de que el cliente no sea potencial, de aquí podremos discriminar los clientes no potenciales y analizar a los potenciales.

nom_cliente	iniciales	region	sexo	objetivo	Predicted: EVENT for OBJETIVO	Predicted: NO EVENT for OBJETIVO
DIAZ CERDA ALEJANDRO	R M	Centro	F	1	0.0104365754	0.9895634246
DIAZ CERDA ALEJANDRO	P H	Centro	F	0	0.2295573893	0.7704426107
DOMINGUEZ ZABALETA ROSA	D H	Centro	F	1	0.4860004778	0.5139995222
	W M	Centro	F	1	0.2295573893	0.7704426107
MEDINA SOSA JORGE	F A J	Centro	F	1	0.1250916061	0.8749083939
DAVILA VENEGAS JOSE JARIN	D I	Norte	F	1	0.2295573893	0.7704426107
CONTRERAS JOSE ANTONIO MANDUJANO	G S	Centro	F	0	0.2295573893	0.7704426107
PADILLA RIVAS CONRADO	D	Norte	F	0	0.1188637366	0.8811362634
LOPEZ ESCALANTE DANIEL	D C	Centro	M	0	0.0343836428	0.9656163572
	J F		F	0	0.0775024005	0.9224975995
CASTILLO PEDRO LUIS GALLEGOS	S	Norte	NO DEFINIDO	0	0.2295573893	0.7704426107
SOBERANIS DE LA BARRERA DANIEL	J W	Este	M	1	0.0407342132	0.9592657868
MARTINEZ ABURTO HERON	A P	Centro	F	0	0.758002678	0.241997322
MENDOZA CABRAL DAVID	B K	Este	F	1	0.2295573893	0.7704426107
HERNANDEZ ORNELAS DULCE MARIA	H J H	Este	M	0	0.2295573893	0.7704426107
DE LA O DIAZ EDSON CARLOS	A A	Centro	F	1	0.6918799595	0.3081200405
GONZALEZ BENITO VILLEGAS	H E	Centro	F	0	0.5626905991	0.4373094009
	C D	Centro	U	0	0.2295573893	0.7704426107
MELLADO GONZALEZ JOSE	J G	Sur	M	0	0.0509374703	0.9490625297
MATA TOMAS PEREZ	N J	Este	M	0	0.2295573893	0.7704426107
IGLESIAS COLIN PEDRO	H	Centro	F	0	0.2295573893	0.7704426107
JORGE ANTONIO DIAZ GARCIA	S	Centro	F	1	0.0222556382	0.9777443618

Valor predicho

El anterior resultado ayudará al tomador de decisiones del área de negocio a generar estrategias de mercado enfocadas a los clientes potenciales que sirvan para generar mayores ingresos y a evaluar cuáles serán los clientes que deberían aumentar su consumo.

La toma de decisiones

"La toma de decisiones es un proceso de pensamiento que ocupa toda la actividad que tiene por fin solucionar problemas"[18], en este caso el problema es la generación de ingresos de las EDVD, cabe mencionar que el presente trabajo no esta enfocado a la toma de decisiones, si no a la ayuda para la toma de decisiones que realiza la persona de negocio.

El soporte para la toma de decisiones, la mayoría de las veces se refiere al término de negocios inteligentes (*Business Intelligence*), y se define como el proceso de interpretación de una cantidad enorme de datos con el propósito de ayudar a la toma de decisiones de negocio importantes.
El soporte de decisiones, esta relacionado la mayoría de las veces con un repositorio de información que contiene la historia del negocio. Los datos son recolectados de una variedad de sistemas transaccionales, los cuales son actualizados en intervalos cortos de tiempo, que permitan analizar una toma histórica de los datos completos de la organización.

La MD es una parte primaria dentro del soporte de decisiones, ya que emplea diferentes métodos matemáticos que permiten formular una hipótesis. Por lo regular cuando realizamos reportes o consultas a nuestra información almacenada, los especialistas del negocio determinan que es lo que quieren encontrar sobre el histórico de los datos, en cambio en el proceso de MD se basa en una pregunta de negocio en específico que va dirigida a una predicción en un futuro próximo y al planteamiento de una hipótesis. Esto se analiza muy claramente en el proceso realizado de MD; ya que se planteo una pregunta de negocio que fue ¿Quiénes son mis clientes potenciales?, se obtuvo una predicción con base en la relación de las variables presentadas y a partir de los resultados surgen nuevos planteamientos de hipótesis, cómo por ejemplo: "La segmentación más importante de mis clientes potenciales se debe hacer por la variable de edad", al aplicar técnicas de MD se puede rechazar o aceptar la hipótesis antes declarada.
Sin embargo el núcleo de la MD es, no solamente contestar a las hipótesis planteadas, si no sugerir nuevas, basadas en el proceso de búsqueda de patrones de conducta existentes.

El poder del soporte de decisiones permite refinar continuamente las estructura y las prácticas para incrementar la productividad y rentabilidad de la empresa. La MD es una de los métodos más avanzados para analizar los negocios empresariales en un periodo de tiempo.

Otro factor importante en la toma de decisiones dentro de una organización, es que esta se debe realizar de dos formas, la primera debe estar enfocada a problemas particulares de las diferentes áreas de negocio, pero a la vez y sería la segunda forma, deben estar enfocados los esfuerzos individuales en las decisiones grupales, para que haya en la organización una mejora grupal e integral y no particular.

Así, como la MD requiere el esfuerzo de muchas áreas de negocio para la aplicación de este proceso, de igual forma el beneficio obtenido esta dirigido no tan solo a un miembro de la empresa, si no a todas las áreas involucradas en este proceso.

A continuación se presentan los beneficios obtenidos al utilizar el proceso de MD propuesto con anterioridad.

[18] John P. Van Gigh, "Teoría general de sistemas, edit. trillas, 1981

Impacto de la aplicación de MD, vista desde la perspectiva del área de negocio

Problema: Obtención de clientes potenciales y generación de mayores ingresos.		
	Alternativa 1 Utilizar el proceso de minería de datos para la obtención de clientes potenciales que provocará el aumento de ingresos.	Alternativa 2 No utilizar el proceso de minería de datos.
En los consumidores	Se tendrá un conocimiento de los clientes más rentables, que ayudará a satisfacer mejor sus demandas.	No se tendrá conocimiento del cliente potencial, por lo que no habrá un cambio drástico en forma de trabajo de la empresa.
Sobre el área de mercadotecnia	Se conocerá a qué clientes se deben enfocar sus esfuerzos y satisfacción, teniendo como soporte una herramienta estadística y darán mejores resultados en un tiempo optimo.	Las estrategias de mercado no tendrán un soporte matemático para realizar su publicidad y se llevarán a cabo al utilizar su método tradicional, y no presentaran resultados confiables, ni óptimos.
En el área de ventas y finanzas	Se superarán las metas establecidas de ventas, que provocará mayor estabilidad económica.	Disminuirán las ventas o se mantendrán sin obtener un ingreso superior a lo establecido.
Sobre el área de sistemas	Habrá unificación de las bases de datos que ayudará no tan solo al proceso de MD, sino a la entrega rápida de la información para cualquier área de negocio.	Se mantendrán los sistemas descentralizados, enfocadas a cada área de negocio.
Competencia	Se mantendrá una competencia equitativa con las empresas del mismo giro que manejen dicha tecnología.	La empresa será desplazada por la competencia que trabaje con minería de datos.

La tabla anterior presenta el impacto de la aplicación de un proceso de minería de datos enfocado a la obtención de clientes potenciales basándose en diferentes perspectivas de negocio, las cuales ayudarán a visualizar los beneficios a largo y corto plazo que la empresa obtendrá con la aplicación de un proceso de MD.

Una vez que se planteo el panorama de la MD y la toma de decisiones, se presentará a continuación un enfoque de mercado de la MD, ya que los esfuerzos de todo el proceso de MD va dirigido obviamente al área de mercadotecnia para este caso de estudio en particular, y se mencionará cuáles son los pasos a seguir una vez obtenidos los clientes potenciales.

Importancia de la minería de datos en el mercado

Los resultados obtenidos con anterioridad deben ser presentados al área de mercadotecnia para que estos tomen decisiones y creen estrategias dirigidas a la satisfacción del cliente, y de igual forma a la captación de nuevos clientes, ya que al final de todo estudio mercadotécnico "el objetivo fundamental es servir al consumidor final,"[19] al obtener información respecto a quiénes son y dónde están ubicados, sus ingresos y deseos, sus motivaciones y actitudes y sus habilidades de compra. Lo que se ha llevado a cabo en el anterior análisis fue la obtención de los clientes potenciales, lo que permitirá a la empresa, cuando esta opere en aquel sector presentado, modificar de manera adecuada las estrategias de mercado y publicidad enfocadas a este segmento, al tomar en cuenta las reacciones y motivaciones obtenidas.

A pesar de que hasta el momento se obtuvo una base de datos de los clientes potenciales, hace falta conocer las características de los clientes que permita segmentar mi población de acuerdo a los gustos de los clientes, ya que estos tienen comportamientos diferentes, por ejemplo se puede encontrar :

- Un grupo dirigido por la costumbre de consumidores fieles a una marca, que tienden a quedar satisfechos con el producto o marca comprados la última vez.
- Un grupo de clientes conscientes y sensibles a los reclamos racionales.
- Un grupo de consumidores conocedores del precio, que deciden por comparación económica.
- Un grupo de clientes impulsivos que compran según la apariencia física del producto y nos son muy sensibles al nombre de la marca.
- Un grupo de consumidores que reaccionan emocionalmente y responden a símbolos de un producto y son muy impresionables por las imágenes.
- Un grupo de nuevos clientes que todavía no han estabilizado las dimensiones psicológicas de su comportamientos.

Lo más importante dentro de concepto de MD es que la empresa tenga conocimiento de quiénes son sus clientes potenciales y que tipo de características tienen, para que en conjunto se pueda preparar la publicidad de acuerdo al comportamiento de la población consumidora, y satisfacer segmentos específicos de clientes.

El análisis de Cluster, detallado en el subtena 1.3, una de las técnicas de MD mencionadas con anterioridad nos ayuda a crear segmentos de acuerdo a las características similares de los clientes.
"Siempre que el mercado para un producto lo constituyan dos o más compradores, el mercado está en posibilidad de que se le divida en segmentos , es decir, se le divida en grupos significativos de clientes. "[20]

La finalidad de la segmentación afirma Salvador Mercado en su libro de Mercadotecnia, es determinar diferencias entre compradores, que pueden tener consecuencias en la elección entre ellos o en venderles y esto con el fin de que los ejecutivos de *"marketing"* tengan bases para una mejor toma de decisiones.

[19] Salvador Mercado H, "Mercadotecnia, Principios y aplicaciones para orientar la empresa hacia el mercado", Edit, Limusa, 1991, p 67
[20] Idem, p 75

Una de las ventajas de la obtención de los clientes potenciales y su segmentación son:

- El mercado se puede examinar más fácilmente.
- Se está en mejores posibilidades de hacer comparaciones con la competencia.
- Al conocer los segmentos, se aprecian mejor el mercado y sus características.
- Se pueden hacer ajustes a los productor (innovaciones, atractivos, mejoramiento de calidad, etc.)
- Se enfoca la publicidad al segmento de los clientes de interés.

El objetivo de cualquier empresa es generar mayores ingresos, y una de las técnicas más usadas hoy en día es la publicidad, la cual se define como "el conjunto de actividades que se ocupan de informar sobre la existencia y cualidades de bienes y servicios, de tal forma que estimule su adquisición",[21] el proceso de reunir información y contestar a las preguntas de negocio no es otra cosa que la preparación para la ejecución de las decisiones y en este caso en particular, que se encuentra orientado a las EDVD, la toma de decisiones va relacionada con las estrategias de mercadeo que se van a ejecutar, por ejemplo, con que clientes va a invertir para aumentar sus ventas, a qué segmento de consumidores va a enfocar su publicidad, etc.

La publicidad tiene un poder muy grande, pues bien planeada con un análisis previo de mercado basado en la MD, con material de punto de venta adecuado, se logra interesar a los consumidores a adquirir el producto.

Así, las personas de mercadotecnia deben tener conocimiento completo del consumidor, sus gustos, sus preferencias, sus posibilidades, acoplado a todo esto al conocimiento de todas las otras funciones básicas de la compañía y en posición de ordenar lo que debe hacerse, a todo el personal relacionado con el producto desde su diseño hasta su venta, para tomar mejores decisiones en la empresa.

[21] Idem, p 389

3.2 Alcances y Limitaciones del caso práctico

Según Ackoff, un límite es una cantidad que una variable no puede sobrepasar, por lo que en este subtema se mencionarán los límites del caso práctico propuesto y los alcances del mismo.

Alcances del caso práctico

Los alcances del caso práctico van muy ligados con el plan de actividades presentadas en el capítulo II, en el cual se enumeraron las diferentes tareas que se llevaron a cabo para el proceso de MD. Se alcanzo el objetivo deseado que es, obtener los clientes potenciales de una EDVD, lo que ayudará a las personas del área de negocio a la toma de decisiones para enfocar sus esfuerzos a los clientes más rentables lo que en un tiempo determinado se convertirá en mayores ingresos para dicha empresa.
Haré mención dentro de los alcances del proyecto de uno de los mandamientos de Peter Druker que dice " Haga del aprendizaje organizacional una religión de la empresa. Al final, la única ventaja competitiva sostenida será su capacidad para aprender más rápido y mejor que sus competidores, y convertir dicho aprendizaje en nuevos productos, servicios y tecnologías antes de que su competencia imite su innovación más reciente." Debemos hacer uso de la tecnología para obtener resultados más rápidos y confiables, en este caso una de las mejores herramientas de MD según la revista estadounidense DM Review es SAS/Enterprise Miner utilizada en este proyecto, el alcance mencionado es solamente una parte de lo que se puede realizar con la herramienta, la innovación en las herramientas y aplicación de nuevos métodos de obtención de resultados debe ser un principio en todas las empresas, por lo que siempre se debe promover innovaciones para ser una empresa competitiva en el mercado.

Limitaciones del caso práctico

Una de las principales limitaciones del caso práctico fue la aplicación de diferentes técnicas de MD; ya que no se contaba con la experiencia suficiente para aplicarlas, por lo que se debe considerar que los resultados obtenidos fueron solamente generados al tomar en cuenta solamente regresiones logísticas y árboles de decisión, sin embargo faltaron por aplicar redes neuronales y algoritmos genéticos, los cuales pudieron haber generado mejores resultados a los presentados en este capítulo.

En este caso práctico solamente se obtuvieron los clientes potenciales al analizar 6 meses de historia, este proceso debe ser iterativo cada vez que haya nueva información que analizar; por lo regular se lleva a cabo cada mes, la minería de datos como proceso de planeación requiere de actualización constante, y es una actividad que se encuentra en constante cambio, me permitiré mencionar una frase de Ackoff que describe el proceso de MD manejándolo como proceso de planeación:

"Debido a que los sistemas con un propósito y sus medios ambientes cambian continuamente, no existen planes que conserven su valor con el tiempo. Por tanto, los planes se deben actualizar, extender y corregir frecuentemente si no es que continuamente"[22]

Un proyecto puesto en producción necesita el apoyo de todas las áreas de negocio para llevar a cabo la concentración de la información, tenerla accesible, además requiere de un cambio continuo dependiendo de las preguntas de negocio que se quieran responder.

[22] Ackoff Rusell , "El arte de resolver problemas", Edit. Limusa, pp.237-239

Los análisis que se podrían llevar a cabo para obtener un mayor conocimiento del cliente son los siguientes:

- Segmentación de clientes.- en este caso se podría llevar a cabo la segmentación de clientes, para saber que características específicas que tienen aquellos clientes potenciales seleccionados, al utilizar técnicas de Agrupamiento o *Clustering* de la MD.

- Obtención de relaciones de compra de productos.- este análisis necesita de la lista de productos que adquiere el cliente para analizar si existe alguna asociación, entre los productos que se adquieren y de esta forma poder ofrecer ofertas de productos en conjunto, se pueden utilizar técnicas de MD de Asociación de variables para este fin.

La principal limitación que puede presentarse en un proyecto de MD, es la falta de continuidad, tanto de entrega de información y de actualización de la misma, por lo tanto es necesario que se lleve a cabo un tipo de planeación continua que incluya los principios holísticos mencionados por Ackoff que son:

El principio de la coordinación que establece que ninguna parte de una organización puede planearse con eficiencia si se plantea independientemente de las demás unidades del mismo nivel. Así, todas ellas deben planearse simultánea e interdependientemente.

El principio de la integración establece que la planeación realizada independientemente en cualquier nivel de un sistema no puede ser tan eficiente como la planeación llevada a cabo interdependientemente en todos los niveles.

"Cuando los principios de coordinación e integración se combinan, obtenemos el principio holístico, el cual enuncia que, mientras más partes y niveles de un sistema se planeen simultánea e interdependientemente, mejores serán los resultados."[23] Por lo que mientras haya una integración vertical y horizontal en la empresa, se obtendrán mejores resultados.

3.3 Tendencias de la minería de datos

Avance tecnológico

"El cambio constante hoy en día es sinónimo de competencia, el reto de incertidumbre; se le considera más bien una evolución en lugar de una revolución." [24] Por lo que cuando las organizaciones compiten para llegar a ser más eficientes, deben adaptarse al cambio que el ambiente de mercado les presente, ya sea en reducción de personal, actualización de herramientas de software, cambio en las políticas o procedimientos de la empresa, etc.

Para el presente trabajo es inminente el cambio tecnológico que lleva consigo, ya que si bien es cierto que la minería de datos no es un tema actual en el mundo internacional, para la empresa mexicana si lo es, y no por falta de personal capacitado, si no por falta de organización de las diferentes áreas de negocio involucradas en un proyecto de MD; el cambio tecnológico en las herramientas de MD va surgiendo año con año, es real que hace 3 años, menciona la empresa SAS Institute, no se contaba con una herramienta que manejara algoritmos genéticos en el proceso de MD , en el año del 2001 se creo y salió al mercado; uno de los presentes problemas actuales de la minería de datos y de muchas herramientas tecnológicas, es que

[23] Russell L. Ackoff, "Planificación de la empresa del futuro", edit. Limusa
[24] M:A: Magali Chaín Palavicini, "El manejo del cambio estrtégico en las organizaciones mexicanas".

muchas veces se cuenta con el software más actualizado, pero el conocimiento sobre las nuevas innovaciones que surgen se adquiere un poco tarde, o solamente lo tienen personas selectas.
A pesar de que las herramientas de MD en el mercado ocupan un lugar reducido, al mismo tiempo las aplicaciones de MD para el mercado han crecido exponencialmente. A continuación se presentarán algunas de las tendencias de MD que existen en el mercado hoy en día:

Minería de Texto

Una de las nuevas aplicaciones hoy en día que ha surgido en el mercado es la Minería de Texto o "Text Mining" , lo que se conoce como un subgrupo de la tecnología de minado de la información, que es un componente del concepto de Administración del conocimiento, en este caso se refiere a la colección de habilidades, experiencias y la sabiduría de la organización.

A diferencia de la minería de datos que se ha descrito en el presente trabajo, la Minería de Texto trabaja con información almacenada en documentos de texto. Específicamente, la minería de texto se refiere al proceso de búsqueda a través de datos en texto, o en Internet y sus derivados, de información que tenga significado para la organización.
La Minería de Texto es particularmente relevante porque existen enormes cantidades de conocimiento que reside en documentos de texto que están almacenados y que las organizaciones no toman ventaja de esto. El advenimiento de Internet y la publicación en línea se han incrementado notablemente por lo que se requiere de su revisión y análisis sencillo. La Minería de Texto dirige este problema, definiendo técnicas y herramientas diseñadas para analizar y entender este tipo de información dinámica.

Minería en Internet

Internet es un servicio que contiene información global de servicios, noticias, anuncios, información de clientes, administración financiera, educación, gobierno, comercio electrónico, y mucho más servicios de información. Internet también contiene una colección rica y dinámica de ligas de información a otras páginas.
"Minería en Internet es una tarea desafiante, que busca patrones de acceso a las páginas de Internet, estructura de estas, y contenido regular o dinámico."[25] En general la minería en Internet puede ser clasificada en tres categorías:

- El minado del contenido en Internet
- Minado de la estructura de las páginas en Internet
- El minado del uso de Internet

Alternativamente, las estructuras de las páginas en Internet, pueden ser tratados como parte de la minería de contenido, por lo que se puede simplificar el análisis.

Las dos tendencias de la minería de datos presentadas anteriormente, muestran que si en México la minería de datos es un concepto que esta madurando poco a poco, nos queda mucho camino por recorrer, por lo que se debe de estar siempre a la vanguardia de las nuevas tecnologías y recordar que la única constante dentro de cualquier organización, proyecto, sistema , herramienta, etc, es el cambio.

El cambio tecnológico debe ir de la mano del conocimiento, solamente así, el factor humano podrá hacer un uso adecuado de la herramienta que tiene en su poder.

[25] Jiawei Han, Micheline Kamber, "Data mining, Concepts and Techniques", Edit. Morgan Kaufmann Publishers, 2001, p435.

Presentación de la información

Al analizar el avance tecnológico en cuanto a la entrega de información, se observa que la comunicación entre los diferentes sistemas operacionales se realiza actualmente vía Internet, desde la compra de un arreglo floral, hasta la transacción internacional entre cuentas bancarias, una de las tendencias de la minería de datos, es la presentación de todos los resultados vía web de forma sencilla, en la cual se cree una interfaz de trabajo que permita la presentación de resultados vía web de forma agradable al usuario.

El poder del conocimiento

Un punto importante que se debe añadir a las tendencias de la minería de datos, es la vasta información que se puede obtener sobre el cliente, el cual requiere de bienes o servicios y que en muchos casos la información es comercializable, por lo que en un futuro no existirá privacidad en cuanto a la adquisición de productos necesarios, modos de vestir, carro de preferencia, deportes favoritos, pasatiempos y enfermedades, el panorama cambiará drásticamente de una fuerte demanda por parte del cliente, a una gran oferta por parte del proveedor, se contará con una base de datos de conocimiento enorme que la minería de datos podrá aprovechar; para este momento también se pretende que cambie la forma de trabajo, para la MD lo indispensable es obtener una consolidación de la información, la cual es muy difícil de conseguir en una empresa, ya que se debe de obtener el consentimiento de las diferentes áreas de negocio, en un futuro se tendrán los datos y se demandará la explotación de los mismos.

Cuando se usa la MD para obtener conocimiento de los clientes, se esta utilizando la más alta tecnología, la MD trata de predecir cuáles son la acciones que se tomarán en un futuro cercano, además se aprende qué es lo que la persona hizo en el pasado para predecir sus necesidades en el futuro.
"Una promesa de la MD es regresar a los negocios el enfoque de atención al cliente y proveer de procesos eficientes de negocio."[26]

3.3.1 Planeación de escenarios

En el tema anterior se menciono las tendencias en un futuro de la minería de datos, en el presente subtema se continuará con este tópico pero desde la perspectiva de la planeación de escenarios.

La planeación de escenarios es una herramienta de pensamiento para su uso en una conversación estratégica. No pretende predecir lo impredecible y considera futuros múltiples.

La planeación de escenarios contribuye a:

- Creación de una estructura de los eventos y patrones en el entorno
- Identificación de la incertidumbre irreducible
- Creación de un proceso de conversión dialéctica
- Aprovechamiento del conocimiento disponible de cada miembro de la organización
- Aportación de perspectivas externas
- Adaptación de los puntos anteriores a las consideraciones estratégicas corporativas.

[26] Michael J:A. Berry, Gordon Linoff, " Mastering Data Mining. The Art and Science of Customer Relationship Management, Edit. John Wiley & Sons, Inc, 2000, p.483.

La planeación de escenarios coadyuvara a la visión de proyectos de MD a futuro en donde se plantearan las variables que afectan a que un proyecto de MD se lleve a cabo, lo que permitirá tomar medidas preventivas de acuerdo a las tendencias futuristas que se describirán a continuación.

El análisis de escenarios (herramientas estratégicas) crean un pensamiento más eficaz. La planeación de escenarios es vital para la tarea ejecutiva normal cotidiana y es una manera de pensar que penetra en la mente institucional; está basada en las siguientes suposiciones de mero sentido común:

- Poseer estrategias sanas reduce la complejidad de la tarea ejecutiva
- Discutir la estrategia es parte natural de cualquier tarea ejecutiva
- No hay dificultad en una buena estrategia basada en el pensamiento de sentido común
- El invertir tiempo en estructurar el debate estratégico trae consigo un aumento de eficiencia en el aspecto de los aspectos cotidianos.

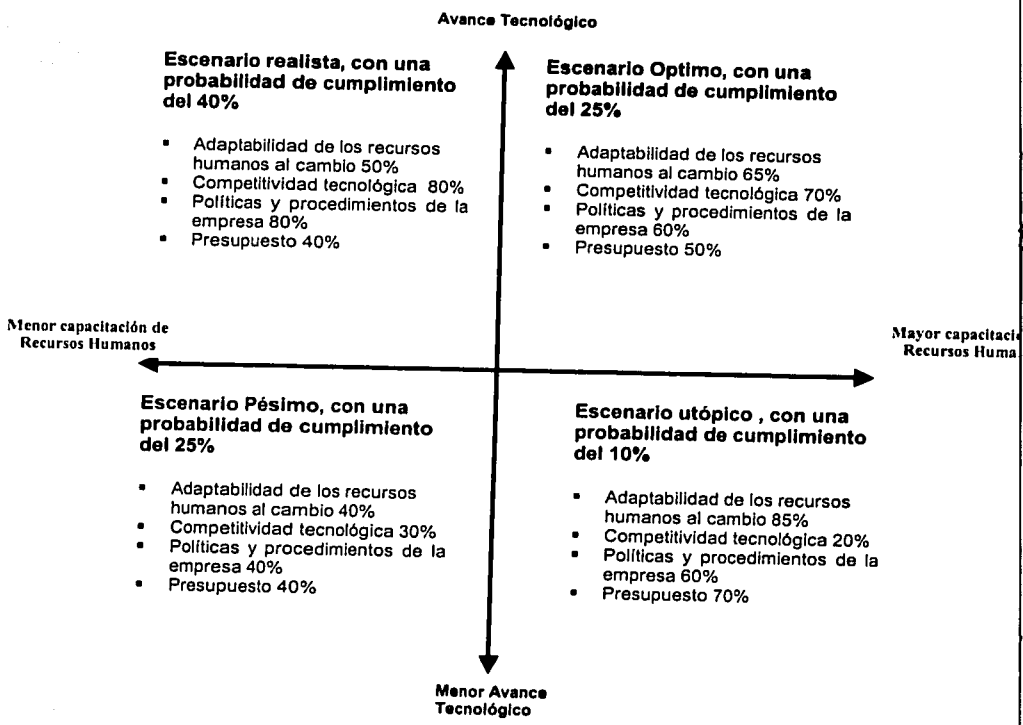
La primera fase para la planeación de escenarios es definir las variables primarias y secundarias que afectan de manera directa el proyecto y las cuales se definirán con base en el caso práctico que se expuso en este trabajo de investigación.

Primarias:
Avance Tecnológico
Capacitación de recursos humanos

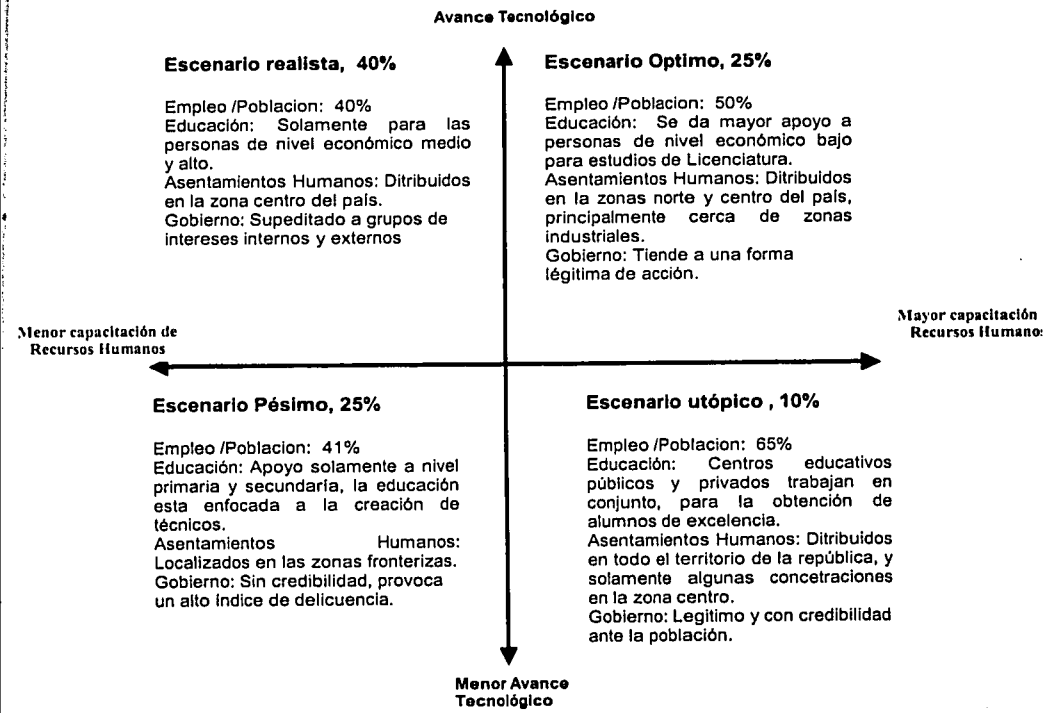
Secundarias:
Presupuesto
Políticas y procedimientos de la empresa
Competitividad de tecnologías
Adaptabilidad de los recursos humanos al cambio

Una vez que se definieron las variables de análisis se continuará con el desarrollo de los escenarios a corto (5 años) y largo (10 años) plazo.

Diagrama de escenarios a corto plazo
(año 2007)



Condiciones generales para que los escenarios anteriores se cumplan (año 2007)



ESCENARIO OPTIMO, con una probabilidad de cumplimiento del 20%

Adaptabilidad de los recursos humanos al cambio, 65%	La demanda de empleo y la competitividad con empresas extranjeras, exige que el personal asignado a proyectos de minería de datos(MD) se encuentre cada vez mayor capacitado en las herramientas de sistemas enfocadas a MD, al igual que en los nuevos métodos y técnicas usadas, ya que el concepto de MD esta entrando fuertemente en todas las medianas y grandes empresas de México. Los recursos humanos son los que se preocupan por su superación y crecimiento profesional por lo que se encuentran a un nivel competitivo en el mercado, ya que las empresas están dedicadas simplemente a genera mayores ingresos con base en resultados y aunque hay un apoyo al recurso humano todavía no lo tienen como el factor de mayor importancia en la empresa.
Competitividad tecnológica, 70%	La competitividad tecnológica esta creciendo ya no tanto en número de proveedores generadores de esta, ya que cada vez existen mayores fusiones de empresas de tecnología, por lo que quedan pocos proveedores pero mayores retos entre estos, ya que deben generar el mejor producto al mercado, tomando en cuenta la competencia no tan solo nacional si no internacional, y para la MD el mejor producto o herramienta de preferencia es aquel que cuenta con la mayoría de métodos y técnicas matemáticas y estadísticas de las cuáles muchas son innovaciones que van surgiendo a través de los años.
Políticas y procedimientos de la empresa, 60%	Dentro de los proyecto de MD se han ido tomando en cuenta la mayoría de las políticas y procedimientos de la empresas, sin embargo aún falta integrar parte de estos, ya que un proyecto de MD necesita el apoyo integral de toda la empresa, pero existen algunas áreas celosas de la información referente a su tipo operación por lo que los proyectos de MD se han realizado con la información que ha estado disponible y se han enfocado en las áreas que han compartido la información.
Presupuesto, 50%	Dado el nivel competitivo existente en las diferentes empresas, uno de los métodos que esta dando respuestas a los problemas surgidos en los negocios empresariales es la MD, por lo que las empresas han dedicado su presupuesto a la inversión en este tipo de proyectos.

ESCENARIO PÉSIMO, con una probabilidad de cumplimiento del 20%

Adaptabilidad de los recursos humanos al cambio, 40%	El precio de la tecnología ha ido aumentado considerablemente, por lo que solamente las grandes empresas pueden hacer uso de esta y contar con herramientas de MD, además se considera que las Pymes ocupan un porcentaje alto en la industria, aproximadamente el 90%, lo que provoca que la mayoría de los recursos humanos tengan miedo al cambio tecnológico, por la poca actualización en tecnología que se realiza, además del temor a ser reemplazados obliga a que se genere un ambiente contra la nueva tecnología.
Competitividad tecnológica, 30%	El aumento de precios en la tecnología, ha ido disminuyendo la competencia, ya que las empresas prefieren mantener tecnología obsoleta a invertir en nueva, lo que ha provocado que la brecha tecnológica haya crecido de manera impresionante, por lo que se ha producido la creación de tecnología personalizada par los negocios de mayores ingresos y deja atrás el poder de elección entre diferentes tecnologías del mercado, los proyectos de MD son dirigidos y personalizados para las grandes empresas.
Políticas y procedimientos de la empresa, 40%	Las políticas y procedimientos están enfocados a generar mayores ingresos y a reducir los costos, y se preocupan muy poco por el desarrollo de los recursos humanos. No se utilizan estrategias de mercadeo (que proporciona la MD) para la captación de nuevos clientes ni para la retención de los clientes de interés.
Presupuesto, 40%	El presupuesto de la mayoría de las empresas esta asignado solamente a proyectos y tecnología de bajo costo, por lo que el conocimiento de la nueva tecnología en estas empresas no ha ido evolucionado, lo que ha provocado un atraso de nivel técnico y profesional en las personas que laboran en empresas de bajo presupuesto, además de no utilizar la MD como herramienta para una mejor toma de decisiones.

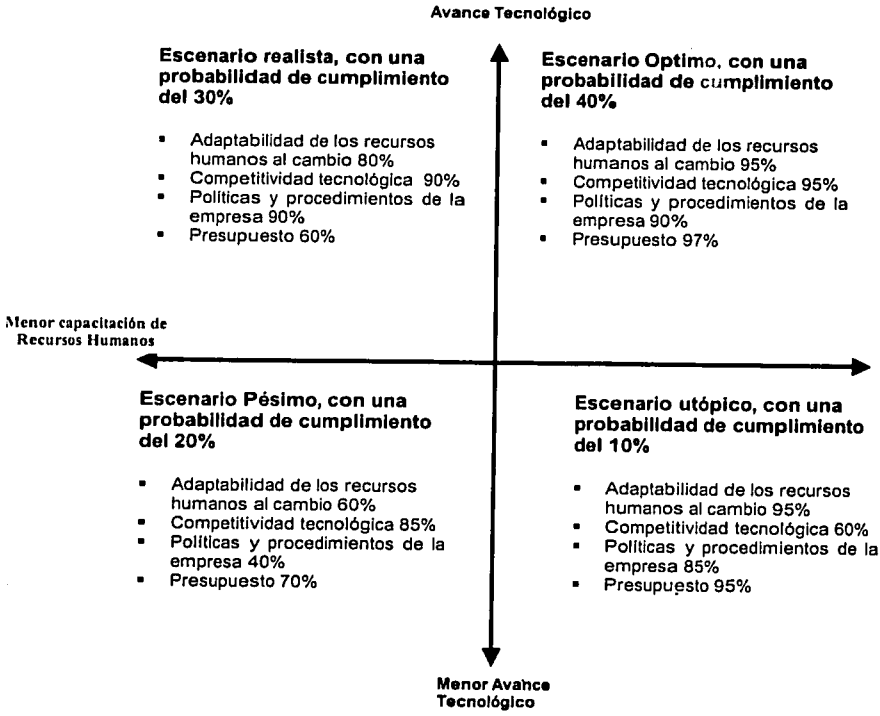
ESCENARIO REALISTA, con una probabilidad de cumplimiento del 40%

Adaptabilidad de los recursos humanos al cambio, 50%	El creciente avance tecnológico ha provocado un incremento de la competencia debido a la necesidad de comercialización de cada nueva tecnología que se ha ido creando año con año, lo que ha generado promociones para las empresas grandes y las Pymes, por lo que las empresas cuentan con la más alta tecnología, y dan por hecho que el personal se adecuara al cambio tecnológico, no tomando en cuenta la capacitación del personal, esto a largo plazo conducirá a una adaptabilidad lenta y a un conocimiento tecnológico limitado que afectará de manera directa el éxito de la empresa.
Competitividad tecnológica, 80%	La competencia tecnológica se ha incrementado por la utilización de tecnología en la mayoría de las empresas, por lo que la guerra comercial de publicidad en medios de comunicación se hace presente al llegar a diferentes niveles económicos de las empresas, haciendo uso de la tecnología satelital para tal fin.
Políticas y procedimientos de la empresa, 40%	Se encuentran enfocados al avance tecnológico y se han olvidado del recurso humano, lo que provocará a a futuro personal descontento y sin motivación.
Presupuesto, 40%	Las empresas no cuentan con presupuesto suficiente para hacer frente a los avances tecnológicos. La comercialización de la tecnología a impulsado su adquisición, y la necesidad de contar con la última tecnología a ocacionado la compra de tecnología poco robusta y provoca que no se exploten todas las técnicas de MD, por lo que no se han obtenido resultados satisfactorios en los proyectos realizados.

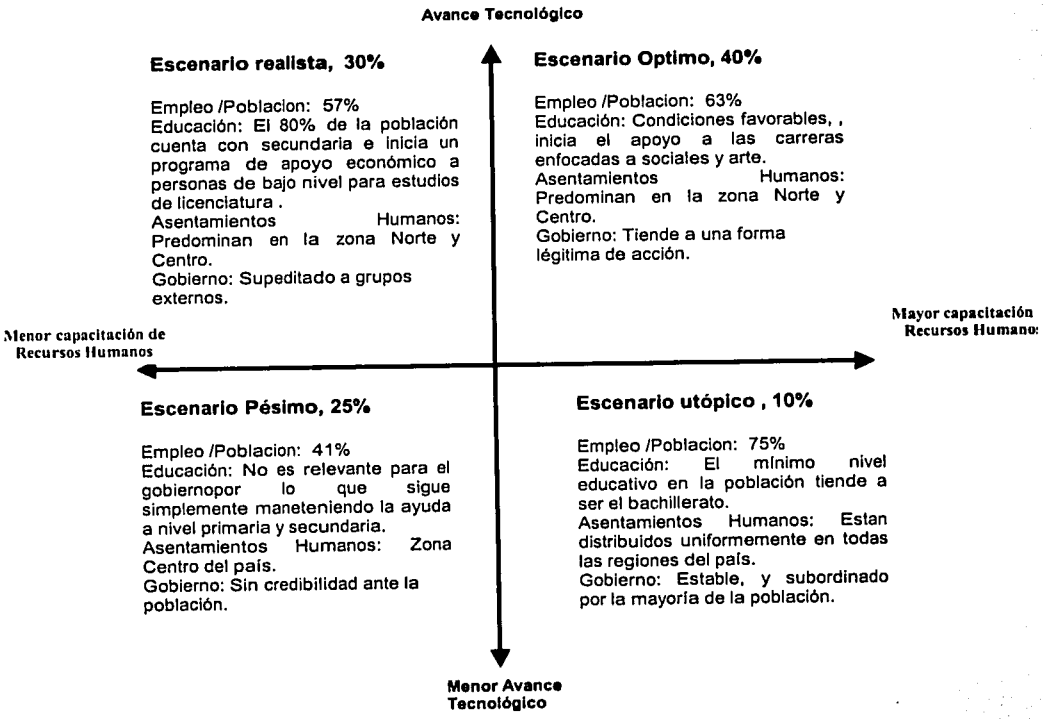
ESCENARIO UTÓPICO , con una probabilidad de cumplimiento del 20%

Adaptabilidad de los recursos humanos al cambio, 85%	Las empresas están enfocando sus esfuerzos a mejorar su productividad y capacitar a su personal constantemente en las herramientas necesarias para desempeñar mejor su trabajo, además de preocuparse por la formación profesional. Por lo tanto existe personal capacitado que puede llevar a cabo los proyectos de MD a la perfección, y esto se ha logrado con ayuda también de la disminución de nuevo tecnología ya que el personal ha aprendido a la perfección el manejo de esta.
Competitividad tecnológica, 20%	La competencia tecnológica ha disminuido gradualmente ya que la fusión de las empresas se ha incrementado hasta llegar a un punto en que existe muy poca competencia en el mercado, lo que ha provocado un decremento en el avance tecnológico por tener un exceso de seguridad en que la compra del producto siempre se va a llevar a cabo, por lo que se decide realizar lentamente la actualización de las diferentes tecnologías, las herramientas de MD en este momento tienen el mismo grado de robustez, por lo que la diferencia solamente se encuentra en el proveedor que la distribuye.
Políticas y procedimientos de la empresa, 60%	Las políticas y procedimientos se han actualizado constantemente de acuerdo a las propuestas de los empleados y a las necesidades que estos tienen, haciéndolas cada vez más flexibles a las necesidades de la empresa, además que se tienen el apoyo de los directivos para este fin.
Presupuesto, 70%	El presupuesto es asignado a cada uno de los proyectos prioritarios de la empresa, entre los que se encuentran los de MD, además, se tiene como prioridad la motivación y la creación de líderes empresariales, por lo que la tarea es asignar presupuesto a los proyectos que a corto y largo plazo produzcan ganancias a las empresas, y uno de los medios para llevar a cabo este fin, es tener al personal motivado y capacitado eficientemente .

Diagrama de escenarios a largo plazo (año 2012)



Condiciones generales para que los escenarios anteriores se cumplan (año 2012)



A continuación se presentará una visión más amplia de los escenarios planteados con anterioridad para el año 2012.

ESCENARIO OPTIMO, con una probabilidad de cumplimiento del 40%

Adaptabilidad de los recursos humanos al cambio, 95%	El factor humano es considerado el sujeto más importante de la empresa, por lo que tiene capacitación constante sobre las herramientas tecnológicas de mayor comercialización de MD, los proyectos de MD serán vistos un proceso necesario en la empresa al igual que la herramienta. El avance tecnológico ha continuado su crecimiento de manera exponencial, no tan solo en herramientas de MD; si no en sistemas operativos, operaciones transaccionales vía Internet, lectura de correo electrónico a través de diferentes medios, etc., por lo que las personas están acostumbradas al cambio de tecnología, lo que permitirá la adaptación rápida y a un menor costo a las diferentes herramientas de MD que se les presenten.
Competitividad tecnológica, 95%	Ya que la adaptabilidad del individuo, va de la mano con el avance tecnológico, la competencia va creciendo, sin embargo, todo dependerá de quien lleve a la perfección una estrategia de planeación y un método de venta al usar toda la tecnología móvil que se encuentre en el momento para llegar en primer lugar con el cliente, y así permita superar a los adversarios, por lo que la competitividad tendrá un nivel de importancia del 95%.
Políticas y procedimientos de la empresa, 90%	Las políticas y procedimientos de las empresas, han sido trascendentes para lograr la unión y organización del factor humano y de las diferentes áreas de la empresa, los usos de procedimientos para llevar a cabo un proyecto de MD se vuelven imprescindibles, por lo que los errores van disminuyendo y el aprendizaje va incrementándose.
Presupuesto, 97%	Las alianzas estratégicas de las empresas, generadas a través de los años, han provocado una leve disminución en el gasto de presupuesto, ya que los aliados colaboran para el desarrollo de proyectos en conjunto lo que provoca la disminución de costos, por lo que la mayoría de las empresas cuentan con un área de MD y con una herramienta actualizada para sus proyectos internos.

ESCENARIO PÉSIMO, con una probabilidad de cumplimiento del 20%

Adaptabilidad de los recursos humanos al cambio, 60%	A pesar que el crecimiento tecnológico ha sido difundido a través de diferentes medios de comunicación, aún existen muchas empresas que por falta de presupuesto, no quieren implementar nuevas tecnologías, ni capacitar al personal en estas adquisiciones, la mayoría de los recursos humanos están acostumbrados a todo el cambio tecnológico generado no tan solo en la empresa, si no al cambio tecnológico que existe en los medios de transporte y comunicación, en el hogar, etc., por lo que la adaptabilidad al cambio de alguna forma esta bien definida y es parte ya de una rutina, por lo que la empresa no se preocupa por capacitar a los recursos humanos, ya que estos tiene la costumbre de enfrentarse a cambios drásticos, sin embargo no se preocupa por los cambios conceptuales, manteniendo al factor humano como especialistas técnicos, a los que les hace falta, capacitación en cuanto a liderazgo, planteamiento de estrategias, metodologías, métodos de planeación, etc.
Competitividad tecnológica, 85%	La competencia tecnológica es parte del la rutina comercial de todo el ambiente empresarial, por lo que en este momento existe una competencia impresionante ya que han surgido nuevas empresas competidoras y ya que el cambio tecnológico se genera cada día y la población esta acostumbrada a este, las empresas se han preocupado por generar mayores estrategias de comercialización utilizando la MD como método esencial sin embargo estos métodos están siendo usados para un bombardeo comercial excesivo.
Políticas y procedimientos de la empresa, 40%	Ya que el crecimiento de la tecnología ha aumentado considerablemente, las políticas y procedimientos están siendo enfocados al desarrollo de empleados técnicos y maquileros, los cuales no tengan acción en la empresa, y solamente los directivos tengan conocimiento de la forma de operación de esta, por lo que los proyectos de MD enfocados a una mejor toma de decisiones están siendo utilizados por un personal restringido.
Presupuesto, 70%	El presupuesto esta siendo utilizado para la adquisición de nueva tecnología enfocada al beneficio propio de los dueños de las empresas, y no hay una preocupación directa por los recursos humanos, y ya que la demanda de trabajo se ha incrementado, los dueños de las empresas despiden y contratan a recursos humanos en tiempos muy cortos según lo requieran.

ESCENARIO REALISTA, con una probabilidad de cumplimiento del 30%

Adaptabilidad de los recursos humanos al cambio, 80%	Los recursos humanos son un factor importante en las medianas y grandes empresas, por lo que la capacitación en las nuevas tecnologías ha sido imprescindible para el crecimiento de estas y aunque no existe una preocupación directa por el factor humano, la capacitación se ha realizado más que por gusto, por necesidad, por lo que también existe un área de las personas encargadas de los proyectos de MD. Solamente las grandes empresas han generado un ambiente de trabajo agradable, motivando al personal, realizando actividades grupales, capacitándolos en cursos de desarrollo personal más que de profesional, por lo que existe diferencia entre los empleados de las grandes empresas y los de las medianas y pequeñas. La demanda del personal difiere enormemente cuando se trata de las Pymes y cuando se trata de las grandes empresas, ya que los requisitos ya no tan solo son a nivel técnico, sino cultural y de liderazgo.
Competitividad tecnológica, 90%	La tecnológica es indudablemente un factor trascendental, ya que es la que guía al éxito o fracaso de los proyectos emprendidos, las empresas están acostumbradas a que las respuestas a sus problemas estén resueltos por un medio tecnológico en un tiempo de respuesta mínimo, por lo que las empresas son demandantes de la tecnología. La existencia de múltiples proveedores en su mayoría extranjeros, es notable y no tan solo la presencia de productos norteamericanos, si no europeos y asiáticos han incrementado su presencia.
Políticas y procedimientos de la empresa, 90%	Tienen su enfoque en el seguimiento de un método para cualquier proyecto o acción a seguir dentro de la empresa, y es un requisito necesario el que todo las operaciones de la empresa estén sustentadas en una planeación previa y se utilice un método en específico en todo proyecto realizado, por lo que al realizar uno de MD, estos deben estar sustentados en un método de planeación.
Presupuesto, 60%	El presupuesto es asignado de manera muy cuidadosa a cada uno de los proyecto emprendidos, y aunque se asigna presupuesto a la capacitación del personal tanto técnica como profesional esto no es aún una prioridad en todas las empresas.

ESCENARIO UTÓPICO , con una probabilidad de cumplimiento del 10%

Adaptabilidad de los recursos humanos al cambio, 95%	Las empresas han enfocado sus esfuerzos a la capacitación de sus recursos en cuanto a los avances tecnológicos, conocimiento de las nuevas tendencias, además de formarlos con cursos de desarrollo humano, que permitirá el desarrollo de un personal altamente capaz de adecuarse a cualquier tecnología, forma de trabajo o estructura organizacional que se le presente. Por lo que la calidad de vida de las personas se ha incrementado, y ha originado la motivación y el deseo de realizar un buen trabajo.
Competitividad tecnológica, 60%	La competitividad tecnológica ha disminuido en gran medida ya que el personal con que se cuenta tiene la visión suficiente para decidir que talvez los cambios tecnológicos tan frecuentes pueden evadirse, si es que se toma la decisión correcta de un cambio tecnológico y se escoge la tecnología adecuada que tenga una funcionalidad a largo plazo. Además la empresa no se deja llevar por lo comercial que pueda ser la nueva tecnología, si no lo funcional que pueda a portar al negocio.
Políticas y procedimientos de la empresa, 85%	Enfocados al desarrollo humano, profesional y cultural de los recursos de la empresa. En este escenario no solamente se llevan a cabo procedimientos y políticas orientadas a los procesos de las empresas, si no que se crean políticas para la superación del individuo, procedimientos en cuanto a trato de personal que tenga como fin el crear un ambiente agradable y de respeto en el área de trabajo.
Presupuesto, 95%	Disponible para cualquier inversión ya sea tecnológica o de desarrollo humano. Para este momento los dirigentes de las empresas han sido educados de forma profesional y humanamente por lo que los recursos humanos son llevados a un nivel de eficiencia y liderazgo en la empresa.

"La creación de escenarios presupone escribir la historia anticipadamente como la haría un historiador actual con algo que ya paso pero en este caso apenas tiende a que podría suceder o no y esto hace que la empresa adquiera conciencia del rumbo en el que va, para adaptarse mas rápidamente a los cambios y responder a tiempo real a las circunstancias en las que se encuentre debido a las exploraciones previas del terreno. "[27]

De esta forma la planeación de escenarios presentada con anterioridad dará una visión general de que acontecería a un corto y largo plazo, sin embargo el mundo esta en constante cambio que podría darse el caso que alguno de los escenarios se cumpliera o ninguno de ellos se llegará a pareces en lo más mínimo que acontecerá con la MD a futuro.

[27] Kees Van Der Heijden, Escenarios "El arte de prevenir el futuro" , Capitulo 6, Edit. Panorama, 1998.

Conclusiones

A lo largo de los años la idea del dominio político, social y económico ha ido evolucionado, si bien es cierto el poder hace algunos años estaba basado solamente en el concepto de la riqueza individual o grupal, ahora cambia la perspectiva sobre la persona de mayor poder, dándole importancia a la persona poseedora del mayor conocimiento posible. De ahí, que las empresas dediquen sus esfuerzos a obtener conocimiento sobre sus clientes, muy independientemente del giro al que pertenezcan, a través de la MD.

En la sección 3.1 se presentaron los resultados obtenidos al emplear el método de MD, por lo que se observaron las características que influyen para que el cliente sea potencial, el estudio de MD en este trabajo concluye al presentar una tabla en la cual se analizan los clientes a través de la probabilid que tienen de ser potenciales, así los clientes que tengan una probabilidad alta de ser potenciales, serán los que la empresa considere para dirigir estrategías de mercado y publicidad. Dicha probabilidad fue obtenida al aplicar un método estadístico que fue el árbol de decisión, mencionado en el apartado 1.3.1 y aplicado en la sección 2.2.2.4, así mismo, la tabla resultante y el árbol de decisión estan presentados en el apartado 3.1. El estudio de MD dentro de una empresa de este giro, no termina aquí ya que este es el principio del análisis de los clientes, por lo que se puede continuar con agrupaciones de clientes según sus características y realizar segmentación de mercado, también se puede continuar con el análisis de ventas cruzadas como se cito en la sección 3.3, ya que se puede analizar los productos comprados por los clientes y ver si existe alguna frecuencia en la compra de varios productos para poder producir una oferta cojunta, etc; la MD tiene diferentes aplicaciones, por lo que, lo presentado en este trabajo solamente es uno de los diferentes análisis que se puede realizar en las EDVD con el fin de generar mayores ingresos a estas o si fuera el caso, con el propósito de investigación.

Las herramientas de MD han evoulucionado mucho y se pretende que en unos años más éste concepto sea utilizado por la mayoría de las empresas mexicanas, haciendo uso de una herramienta que facilite el análisis; si no fuera este el caso de la mayoría de la empresas en México, se pensaría en un escenario pesimista para la empresa, ya que no tendría un nivel competitivo con las demás.

Conclusiones Generales

La minería de datos(MD) definida en la sección 1.3, ha llegado a unir las diferentes técnicas que ya existían con anterioridad como son: regresiones lineales, árboles de decisión y redes neuronales para un fin en común que puede ser, el resolver un problema de negocio a través de un análisis supervisado o no supervisado, descritos ambos en el subtema 1.3.

De acuerdo al objetivo planteado, la presente investigación cumple con el establecimiento de una guía, detallada en el capítulo II, que permite la obtención de clientes potenciales de las EDVD, al aplicar el método de MD, sin embargo no se logra obtener los beneficios reales para la empresa de este giro, presentada en el caso práctico, detallado en el capítulo II, ya que para este fin, el proyecto debería estar en producción y, además, tendría que evaluarse después de la aplicación de alguna estrategia de negocio dirigida a la población potencial para medir la respuesta obtenida por los clientes, sin embargo, se elaboraron supuestos, mencionados en la sección 3.1, sobre los beneficios obtenidos al utilizar una herramienta de MD.

A través, de los tres capítulos presentados se dio a conocer las diferentes técnicas de MD existentes, cabe señalar que no se ocuparon todas las técnicas en el caso práctico, sin embargo, la descripción de estas técnicas da la pauta para su posterior aplicación. Se describió el método de planeación propuesto por el sistema SAS el cual nos guía para la ejecución de un proyecto de MD desde la definición del problema de negocio que se quiere analizar hasta la implementación y revisión de los beneficios obtenidos.

En el transcurso del presente trabajo se presento un caso práctico, analizado en el capítulo II, enfocado a las empresas dedicadas a las ventas a detalle (EDVD), que, en conjunto con el proceso de planeación descrito con anterioridad guía al consultor para la aplicación del método de minería de datos(MD) propuesto por el sistema SAS para la obtención de los clientes potenciales, los resultados obtenidos fueron parte de una prueba de concepto realizada para una empresa del giro mencionado, por lo que no se trabajo con todos los datos que la empresa maneja, sino que, estos fueron seleccionados y propuestos por la empresa misma. Los resultados obtenidos son representativos si y solo si, las variables que se incluyeron en el trabajo, son las que la empresa maneja al igual que el volumen de información requerida, de otra forma se estaría excluyendo información y los resultados no serían confiables.

En el capítulo III se presentaron las reglas de decisión tomadas en cuenta para la determinación de los clientes potenciales, el resultado final del presente trabajo fue la obtención de la probabilidad de que el cliente fuera potencial para cada uno de los casos analizados, esta probabilidad fue generada al aplicar las técnicas de minería de datos como son: Regresiones logísticas y los árboles de decisión. Los resultados del presente trabajo no se deben generalizar para todas las EDVD ya que cada empresa cuenta con información almacenada que puede variar desde el tamaño de datos que esta manejando, la limpieza de la información almacenada, hasta el comportamiento que estos puedan tener, ya que todo esto influye en el proceso de MD que se desee aplicar.

La investigación presentada dio respuesta a la pregunta de quiénes son mis clientes potenciales, al asignar probabilidades de este hecho a cada uno de los clientes analizados, a partir de estos resultados el consultor de MD puede seguir con el análisis de la información, al aplicar técnicas de MD para encontrar las características que identifican a los clientes con mayor probabilidad de ser potenciales, cómo son los gustos y deseos del cliente, lo que quiere, y si fuera posible hasta cuánto quiere pagar por el producto ofrecido, etc; así, el conocimiento obtenido para la empresa ayudará a guiar sus estrategias de publicidad y mercadeo a los clientes objetivo.

Este trabajo debe considerarse una ayuda para el consultor de MD que desee iniciar un proyecto enfocado a las EDVD, y que requiera de una guía para lograrlo.

Hoy en día la mayoría de las empresas en México están creando una área de MD, sin embargo a veces no tienen los elementos necesarios para hacerlo, por ejemplo: una herramienta de MD, una persona con conocimiento de negocio, una persona que tenga conocimiento sobre los datos que se van a manejar(del área de sistemas) y un analista de la información que sería el consultor de MD, en lo antes mencionado, se observa que cada una de las personas mencionadas son de diferentes áreas de aplicación, esto es porque se incluyen a personas de sistemas, a un actuario o matemático con especialidad en MD para el rol de analista de la información y el analista de negocio que debe ser un especialista con experiencia en el área de aplicación, en la realización de un proyecto de MD debe haber una organización entre las diferentes áreas de la empresa, ya que un proyecto no tan solo se aplica una sola vez, si no que debe actualizarse constantemente de acuerdo a la nueva información almacenada día a día.

Así pues, el proyecto de MD requiere de un cambio continuo dependiendo de las preguntas de negocio que se quieran responder, además, también depende de los cambios internos de la empresa ya que anteriormente se menciono, que un proyecto esta ligado a diferentes áreas de negocio y dirigido para este caso en particular al área de mercadotecnia, por tanto, en la medida que las empresas cuenten con herramientas de MD, podrán asegurar su permanencia en el mercado y mantener su nivel competitivo.

Fuentes de consulta

18. John P. Van Gigh, "Teoría general de sistemas, edit. trillas, 1981
19. Salvador Mercado H, "Mercadotecnia, Principios y aplicaciones para orientar la empresa hacia el mercado", Edit, Limusa, 1991.
20. Idem, p 75
21. Idem, p 389
22. Russel L. Ackoff, "El arte de resolver problemas", edit. Limusa
23. Russell L. Ackoff, "Planificación de la empresa del futuro", edit. Limusa
24. M:A: Magali Chaín Palavicini, "El manejo del cambio estrtégico en las organizaciones mexicanas".
25. Jiawei Han, Micheline Kamber, "Data mining, Concepts and Techniques". Edit. Morgan Kaufmann Publishers, 2001.
26. Michael J:A. Berry, Gordon Linoff, " Mastering Data Mining. The Art and Science of Customer Relationship Management, Edit. John Wiley & Sons, Inc, 2000.
27. Kees Vand Der Heijden, "Escenarios, El arte de prevenir el futuro", edit. Panorama, 1998.

Otras fuentes de consulta

1. Olivia Parr Rud, "Data Mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management", Edit. John Wiley & Sons, 2000.
2. Página principal de SAS Institute, http://www.sas.com
3. M. en A. Ignacion Martín Lizárraga Gaudry, Documento " Tendencias y Escenarios", 2001

Glosario

Applet.- Un applet es una mini-aplicación, escrita en Java, que se ejecuta en un browser (Netscape Navigator, Microsoft Internet Explorer, ...) al cargar una página HTML.

Arbol de decisión.- Forma de representación gráfica, de una colección de reglas jerarquicas, que se divide grupos de valores.

Análisis supervisado: Es una colección de tecnicas que utiliza una variable dependiente para la predicción de valores. Todos los tipos de regresiones y técnicas de clasificación son de tipo supervisado.

Análisis no supervisado: Este termino se refiere a una colección de técnicas que agrupa los datos sin utilizar una variable dependiente. El análisis de grupos (Cluster) es un ejemplo.

Agrupamiento(clustering): El proceso de dividir un conjunto de datos en grupo mutuamente excluyentes cada uno de los cuales comparten características similares que los ayudan a identificarse y que formarán parte del proceso de clasificación.

Árbol de decisión.- Es una estructura de árbol que representa un conjunto de decisiones, las cuales generan reglas para la clasificación de un conjunto de datos.

Algoritmos Genéticos.- Los Algoritmos Genéticos son métodos adaptativos que pueden usarse para resolver problemas de búsqueda y optimización.

Análisis multidimensional.- Se basa en herramientas que ayudan al usuario a analizar la información en diferentes niveles o dimensiones, por ejemplo: dimensión geográfica (país, estado, ciudad, region, etc.), dimensión de tiempo (año, mes, cuarto)

Bit. Binary Digit.- Unidad mínima de información utilizable por un ordenador. Teniendo en cuanto que el funcionamiento es por medio del sistema binario, los únicos valores que puede contener un bit es el 0 y 1.
Byte.- Es la unidad mínima de información, y está compuesta por ocho bits.

CGI (Common Gateway Interface).- El Common Gateway Interface es un estándar que define la interacción entre aplicaciones externas y servidores web. Así, un programa CGI recibe una información del servidor web y le devuelve a éste una página HTML.

Data warehouse(almacén de datos) .- Es una colección de bases de datos integradas, orientadas a temas, diseñadas para apoyar la función de los sistemas para el soporte de decisiones, donde cada unidad de datos es relevante en algún tiempo en específico.(Bill Inmon, 1991)

Data mart.- Es un subconjunto de datos del data warehouse, que a poya las necesidades de una área de usuarios específica. Los datos en un data mart son altamente resumidos, aún más que en el data warehouse.

Datos.- Valores coleccionados , tipicamente organizados para el análisis. Simplemente, los datos son hechos o transacciones realizadas.

Datos categoricos.- es un conjunto de datos que se ajustan a categorias discretas (que es lo opuesto a lo continuo), Los datos categoricos, pueden ser datos no ordenados de tipo nominal como el sexo o ciudad o datos ordenados llamados ordinales, como temperatura alta, media y baja.

Datos faltantes (datos missing) .- Los datos pueden ser faltantes por que no se obtuvo el dato de la información pedida por lo que es desconocida o fue borrada.

EDVD (Empresas dedicadas a las ventas a detalle o retail).- Las EDVD incluyen todas las funciones involucradas en vender (o alquilar) bienes y servicios a usuarios finales, incluyendo hogares, individuos, y otros que compran bienes y servicios de consumo final

Exploración de datos.- Busca en los datos relaciones antes no detectadas, a través de herramientas de visualización de la información.

GUI(interfaz gráfica de usuario).- Se le llama GUI a toda aplicación que cuente con una interfaz gráfica que ayude al usuario a interactuar de manera sencilla con la aplicación.

Gigabyte. Giga o Gb. Es una unidad de medida de memorias. Equivale a 1.024 Mb.

Hardware: Todos los componentes físicos e intangibles de la computadora, se denomina hardware.

JSP (Java Server Pages) .- es una tecnología que nos permite mezclar HTML estático con HTML generado dinámicamente.

Kilobytes. Kb. Unidad de almacenamiento equivalente a 1.024 bytes.

Minería de datos.- es un proceso que tiene como objetivo descubrir patrones ocultos en la información, usando diferentes técnicas estadísticas y matemáticas.
Técnicas de minería de datos: Algoritmos diseñados para el análisis de datos.

Muestra.- Crea un subgrupo de datos, que representa al conjunto de información de análisis, a través, de diferentes mecanismos, como puede ser de forma aleatoria.

Redes Neuronales.- Modelo predictivo no lineal que aprende a través del entrenamiento y asemeja una estructura biológica.

Regresión Lineal.- Es una técnica estadística usada para encontrar las mejores relaciones lineales entre una variable(dependiente) y sus predicciones (independientes).

Regresiones Logísticas: Es una regresión lineal que predice las proporciones de una variable categórica en una población.

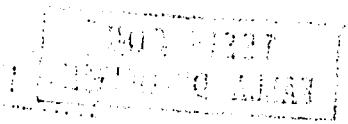
Servlet.- un servlet es una aplicación que se ejecuta en un servidor Web y queda a la espera para resolver peticiones efectuadas por los clientes. Mediante los servlets se puede tener acceso a otros servidores y acceder a la información que en ellos haya contenida, por ejemplo una base de datos.

Series de tiempo.- Es el análisis de una secuencia de elementos hecha en un cierto intervalo de tiempo.

Sistema SAS .- Es una herramienta computacional que se encarga de la entrega de información que ayuda a la toma de decisiones.

Software: Es todo lo intangible que tiene la computadora.

TeraByte. Unidad de medida de almacenamiento que equivale a 1.024 Gigabytes.



Variable objetivo.- es una variable que describe el problema de negocio al que se quiere dar solución, por lo regular una variable objetivo es binaria (0 y 1), en un ejemplo enfocado a Telefonía, el 1 podría significar cliente cancelado y el 0 cliente no cancelado. Esta variable objetivo se define en conjunto con las personas de negocio, ya que ello dictan las reglas para la asignación de valores.

Valores missing.- Los valores missing o valores faltantes son aquellos que representan ausencia de valore, no se incluyen los blancos, ni el 0 en esta definición.

Visualización de datos: es una técnica de minería de datos que se encarga de la representación gráfica de los datos, que nos ayuda a conocer información que no podríamos apreciar con simples números.