ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Instituto de Ciencias Matemáticas

Ingeniería en Estadística informática

"Obtención del perfil de un cliente fiel en una tienda departamental mediante el diseño de un Data Warehouse y Árboles de Decisión"

TESINA DE GRADO

Previa a la obtención del Título de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Presentado por:

DANIEL SANTIAGO AGUIRRE MOROCHO
JUAN CARLOS GALLO ANTE
VICTOR ROLANDO VALENCIA VALVERDE

GUAYAQUIL – ECUADOR
AÑO
2010

AGRADECIMIENTO

Agradecemos a Dios, a nuestros padres, hermanos y amigos que han hecho posible el desarrollo de esta tesina. A nuestro director por la paciencia y aporte de sus conocimientos en este trabajo.

DEDICATORIA

Dedicamos esta tesina a nuestros padres, hermanos y amigos por su apoyo incondicional.

TRIBUNAL DE GRADUACIÓN

Ing. Fabricio Echeverria	**Ing. Carlos Martín**
DIRECTOR DE TESIS	**DELEGADO**

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de esta tesina de grado, nos corresponde exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL"

(Reglamento de graduación de la ESPOL)

Daniel Santiago Aguirre Morocho

Juan Carlos Gallo Ante

Víctor Rolando Valencia Valverde

RESUMEN

Las exigencias diarias y las situaciones del medio involucran a las empresas y directores en una urgencia de encontrar soluciones inmediatas y tomar decisiones ante diversos escenarios posibles que se presentan usualmente. Una adecuado uso de la información y la correcta toma de decisiones en la actualidad es la mejor arma en un mundo competitivo.

El presente trabajo utiliza la técnica de minería de datos y los procesos para el diseño y construcción (extracción, transformación y carga de datos) de un data warehouse y establece un árbol de decisión como método central para determinar el perfil de un cliente fiel en una tienda departamental.

La metodología a seguir en el desarrollo de esta tesina será, primero describir la situación actual, problemática encontrada y la propuesta como alternativa de solución en la tienda departamental.

En el capitulo dos encontramos el modelo multidimensional, conceptos básicos y fundamentales, el modelo de minería de datos, el diseño del data mart y el diseño de un data warehouse que almacene la información extraída.

El capitulo tres se concentra en la metodología árboles de decisión, sus ventajas y desventajas. Comprende el diseño y la arquitectura del árbol, su modelamiento y análisis para la toma de decisiones y la determinación del perfil del cliente fiel.

Por último se presentan las conclusiones y recomendaciones basadas en los resultados de los análisis realizados en los capítulos anteriores.

INDICE GENERAL

CAPÍTULO 3

DISEÑO DEL ÁRBOL PARA LA TOMA DE DECISIONES

CAPÍTULO 4

CONCLUSIONES Y RECOMENDACIONES

INDICE DE TABLAS

INDICE DE FIGURAS

CAPÍTULO 1

TIENDAS DE RETAIL

1.1 Introducción

Sin duda, el diario vivir y las exigencia del medio involucra a las empresas en una necesidad imperiosa de encontrar una serie de respuestas inmediatas ante diversas situaciones o casos que regularmente se presentan. Sin embargo la mayor parte del tiempo la toma de decisiones en las organizaciones son responsabilidad de los altos mandos (alta gerencia) permitiendo poca participación de los mandos intermedios.

*"…siempre la última apelación y la última palabra deberán estar en las manos de un solo hombre. Este hombre posiblemente deberá turnarse con frecuencia con otros. No deberá hacer ningún trabajo especial, salvo aprobar o desaprobar algo ya estudiado por completo y condensado al punto que falte solamente escoger entre el sí o el no***"**[1].

La toma de decisiones en el ámbito laboral se ha convertido en una competencia riesgosa y constante que conlleva a la alta dirección en tratar de mitigar el riesgo lo mayormente posible para lo cual se necesita una cantidad de información sobre un tema específico, siendo necesario acceder a bases de datos.

El propósito es desarrollar y mantener los sistemas de datos haciéndolos disponibles y fáciles para los altos mandos de las empresas, en términos de Data warehouse (Almacén de datos), es brindar la solución consolidada para que todos puedan acceder a la información con los reportes necesarios, dando respuesta a necesidades de diferentes tipos de usuarios para la toma de decisiones.

---

[1] Matteucci Zatinni, 1949

A lo largo de este primer capítulo, comentaremos acerca de las tiendas departamentales, sus antecedentes y actual realidad. Revisaremos la problemática del caso particular en estudio y definiremos la propuesta a utilizar como herramientas para la toma de decisiones riesgo mínimo.

1.2 Antecedentes

Las cadenas de retail asumen una presencia cada vez más importante en el comercio. Definamos retail como la comercialización al por menor. Usualmente utilizado para referirse al rubro de supermercados y tiendas por departamentos[2].

El detal o retail es aplicado a un sector económico que engloba a las empresas especializadas en la comercialización masiva de productos o servicios uniformes a grandes cantidades de clientes. En países más desarrollados, su relevancia es significativa y cada vez más especializada.

---

[2] Diplomas Postitulos, Ingeniería Industrial, Universidad de Chile, 2010

En el negocio del retail se pueden incluir todas las tiendas o locales comerciales que habitualmente se encuentran en cualquier centro urbano con venta directa al público, sin embargo su uso se halla más bien ligado a las grandes cadenas de locales comerciales. El ejemplo más común del retail lo constituyen los supermercados; las tiendas por departamentos, casas de artículos para el hogar, farmacias, entre otras.

La complejidad del retail viene dada por la amplia variedad de artículos y tipos de artículos que ofrecen, así como el nivel de operaciones efectuado. Las operaciones de venta del retail generan una cantidad de datos que puede resultar abrumadora para aquellos ajenos al negocio. Por otra parte, se aprecia crecientemente que las empresas con mayor proyección a nivel latinoamericano son aquellas que han consolidado un tamaño y procesos de gestión comerciales y logísticos altamente eficientes.

Según el estudio “Share of Mind” realizado por la revista Marka en el 2005, desde el año 1.919 el Ecuador ha dado acogida a un segmento muy particular en el sector comercial de la Economía. El sector de tiendas departamentales ha dado cobertura a gran parte de las necesidades de los consumidores pasando a formar parte fundamental y actualmente un papel muy importante en los índices

de consumo, este sector de tiendas por departamento ha mostrado un gran crecimiento tanto en variedad como en número de opciones para el mercado.

Las tiendas departamentales o tiendas por departamentos son establecimientos de grandes dimensiones que ofertan una variedad de productos encaminados a cubrir una amplia gama de necesidades: alimentación, confección, hogar, decoración, etc. Por lo general, se sitúan en el centro de las ciudades y suelen tener varias plantas, dividiendo su superficie comercial en secciones. Se diferencian fundamentalmente del centro comercial, porque los grandes almacenes pertenecen a una única empresa y es una sola tienda de enorme tamaño y de los hipermercados porque la alimentación no es su mayor prioridad en la venta.

1.3 Planteamiento del Problema

En referencia de nuestro proyecto, muestra un caso particular para una prestigiosa empresa del medio, con más de 60 años en el mercado, sostiene su actividad comercial a través de 20 tiendas distribuidas en las ciudades de Guayaquil y Quito. La empresa para incentivar las ventas beneficia a sus

clientes semestralmente realizando campañas promocionales en sus diferentes líneas de productos mediante sus tiendas departamentales.

En la actualidad, la campaña más importante es la dirigida a sus clientes fieles (Cliente que desde la apertura de su cuenta consuma frecuentemente y no posea deuda significativa con la empresa), en un día específico (Día del Socio) por lo que económicamente representa en asistencia y ventas en un solo día de promoción. Esta campaña oferta un porcentaje de descuentos en toda la mercadería de Moda y Hogar, se efectúa dos veces en el año por lo general en los meses de abril y noviembre.

Cada año, al momento de planificar la campaña “Día del Socio”, existe una diferencia entre las áreas de Crédito y de Marketing, la que consiste en decidir qué cantidad de clientes deben ser invitados con el objetivo de mejorar la tasa de respuesta e incrementar el nivel de ventas en relación al año anterior. En la tabla 1.1 se muestra los resultados obtenidos en los tres últimos años, en el cual se observa que la simple acción de invitar más clientes no asegura una mejor tasa de respuesta sin embargo puede suponer una relación hacia un mayor nivel de ventas.

Tabla 1.1 Tasa de respuesta de clientes vs ventas

Año	Clientes Invitados	Tasa de Respuesta	Compras
2007	90.000	3,6%	$ 190.000,00
2008	35.000	6,8%	$ 155.000,00
2009	140.000	3,4%	$ 332.000,00

Fuente: Base de datos de la tienda departamental

Elaborado por: Aguirre, Gallo y Valencia

Sin duda, los distintos enfoques en las áreas antes mencionadas, se presentan claros y fuertes al momento de tomar una decisión, esto inconscientemente provoco cierto desvío de atención al objetivo principal de la campaña y empresa. Pero por otro lado, trae a la superficie la necesidad imperiosa de preparar una decisión importante acorde a los lineamientos de la empresa la cual permita mitigar el riesgo al tomar decisiones no acertadas que generen gastos innecesarios o exagerados por parte de la empresa.

1.4 Propuesta

Siendo para la gerencia un problema muy complejo de solucionar, ya que confronta a dos de sus departamentos claves en el negocio, utilizaremos herramientas para la toma de decisiones que minimicen el riesgo al momento de escoger una solución. El objetivo es la creación de un modelo empresarial para

la toma de decisión bajo incertidumbre, para el efecto mediante una correcta aplicación de la minería de datos y un acertado modelamiento de los datos construiremos un árbol de decisión el cual proporciona una representación gráfica del problema y muestra la naturaleza secuencial del proceso de toma de decisiones.

Una de las grandes ventajas de los árboles de decisión es que, en su forma más general, las opciones posibles a partir de una determinada condición son excluyentes. Esto permite analizar una situación y siguiendo el árbol de decisión apropiadamente, llegar a una sola acción o decisión a tomar.

Luego, a partir del árbol se generará un Modelo de Decisión, el cual nos permitirá determinar el mejor perfil de clientes que deben ser invitados a la promoción “Día del Socio”, para lo cual utilizaremos dos indicadores, la tasa de respuesta de los clientes y sus respectivas ventas de el día especifico para así poder satisfacer a las áreas involucradas al momento de planificar dicha campaña.

Para el estudio, utilizaremos la base de datos histórica de los clientes actualizados y organizados al mes de diciembre del años 2009, y la base de

datos de ventas obtenidas desde el mes de enero del año 2006 hasta el mes de diciembre 2009.

CAPÍTULO 2

EL MODELO MULTIDIMENSIONAL

2.1 Introducción

El presente estudio describe el modelo de minería de datos y los procesos para el diseño y construcción (extracción, transformación y carga de datos) de un Data Warehouse, enfocado a encontrar “El Cliente Fiel” para una campaña en una tienda departamental.

La arquitectura de un Data Warehouse se origina en la obtención de información de las distintas fuentes (Excel, Access, SQL, Oracle, etc.) seguido del filtrado y agrupación de datos para posteriormente efectuar la carga de la información actualizada. Dicha arquitectura se orienta a evitar problemas que muchas veces se originan por la pérdida, duplicidad, carga irrelevante o poca consistencia de la información.

En el siguiente capítulo encontramos los conceptos básicos y fundamentales, el modelo de minería de datos y el diseño del Data Mart, con lo cual definimos como objetivos principales; el diseño de una herramienta capaz de extraer información, la limpieza y depuración de elementos no significativos de dicha información, el enriquecimiento de la información con datos más relevantes y el diseño de un Data Warehouse que almacene la información extraída.

2.2 Definición de conceptos utilizados

2.2.1 Minería de Datos

La minería de datos (en inglés, *data mining*) se define como la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de datos[1].

[1] DAEDALUS - DATA, Decisions and Language, S.A., 1998

2.2.2 Data Warehouse

Un Almacén de Datos o Data Warehouse es una base de datos corporativos que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta[2].

2.2.3 Data Mart

Un Data Mart es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento. Un Data Mart puede ser alimentado desde los datos de un Data Warehouse, o integrar por si mismo un compendio de distintas fuentes de información[3].

[2] SINNEXUS Business Intelligence + Informática estratégica, 2007

[3] SINNEXUS Business Intelligence + Informática estratégica, 2007

2.2.4 Modelo Multidimensional

En un modelo de datos multidimensional los datos se organizan alrededor de los temas de la organización. La estructura de datos manejada en este modelo son matrices multidimensionales o hipercubos[4].

2.2.5 Modelo de Consulta

Estos modelos permiten dar una idea más clara de lo que un diseño multidimensional persigue (dimensiones, medidas e indicadores), que no es otra cosa que ofrecer una vista más entendible y familiar al analista del negocio[5].

2.2.6 Modelo Estrella

Esquema de la estrella es la arquitectura de almacén de datos más simple. En este diseño del almacén de datos la tabla de Variables (Hechos) esta rodeada por Dimensiones y juntos forman una estructura que permite implementar mecanismos básicos para poder utilizarla con una herramienta de consultas OLAP Online Analytical Processing[6].

---

[4] Tamayo, M. y Moreno, F., 2006

[5] Bustamante, J., Rodriguez J., Echeverría F., 2009

[6] ETL-Tools.Info, 2006

2.2.7 Esquema en Copos de Nieve

Esquema en copo de nieve (bola de nieve) es una variedad más compleja del esquema estrella. El afinamiento está orientado a facilitar mantenimiento de dimensiones. Lo que distingue a la arquitectura en copo de nieve de la esquema estrella, es que las tablas de dimensiones en este modelo representan relaciones normalizadas (3NF) y forman parte de un modelo relacional de base de datos[7].

2.2.8 Carga de Datos

Inserción sistemática de datos en el componente de almacenamiento físico[8].

2.2.9 Cubo

Este proceso consiste en obtener datos relevantes entre la gran cantidad de información contenida en el sistema. Se pueden agregar múltiples dimensiones para realizar los cruces que permitirán extraer, en forma rápida y eficiente, la información que se requiere[9].

---

[7] ETL-Tools.Info, 2006

[8] Data Warehouse. Velasco, 2004

[9] DYBOX Information Technology, 2000

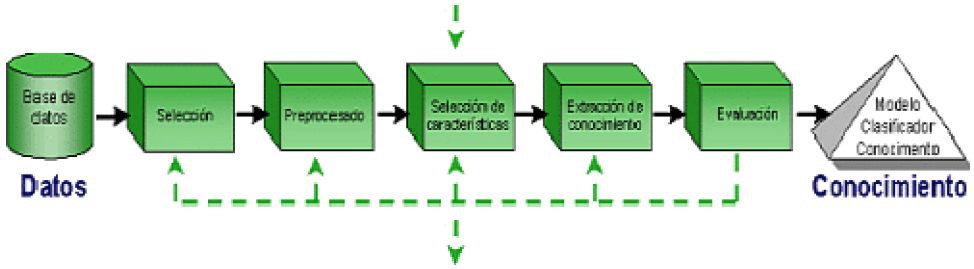
2.3 Modelo de Minera de Datos

La minería de datos se define también como el análisis y descubrimiento de conocimiento a partir de datos. En otras palabras, la minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos, hace uso de técnicas que pueden aportar información útil a través de métodos estadísticos complementados con métodos y algoritmos informáticos.

De acuerdo a DAEDALUS, los pasos a seguir para la realización de un proyecto de minería de datos son siempre los mismos, independientemente de la técnica específica de extracción de conocimiento usada.

El proceso de minería de datos se compone de las siguientes fases:

Figura 2.1: Fases dentro de un proceso de Minería de Datos



Fuente: DAEDALUS – DATA

- **Selección y preprocesado de datos**

  El formato de los datos contenidos en la fuente de datos (base de datos y Data Warehouse) nunca es el idóneo y la mayoría de las veces no es posible ni siquiera utilizar ningún algoritmo de minería sobre los datos "en bruto".

  Mediante el pre procesado se depuran los datos (de forma que se eliminan registros incompletos, o que no cumplan con las necesidades y el algoritmo que va a usarse), se obtienen muestras de los mismos (en busca de una mayor velocidad de respuesta del proceso), o se reduce el número de valores posibles (mediante redondeo o clustering).

- **Selección de variables**

  Aún después de haber sido pre procesados, en la mayoría de los casos se tiene una cantidad ingente de datos. La selección de características reduce el tamaño de los datos eligiendo las variables más influyentes en el problema, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería.

  Los métodos para la selección de características son básicamente dos:

1. Aquellos basados en la elección de los mejores atributos del problema
2. Y aquellos que buscan variables independientes mediante test de sensibilidad, algoritmos de distancia o heurísticos

- **Extracción de conocimiento**

  Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un pre procesado diferente de los datos.

- **Interpretación y evaluación**

  Una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno

de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

2.3.1. Descripción de las tablas

Dado que el objetivo es establecer la información más relevante que ayude a determinar el perfil de un cliente fiel. Es de gran utilidad en el presente estudio definir las tablas que nos facilitaran los datos a ser investigados, para así facilitarnos la interpretación de los resultados obtenidos. A continuación se presentara la descripción de cada tabla.

2.3.1.1. Maestro

La tabla Maestro contiene todos los clientes de la empresa, sus datos demográficos e información de la cuenta.

Tabla 2.1 Maestro

Maestro
id_cedula
id_cuenta
Nombre
Genero
estado civil
tipo de cliente
tipo de cuenta
estado de la cuenta
Ciudad
fecha apertura
fecha nacimiento
Dirección
Cupo

Fuente: Base de datos de la tienda departamental
Elaborado por: Aguirre, Gallo y Valencia

2.3.1.2. Estatus

La tabla Estatus, resume si la tarjeta ha sido retirada por el cliente o permanece como no entregada, Si el campo es igual a "E" o Blanco significa que la cuenta esta desbloqueada (activada).

Tabla 2.2 Status

Status
id_cliente_status
status_cuenta

Fuente: Base de datos de la tienda departamental

Elaborado por: Aguirre, Gallo y Valencia

2.3.1.3. Ventas

La tabla Ventas contiene todas las compras realizadas mensualmente por el cliente. Esta tabla nos facilita información de 48 meses (4 años), es decir desde Enero 2006 a Diciembre 2009.

Tabla 2.3 Ventas

Ventas
id_cliente_ventas
fecha apertura
mes_ene_06
mes_feb_06
:
:
:
mes_dic_09
meses_no_cliente
meses_ventas
meses_sin_ventas

Fuente: Base de datos de la tienda departamental

Elaborado por: Aguirre, Gallo y Valencia

2.3.1.4. Ciudad

La tabla Ciudad almacena la ciudad, provincia y región a la cual pertenece el cliente.

Tabla 2.4 Cuidad

Ciudad
id_ciudad
nombre_ciudad
provincia
region
tipo_region

Fuente: Base de datos de la tienda departamental

Elaborado por: Aguirre, Gallo y Valencia

2.3.1.5. Día de Corte

La tabla Día de Corte recopila información con respecto al estado actual de la cuenta.

Tabla 2.5 Día de corte

Día_de_corte
id_cliente_corte
lugar_apertura
edad_actual_mora
corte
saldo_neto

Fuente: Base de datos de la tienda departamental

Elaborado por: Aguirre, Gallo y Valencia

2.4 Diseño de Data Warehouse

La creación de un Data Warehouse representa en la mayoría de las ocasiones el primer paso, desde el punto de vista técnico, para implantar una solución completa y fiable de Business Intelligence (Inteligencia del Negocio).

Según (Miranda, 2007), La Inteligencia del Negocio (BI) representa las herramientas y sistemas que juegan un papel clave en el proceso estratégico de la planificación de una compañía. Estos sistemas permiten reunir, almacenar, y analizar los datos corporativos siendo una importante ayuda en la toma de decisiones.

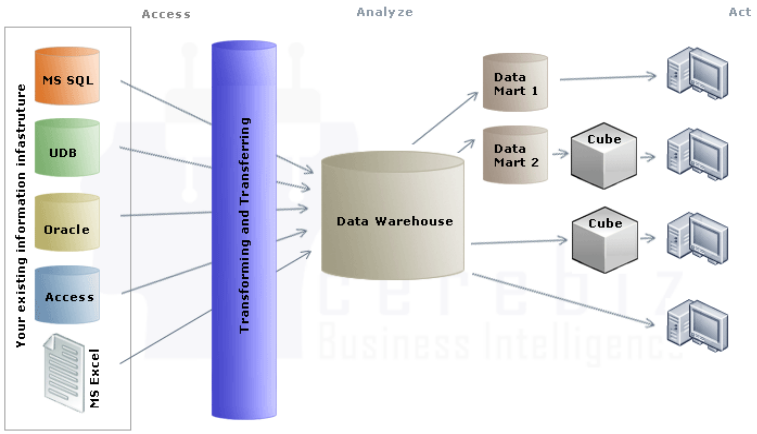
Para comprender íntegramente el concepto de Data Warehouse, es importante entender cuál es el proceso de construcción del mismo (estructura básica de la arquitectura Data Warehouse), denominado ETL (Extracción, Transformación y Carga), a partir de los sistemas operaciones de una compañía:

- **Extracción:** Selección sistemática u obtención de información de las distintas fuentes tanto internas como externas.

- **Transformación:** Proceso para filtrado, limpieza, depuración, homogeneización y agrupación de la información.
- **Carga:** organización y actualización de los datos y los metadatos en la base de datos. (Inserción de datos).

La ventaja principal de un Data Warehouse radica en las estructuras en las que se almacena la información (modelos de tablas en estrella, en copo de nieve, cubos relacionales, etc). Este tipo de persistencia de la información es homogénea y fiable, y permite la consulta y el tratamiento jerarquizado de la misma siempre en un entorno diferente a los sistemas operacionales[10].

Figura 2.2: Arquitectura de un Data Warehouse



Fuente: (SINNEXUS Business Intelligence + Informática estratégica, 2007).

[10] SINNEXUS Business Intelligence + Informática estratégica, 2007

Un Data Warehouse se caracteriza por ser:

- **Integrado:** los datos almacenados en el Data Warehouse deben integrarse en una estructura consistente, por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las distintas necesidades de los usuarios.

- **Temático:** sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales. Por ejemplo, todos los datos sobre clientes pueden ser consolidados en una única tabla del Data Warehouse. De esta forma, las peticiones de información sobre clientes serán más fáciles de responder dado que toda la información reside en el mismo lugar.

- **Histórico:** el tiempo es parte implícita de la información contenida en un Data Warehouse. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por

el contrario, la información almacenada en el Data Warehouse sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el Data Warehouse se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

- **No volátil:** el almacén de información de un Data Warehouse existe para ser leído, pero no modificado. La información es por tanto permanente, significando la actualización del Data Warehouse la incorporación de los últimos valores que tomaron las distintas variables contenidas en él sin ningún tipo de acción sobre lo que ya existía.

2.4.1 Modelo Multidimensional

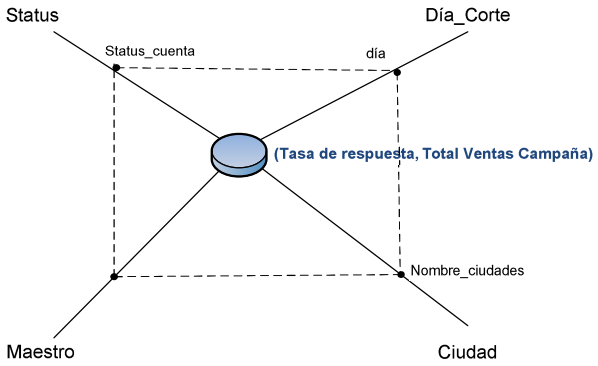
Para el diseño de un Data Warehouse, por lo general se utiliza un modelo multidimensional. Para poder entender la definición presentada así como el modelo multidimensional se deben comprender tres conceptos: cubo, medida y dimensión. Sus principales arquitecturas el Modelo Estrella y Copo de Nieve.

En un modelo multidimensional, se soporta el manejo de una enorme cantidad de datos empresariales y temporales. La estructura de datos

manipulada es de matrices multidimensionales o cubos. Sus componentes principales son las dimensiones y medidas. Las dimensiones están descritas por conjuntos de atributos, son tablas compuestas de niveles y jerarquías, a través de las cuales se agrupa los datos en un nivel deseado. La medida o hecho es un dato numérico que representa una actividad específica de un negocio.

La Figura 2.3, muestra el Modelo de Consulta, en el cual se presenta las dimensiones Maestro, Status, Ciudad, Dia de Corte y Ventas, donde cada dimensión tiene diferentes niveles o hechos, las cuales se enlazan y obtiene las medidas. Para nuestro efecto las medidas son la tasa de respuesta de los clientes y las ventas totales de la oferta (campaña).

Figura 2.3: Modelo Consulta



Fuente: Base de datos de la tienda departamental
Elaborado por: Aguirre, Gallo y Valencia

2.4.2 Esquema relacional de un Data Warehouse

Mediante el esquema relacional, se tiene un soporte donde almacenamos los datos en las tablas de dimensiones y de hechos. Así, se proporciona una vista multidimensional de los datos.

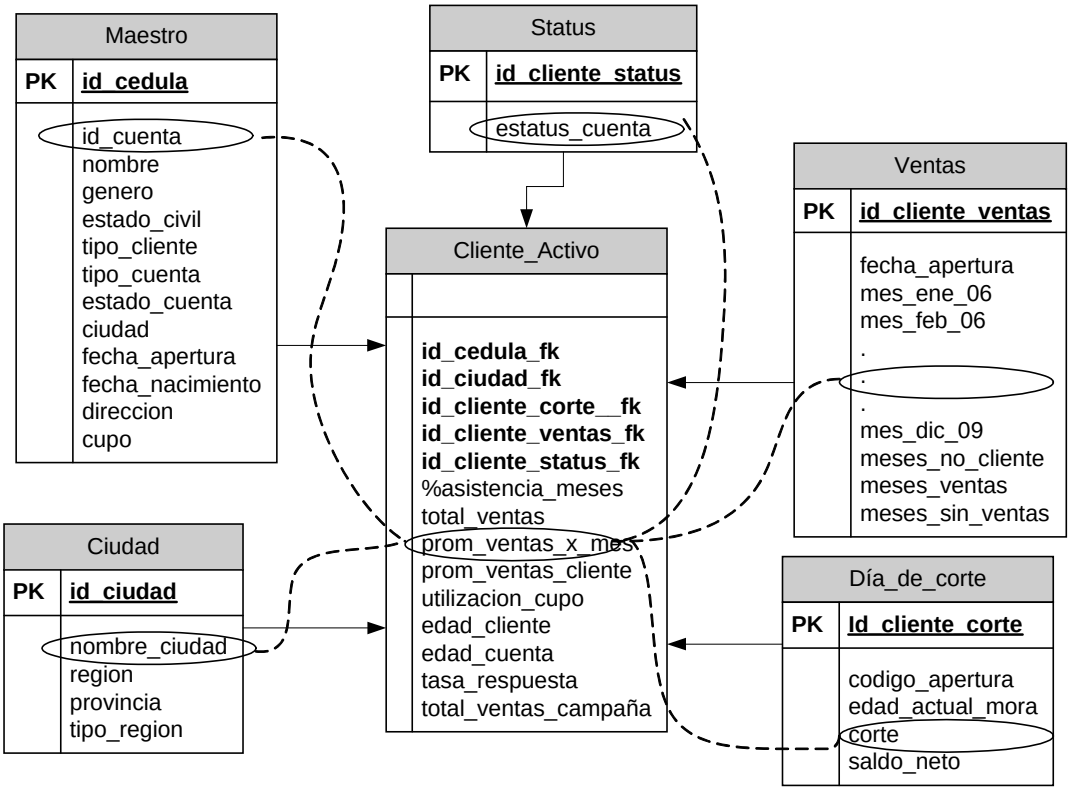
El modelo multidimensional de datos se puede instrumentar por un esquema relacional, donde las dimensiones del cubo son modeladas como relaciones de dimensiones. Los cubos son modelados como funciones del producto cartesiano sobre las dimensiones de las medidas de los datos.

Modelo Estrella

En el modelo estrella, es usado para soportar una operación de datos multidimensionales. Las tablas de dimensiones se relacionan a través de la clave foránea a una sola tabla de hechos. La clave primaria en la tabla de hechos se compone de una relación de las llaves primarias de las tablas de dimensiones. De ser necesario, se puede rediseñar el modelo estrella en un modelo de copo de nieve donde las tablas de dimensiones tienes tablas de sub-dimensiones.

En la Figura 2.4, podemos observar el diseño del modelo estrella. Las dimensiones son las tablas antes mencionas, las cuales se relacionan mediante la clave primaria a la tabla de hechos (Cliente_Activo).

Figura 2.4: Modelo Estrella



Fuente: Base de datos de la tienda departamental

Elaborado por: Aguirre, Gallo y Valencia

2.4.3 Construcción y Carga de Datos

Como se menciono antes, el proceso de construcción es el único responsable de llevar la información de las distintas fuentes al Data Warehouse. La extracción de los datos, su trasformación interna y la carga de los mismos de una forma correcta, aseguran una herramienta ventajosa para la toma de decisiones.

Referente a nuestro estudio, el proceso de extracción implico en la recuperación de datos relevantes de las fuentes; Tablas Maestro, Estatus, Ciudad, Ventas, Día de Corte. Se procedió a la transformación de los datos, se realizaron los siguientes filtros:

De la tabla Maestro, se extrajo aquellos clientes no empleados de la empresa, con un registro único (cedula <>0), cuya cuenta sea de tipo crédito y se encuentre en estado vigente. La tabla Status nos permite depurar las cuentas con estatus bloqueadas. La tabla Día de Corte accedemos a los clientes que se encuentren con mora no mayor a 30 días.

Adicionalmente, creamos variables entre ellas; Edad del cliente, a partir del campo fecha de nacimiento. En la tabla Ventas, consideramos útil la creación de algunos nuevos campos, es decir realizaremos la agrupación de información creando así variables de comportamiento como total ventas al cliente, Edad de la cuenta, % de asistencia mensual, Promedio de ventas por mes, promedio de ventas al cliente y utilización del cupo.

Tabla 2.6 Cliente Activo

Cliente Activo
id_cedula_fk
id_ciudad_fk
id_cliente_corte_fk
id_cliente_ventas_fk
id_cliente_status_fk
%_frecuencia_mensual
total_ventas_cliente
prom_ventas_x_mes
prom_ventas_cliente
utilizacion_cupo
edad_cliente
edad_cuenta
tasa de respuesta
total_ventas_campaña

Fuente: Base de datos de la tienda departamental
Elaborado por: Aguirre, Gallo y Valencia

La creación de esta variable se debe a que son consideradas importantes y las discriminantes que podrían hacer en la realización del árbol de decisión. Finalmente y como último paso realizamos la carga de datos con la información relevante y de esta forma obtenemos la tabla de hecho (Cliente Activo). Ver Tabla 2.6

CAPÍTULO 3

DISEÑO DEL ÁRBOL PARA LA TOMA DE DECISIONES

3.1 Introducción

Este capítulo busca resolver la problemática existente a través de los Árboles de Decisión, conocidos como una herramienta fácil de utilizar y sobretodo comprender. En ciertas ocasiones se considera que para resolver un problema complejo el primer paso es descomponerlos en problemas más simples.

Los árboles de decisión permiten un método en que se pueden desglosar los problemas y la sucesión en el proceso de decisión.

El árbol de decisión permite escoger adecuadamente una estrategia cuando se enfrentan varias alternativas y eventos inciertos al momento de tomar una decisión.

Este capítulo se concentra en la metodología “Árboles de Decisión”, sus ventajas y desventajas. Así también, comprende el diseño y la arquitectura del árbol, su modelamiento y análisis para la toma de decisiones. Finalmente determinaremos el Perfil del Cliente Fiel, las validaciones del modelo y evolución de los principales indicadores.

3.2 Árboles de Decisión

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas[1].

[1] Hernández Orallo J., Ramírez Quintana M. y Ferri Ramírez C., 2004

Los árboles de decisión, son una técnica no paramétrica más utilizada. Frecuentemente se utilizan en la minería de datos como herramienta para analizar datos y realizar predicciones. Esta metodología es usada para predecir variables categóricas (árboles de clasificación) y para predecir variables continuas (árboles de regresión).

En el campo profesional para la toma de decisiones, esta técnica es generalmente usada por su sencillez. El conocimiento que se extrae del árbol se representa de forma clara mediante reglas de decisión y el criterio estadístico aplicando distribuciones de probabilidad en los nodos para llegar a la solución son ventajas que presentan los árboles de decisión asociado a las redes neuronales.

Así también, se detallan ventajas y desventajas importantes a considerar al momento de su diseño e implementación[2]

[2] Portocarrero

Ventajas:

- La regla de asignación son simples y legibles, por tanto la interpretación de resultados es directa e intuitiva.
- Es robusta frente a datos típicos u observaciones mal etiquetadas.
- Es válida sea cual fuera la naturaleza de las variables explicativas: continuas, binarias nominales.
- Es una técnica no paramétrica que tiene en cuenta las interacciones que pueden existir entre los datos.
- Es rápido computacionalmente.

Desventajas:

- Las reglas de asignación son inestables.
- Dificultad para elegir el árbol óptimo.
- Ausencia de una función global de las variables y como consecuencia pérdida de la representación geométrica.
- Los árboles de clasificación requieren un gran número de datos para asegurarse que la cantidad de las observaciones de los nodos hoja es significativa.

Existen diferentes programas para la elaboración de Árboles de Decisión, en nuestro estudio a través de la metodología del SPSS AnswerTree Versión 3.0 determinaremos las tendencias y perfiles, además de poder observar claramente los resultados mediante el diagrama del árbol.

AnswerTree es un sistema de aprendizaje computarizado que crea sistemas de clasificación que se muestran como arboles de decisión. Está diseñado para ser fácil de usar, tanto por estadísticos como por no estadísticos. Con los cuatro eficaces algoritmos de árbol de decisión de AnswerTree, los modelos extraerán de los datos las respuestas que necesita[3]. (SPSS, 2001)

Partiendo del Cubo de información que elaboramos en el capítulo 2, recordemos que nuestro ente de investigación es toda persona que ha abierto una cuenta en la tienda departamental hasta diciembre del año 2008. Definidas las variables predictoras (demográficas y de comportamiento) que serán analizadas mediante el AnswerTree, el primer paso para el desarrollo del árbol es determinar la variable de criterio que será el nodo de partida o inicio.

[3] Guía de usuario de AnswerTree 3.0 – SPSS Inc, 2001

A raíz de la variable Ventas 2009 de la cual obtenemos su promedio, creamos la variable de criterio BM definiendo como Cliente Bueno “B” si ventas 2009 es iguales o superior al promedio, caso contrario (ventas 2009 < Promedio) será considerado Cliente Malo “M”.

3.2.1 Diseño y Arquitectura

Un árbol es un conjunto de nodos y ramas, donde un nodo es un punto de unión y cada rama es un arco conector. Cada nodo representa un subconjunto de la población y es un punto en el que se debe tomar una decisión. De los nodos, salen ramas las cuales simbolizan las decisiones posibles. El nodo raíz constituye toda la población y no tiene ramas entrantes.

La presentación de la información se hace en un diagrama en forma de árbol invertido donde el proceso recursivo, muy esquemáticamente, se traduce en los siguientes pasos[4]. (Schiattino Lemus & Silva Zamora, 2008)

a. El nodo raíz es dividido en subgrupos (dos o más) determinados por la partición de una variable predictora elegida, generando nodos hijos.

[4] Schiattino Lemus & Silva Zamora, 2008

b. Los nodos hijos son divididos usando la partición de una nueva variable. El proceso recursivo se repite para los nuevos nodos hijos sucesivamente hasta que se cumpla alguna condición de parada.

c. Algunos de los nodos resultantes son terminales, mientras que otros nodos continúan dividiéndose hasta llegar a un nodo terminal.

d. En cada árbol se cumple la propiedad de tener un camino único entre el nodo raíz y cada uno de los demás nodos del árbol.

Mediante AnswerTree, se puede diagramar el árbol a través de diversos algoritmos, para que se pueda elegir el mejor modelo para unos datos específicos[5].

- CHAID – Detector automático de interacciones mediantes chi-cuadrado: método que utiliza estadísticos de chi-cuadrado para identificar divisiones óptimas.

[5] Guía de usuario de AnswerTree 3.0 – SPSS Inc, 2001

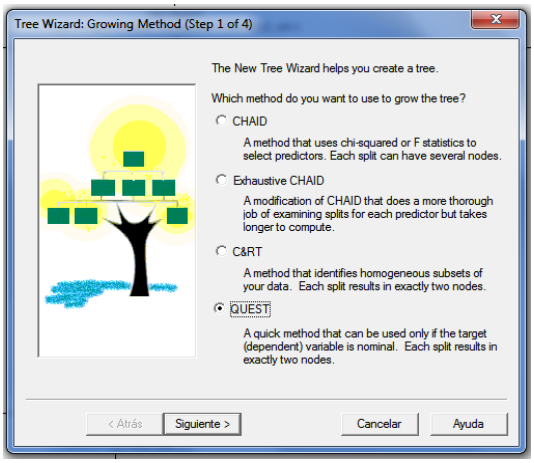
- CHAID exhaustivo – Es una modificación de CHAID que realiza un análisis más detallado de todas las divisiones posibles de cada predictor; por ello tarda más tiempo en entregar el cálculo correspondiente.

- C&RT - Árboles de clasificación y regresión - *es* un completo algoritmo de árbol binario para dividirlos datos y generar subconjuntos homogéneos precisos.

- QUEST - es un Árbol estadístico rápido, insesgado y eficiente: método de cálculo rápido que evita los sesgos de otros métodos, favoreciendo así a predictores con varias categorías.

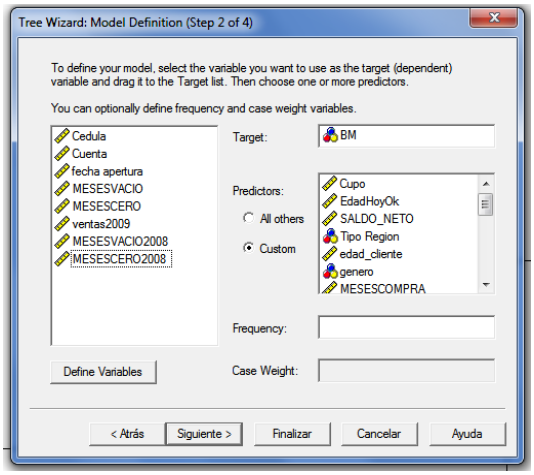
Para el correcto análisis y determinación del perfil de un cliente fiel para una tienda departamental de moda y hogar. Iniciamos la creación del árbol a través del algoritmo QUEST, con la variable de criterio BM establecida, seleccionamos las variables predictoras.

Figura 3.1: Método Quest para la construcción del árbol de decisión



Fuente: AnswerTree ver 3.0

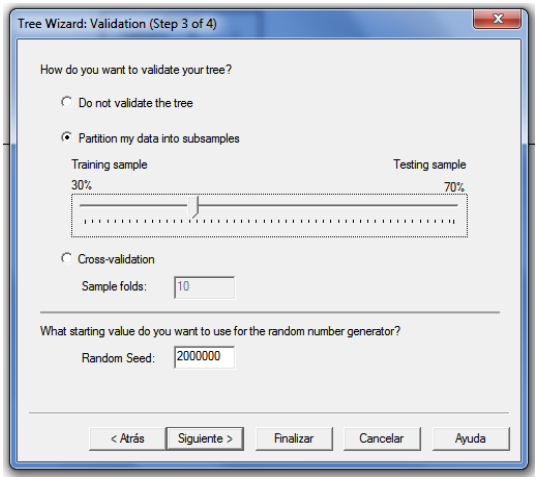
Figura 3.2: Pantalla para la Selección de las variables



Fuente: AnswerTree ver 3.0

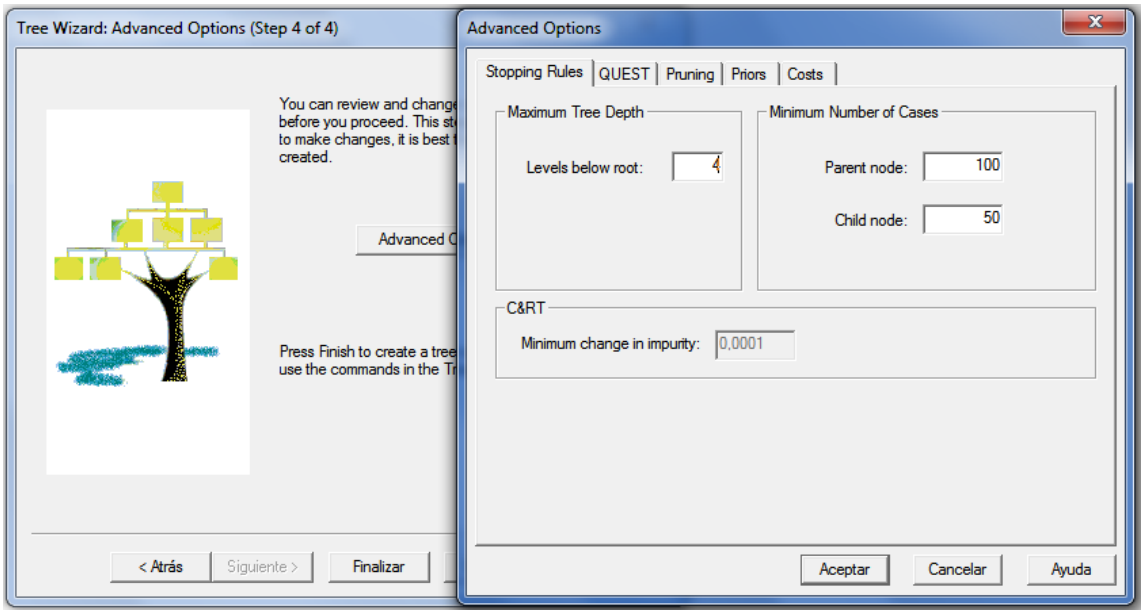
Una vez definida la variable dependiente y predictoras, asignamos una muestra de entrenamiento del 30% y una profundidad máxima de 4 niveles.

Figura 3.3: Pantalla de validación



Fuente: AnswerTree ver 3.0

Figura 3.4: Pantalla de ramificación de árbol de decisión



Fuente: AnswerTree ver 3.0

3.2.2 Modelamiento y Recorrido

El modelamiento es la lectura de los nodos y ramas presentados a raíz de la construcción del Árbol de Decisión. La secuencia temporal se desarrolla de izquierda a derecha. Las ramas que llegan a un nodo desde la izquierda son una detención (parada). Las ramas que salen hacia la derecha todavía pueden ser más ramificadas. De esta forma, definimos los recorridos y reglas de paradas de un árbol de decisión.

Se llama recorrido de un árbol al proceso que permite acceder una sola vez a cada uno de los elementos del árbol para examinar el conjunto completo[6].

1. INORDEN(Sufijo)

   Recorrer el subarbol izquierdo en inorden.

   Examinar la raíz.

   Recorrer el subarbol derecho en inorden.

2. PREORDEN(Prefijo)

   Examinar la raíz.

   Recorrer el subárbol izquierdo en preorden.

   recorrer el subárbol derecho en preorden.

3. POSTORDEN(Posfijo)

   Recorrer el subárbol izquierdo en postorden.

   Recorrer el subárbol derecho en postorden.

   Examinar la raíz

[6] López González, 2008

Existen distintos criterios de parada que pueden provocar la finalización de los algoritmos que realizan árboles de clasificación o regresión. Las causas que pueden provocar la finalización son[7]:

- Se ha alcanzado la máxima profundidad del árbol permitida.

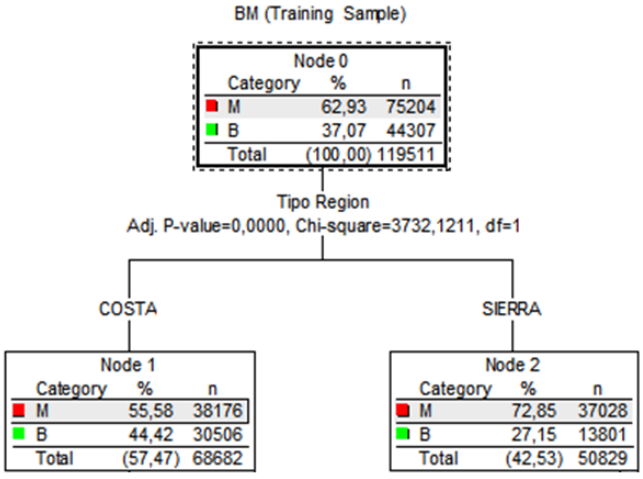
- No se puede realizar más particiones, porque se ha verificado alguna de las siguientes condiciones:
  a. No hay variables explicativas significativas para realizar la partición del nodo.
  b. El número de elementos en el nodo terminal es inferior al número mínimo de casos permitidos para poder realizar la partición.
  c. El nodo no se podrá dividir en el caso en el cual el número de casos en uno o más nodos hijos sea menor que el mínimo número de casos permitidos por nodo.

[7] Puerta Goicoechea, 2002

El nodo raíz del árbol, ver figura 3.5, presenta un desglose de los casos de la muestra tomada. En nuestro conjunto de datos, los casos se encuentran distribuidos entre aquellos que son considerados clientes buenos (37.07%) y aquellos clientes malos (62.93%). Esto también se refleja en la estimación de riesgo, donde al otorgarle a todos los casos el valor de la mayoría (malo), se obtiene una tasa de error del 0.37.

Una vez definido el nodo raíz, procedemos a dividir los datos creando subgrupos con propiedades deseadas (homogéneos).

Figura 3.5: Nodo raíz y Primer nivel



Fuente: Cubo de información

Elaborado por: Aguirre, Gallo y Valencia

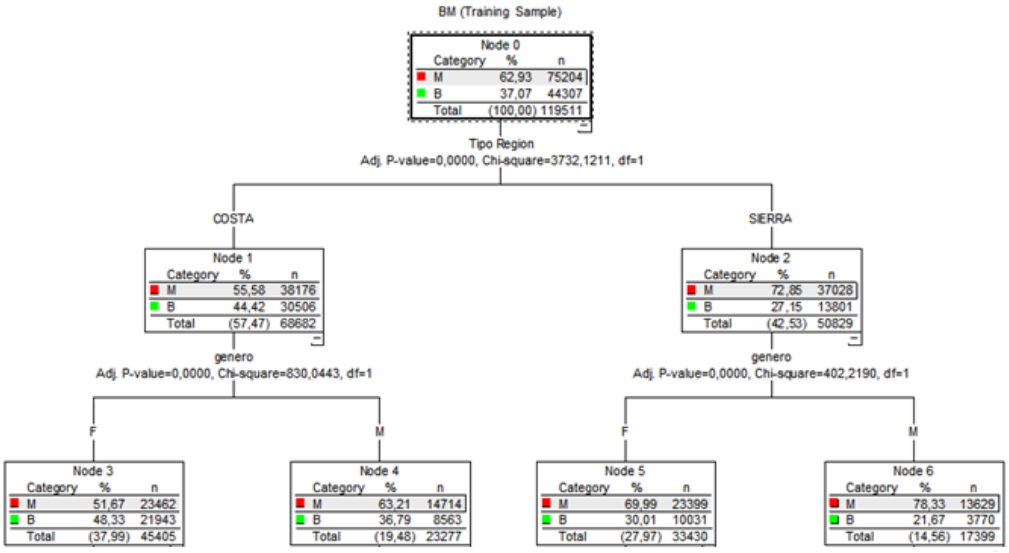
El desarrollo del árbol en su primer nivel nos muestra como variable discriminante a “Tipo Región” (Costa o Sierra), dividiendo el nodo raíz en dos nodos filiales, denotando dos claros sectores o mercados diferentes. En otras palabras, a raíz de la primera variable predictora se generan dos árboles independientes. Un análisis de riesgo refuerza esta conclusión. La estimación de riesgo se mantiene en 0.37, lo cual indica que si se utiliza la regla de decisión basada en el árbol activo, se clasificaran correctamente 100% - 37.07% = 62.93% de los casos.

Si analizamos la rama izquierda, vemos que entre quienes pertenecen a la Región Costa ver figura 3.5 presentan una tase de clientes buenos del 44.42% (30,506 clientes). Por otro lado, la rama derecha (Región Sierra) tiene más probabilidades de tener un cliente malo con un 72,85%.

Al definir un segundo nivel para el árbol, ver figura 3.6, tenemos la variable predictiva “Genero” (Masculino o Femenino). Concentrando nuestra atención a partir del árbol generado por la Región Costa, se pone de manifiesto que los clientes de género femenino son mejores que los masculinos con una tasa de 48.33% vs 36.79%. En la otra rama, la diferencia se mantiene, pero con una menor proporción, las mujeres son

mejores clientes que los hombres con un 30.01% y 21,67% respectivamente.

Figura 3.6: Segundo nivel



Fuente: Cubo de información

Elaborado por: Aguirre, Gallo y Valencia

Hasta el momento los escenarios son claros, la región costa presenta una mejor tasa de clientes buenos que la sierra y en ambas regiones el cliente femenino supera al masculino. Agregamos entonces una variable mas, de esta forma el árbol crece a un tercer nivel. Las variables predictoras para los nodos existentes son diferentes, sin embargo cabe destacar que a partir del presente nivel, la estimación de riesgo disminuye a una tasa de error del 0.25.

Para la rama derecha la situación es diferente ver figura 3.7, para ambos géneros la variable discriminante es “Ventas2008” y su ramificación es casi homogénea. De esta forma, mediante esta rama establecemos una mejor tasa de buenos clientes siempre que sus ventas del año 2008 sean superiores para las mujeres a los $ 464 y para los hombres a $ 472, en proporción determinamos; Femenino 76.28% y Masculino 71.24%. Así también, observamos los clientes cuyas ventas fueron hasta $464 (21.62%) en el caso de las mujeres y para los hombres cuyas ventas fueron no mayor de $ 472 (15.66%).

Las variables antes analizadas aportan información interesante que nos ayudan a determinar el perfil del clientes fiel, sin embargo nuestra definición inicial fue una profundidad del árbol de cuatro niveles. Agregando un nivel más, el resumen de riesgos muestra al árbol con una estimación de riesgo de 0.24, ver Tabla 3.1. Es decir, el árbol garantiza que para cada nodo el 76% de los casos están clasificados en forma exacta.

Para aquellos clientes de la costa, femeninos y con compras realizadas durante 5 meses o menos, la ramificación del árbol finaliza con la variable

"Ventas2008", ver figura 3.8, dividiendo a los clientes con ventas hasta o superiores a $ 398 dólares con una tasa de buenos clientes de 26.76% y 64.51% respectivamente. Mientras que para los clientes que compraron durante más de 5 meses en el año 2008, el árbol ramifica a través del la variable "PromedioCompraCliente" fragmentando el promedio de compras histórico inferior o superior a $ 33, 58.19% y 86.43% respectivamente.

Tabla 3.1 Estimación de riesgo

Training Sample				
Misclassification Matrix				
		Actual Category		
		M	B	Total
Predicted Category	M	61770	15272	77042
	B	13434	29035	42469
	Total	75204	44307	119511
		Risk Statistics		
Risk Estimate		0,240195		
SE of Risk Estimate		0,00123575		

Fuente: Base de datos de la tienda departamental

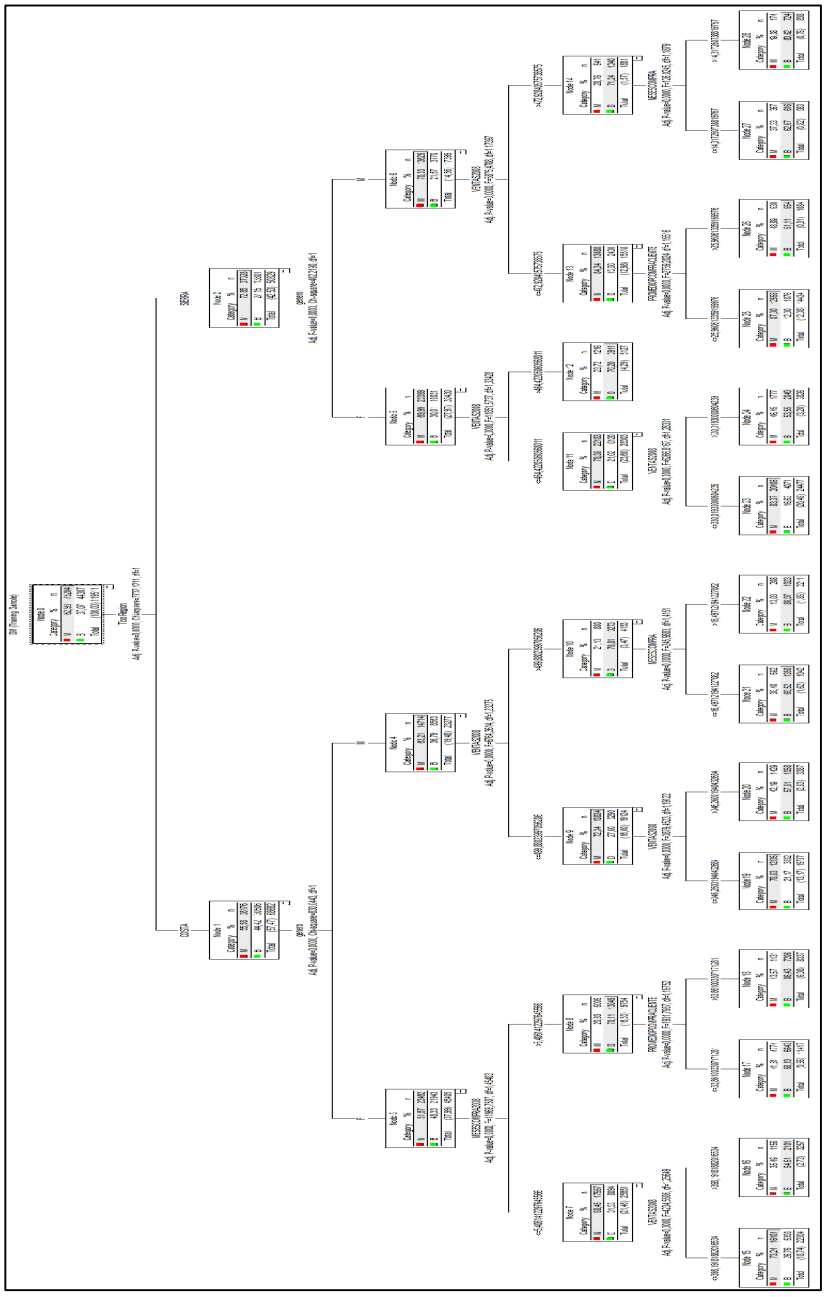
Elaborado por: Aguirre, Gallo y Valencia

En el caso de los hombres de la costa, los clientes con compras superiores a $ 499 tienen una ramificación más a través de la variable "MesesCompra" la cual divide a los clientes que hayan comprado hasta 16 meses o más a través de su existencia como socios de la tienda departamental.

Por otro lado, en el árbol generado de la región Sierra, para el genero femenino nuevamente la variable “Ventas 2008” ver figura 3.9 se manifiesta, segmentando los clientes buenos en 3 rangos bien definidos; Los que compraron en el año 2008 hasta a $ 330 (16.63%); Los clientes con compras superiores a $ 330 e inferiores a $ 464 (53.55%); Y los clientes con compras superior a $ 464 (76.28%).

El género masculino, la variable predictora es “PromedioCompraCliente” separando a los clientes con un promedio de compras histórico máximo de $ 25 o superior al mismo. Mientras que para las compras del año 2008 superiores a los $ 472 dólares, el árbol fragmenta a los clientes a través de la variable “MesesCompra” indicando que el 63% de los clientes ha comprado como mucho 14 meses a través de su existencia como socios de la tienda departamental, mientras que el 80% ha comprado más de 14 meses.

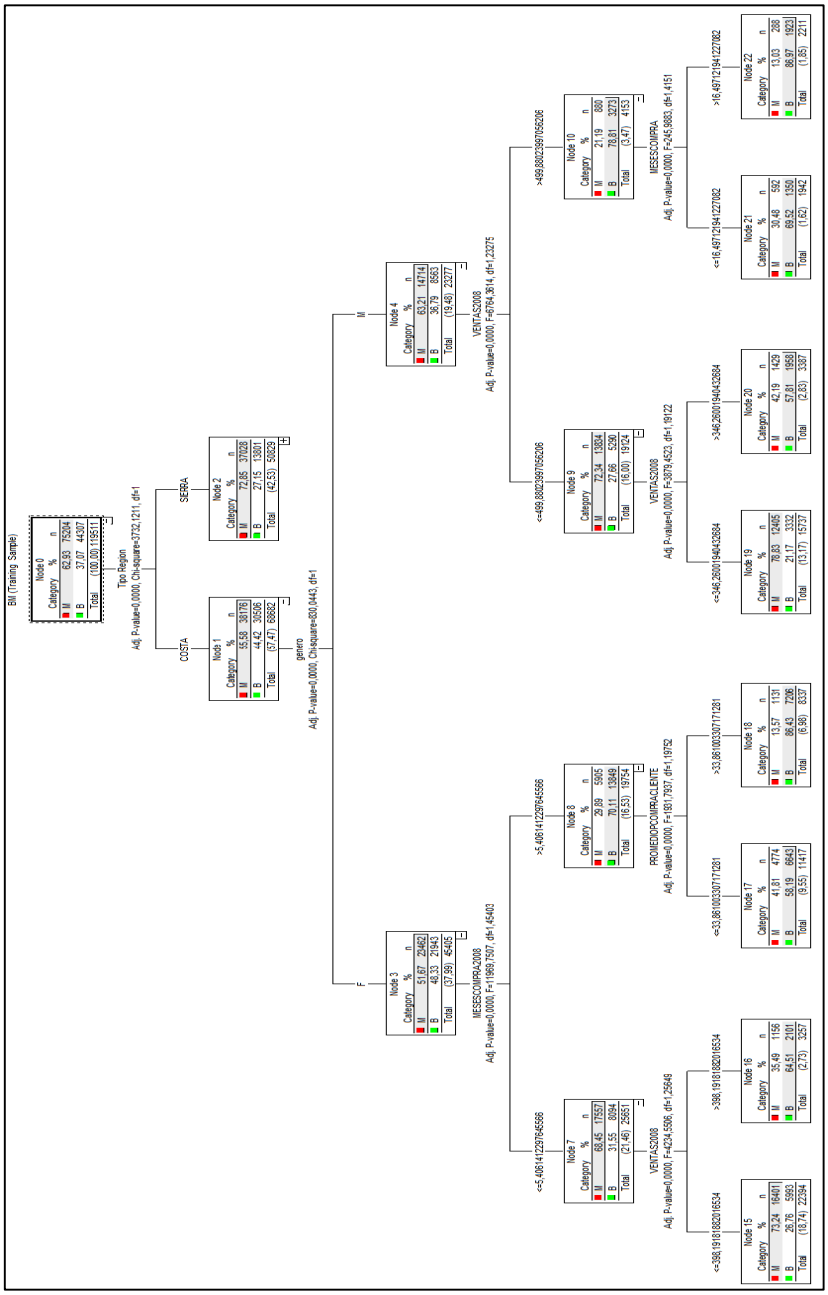
Figura 3.7: Árbol de decisión general



Fuente: Cubo de información

Elaborado por: Aguirre, Gallo y Valencia

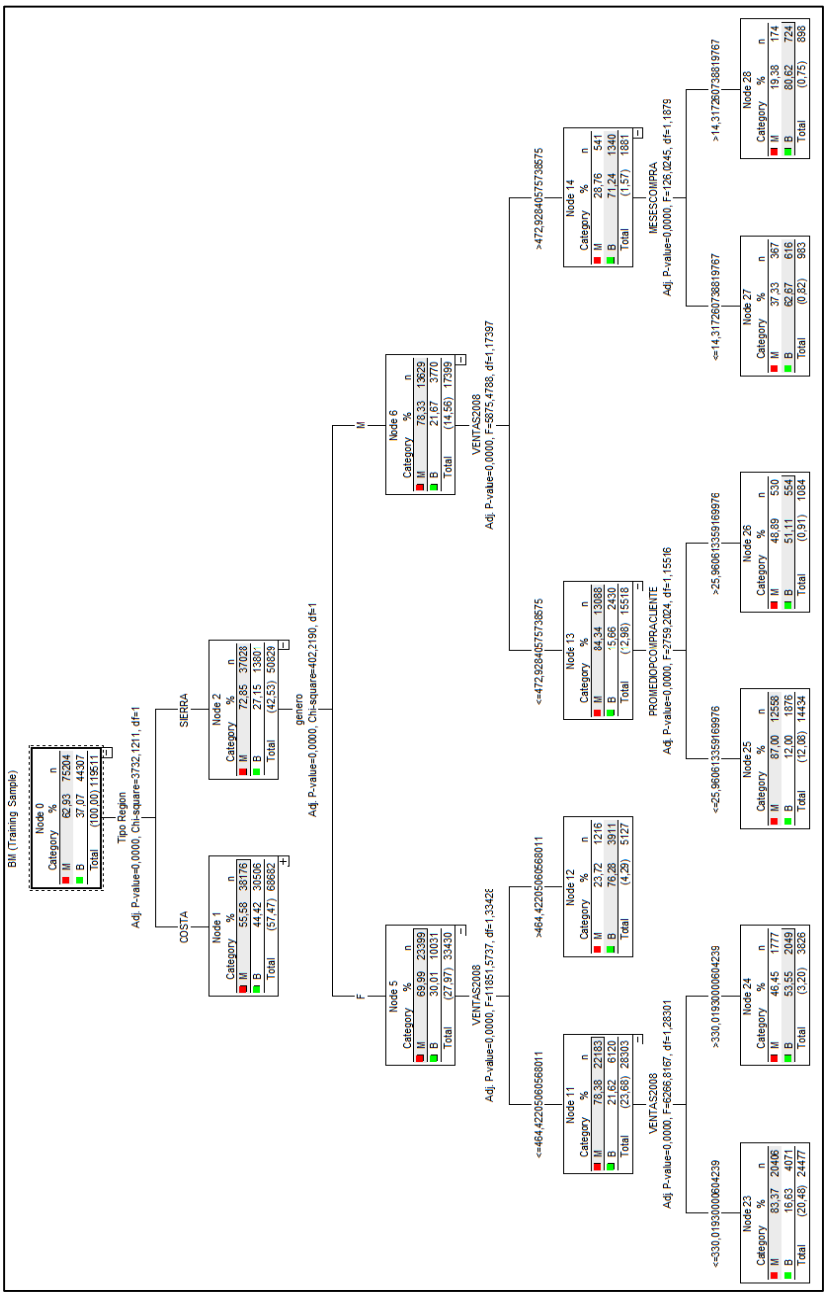
Figura 3.8: Árbol de decisión región costa



Fuente: Cubo de información

Elaborado por: Aguirre, Gallo y Valencia

Figura 3.9: Árbol de decisión región sierra



Fuente: Cubo de información

Elaborado por: Aguirre, Gallo y Valencia

El resumen de ganancias también proporciona una idea más clara del árbol. Este, muestra los nodos que tiene la mayor y la menor proporción de categorías criterio (Bueno o Malo). En este caso, lo interesante es saber que subgrupo de clientes (nodos) tiene mayores probabilidades de ser clasificado como bueno. Ver Tabla 3.2

Tabla 3.2 Resumen de ganancia

Gain Summary

Target variable: BM Target category: B

	Node-by-Node						Cumulative Statistics					
Nodes	Node: n	Node: %	Gain: n	Gain (%)	Resp: %	Index (%)	Node: n	Node: %	Gain: n	Gain (%)	Resp: %	Index (%)
22	2211	1,9	1923	4,3	87,0	234,6	2211	1,9	1923	4,3	87,0	234,6
18	8337	7,0	7206	16,3	86,4	233,1	10548	8,8	9129	20,6	86,5	233,4
28	898	0,8	724	1,6	80,6	217,5	11446	9,6	9853	22,2	86,1	232,2
12	5127	4,3	3911	8,8	76,3	205,8	16573	13,9	13764	31,1	83,1	224,0
21	1942	1,6	1350	3,0	69,5	187,5	18515	15,5	15114	34,1	81,6	220,2
16	3257	2,7	2101	4,7	64,5	174,0	21772	18,2	17215	38,9	79,1	213,3
27	983	0,8	616	1,4	62,7	169,0	22755	19,0	17831	40,2	78,4	211,4
17	11417	9,6	6643	15,0	58,2	156,9	34172	28,6	24474	55,2	71,6	193,2
20	3387	2,8	1958	4,4	57,8	155,9	37559	31,4	26432	59,7	70,4	189,8
24	3826	3,2	2049	4,6	53,6	144,5	41385	34,6	28481	64,3	68,8	185,6
26	1084	0,9	554	1,3	51,1	137,9	42469	35,5	29035	65,5	68,4	184,4
15	22394	18,7	5993	13,5	26,8	72,2	64863	54,3	35028	79,1	54,0	145,7
19	15737	13,2	3332	7,5	21,2	57,1	80600	67,4	38360	86,6	47,6	128,4
23	24477	20,5	4071	9,2	16,6	44,9	105077	87,9	42431	95,8	40,4	108,9
25	14434	12,1	1876	4,2	13,0	35,1	119511	100,0	44307	100,0	37,1	100,0

In versions prior to AnswerTree 3.0 the Gains column was known as Responses and vice versa.

Fuente: Base de datos de la tienda departamental

Elaborado por: Aguirre, Gallo y Valencia

Podemos observar que el nodo 18 representa a los clientes de la región costa, de género femenino que hayan comprado más de 5 meses en el año 2008 y que su promedio de compra histórico sea superior a $ 33. El número de casos de este nodo es 8,337 y el porcentaje de ellos es 7%. La ganancia es el numero de personadas del nodo con una clasificación buena (7,206) y el porcentaje de todas estas personas en el nodo es de 16.3%. Su respuesta es de 86.4% (indica la proporción de casos del nodo que tiene la respuesta criterio "cliente bueno") y el índice es 233.1% (medida de cómo el numero de respuestas criterio del nodo se compra con toda la muestra).

3.3 Análisis y Criterios para la Toma de Decisión

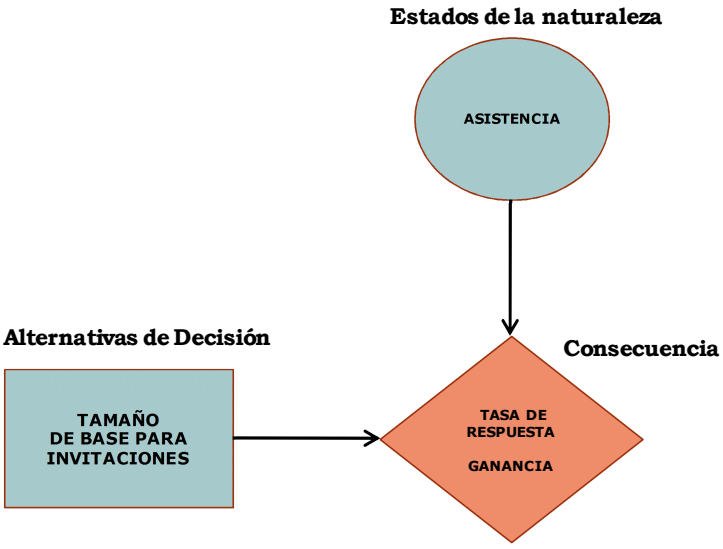
La toma de decisión es fundamental para cualquier actividad humana. En este sentido, somos todos tomadores de decisiones. Sin embargo tomar una buena decisión empieza con un proceso de razonamiento, constante y focalizado, que puedo incluir varias disciplinas[8].

[8] Amaya Amaya, 1998

Habitualmente un sin números de decisiones son tomadas en cualquier empresa, estas en la mayoría de los casos son las responsables del éxito del proyecto y sus resultados futuros. La toma de decisiones proviene directamente del análisis de la decisión el cual explica los diferentes factores que participan y ayuda en la elección de las opciones más adecuadas.

Las alternativas elegidas (las decisiones) y la ocurrencia de un particular evento no controlable (estados de la naturaleza) se pueden relacionar gráficamente a través de la herramienta conocida como diagrama de influencia.

Figura 3.10: Diagrama de influencia



Fuente: Base de datos de la tienda departamental

Elaborado por: Aguirre, Gallo y Valencia

Durante el análisis, existen distintos ambientes (certeza, bajo riesgo, bajo incertidumbre) en los cuales se toman decisiones, estas situaciones se catalogan según el discernimiento y manejo que se tenga sobre las variables involucradas en el problema, considerando que la decisión final que se tome va a estar condicionada por aquellas variables.

De igual forma, al momento de decidir, se puede optar por el criterio Optimista, Pesimista o Conservador que nos permiten evaluar las alternativas. Estos distintos criterios para tomar la decisión están condicionados por los ambientes antes mencionados que se definen de la siguiente forma:

a. Certeza, cuando al momento de tomar la decisión, se conoce el comportamiento de los eventos no controlables siendo el único inconveniente seleccionar la mejor estrategia.

b. Bajo riesgo, cuando se conocen las probabilidades de que suceda cada evento no controlable y se debe probar la estrategia decidida con la probabiliodad de cada evento incontrolable.

c. Bajo incertidumbre, cuando se desconoce y no se puede determinar las probabilidades de los eventos no controlables.

En este ambiente se sugiere optar por algún criterio de decisión definidos a continuación[9].

- El enfoque optimista evalúa cada alternativa de decisión en función del mejor resultado que pueda ocurrir. La alternativa que se recomienda es la que da el mejor resultado posible.

- El enfoque pesimista evalúa cada alternativa de decisión desde el punto de vista del peor resultado que pueda ocurrir. La alternativa de decisión recomendada es la que proporciona el mejor de los peores resultados posibles.

- El enfoque de arrepentimiento para la toma de decisiones no es puramente optimista ni puramente conservador.

[9] Anderson, 2004

3.3.1 Determinación del Perfil de un Cliente Fiel

Sin duda, el mercado competitivo que existe, con lleva a buscar diferentes formas de captación de clientes, a quienes ofrecerle y venderles los productos, pero mas allá de la comercialización que se pueda tener con un cliente, hoy en día las empresas deben conseguir también el mantener la relación 100% con estos clientes, satisfacer sus necesidades de acuerdo al mercado. Es decir convertir a un cliente normal en un cliente fiel.

Conociendo la actualidad de las empresas, definimos a un cliente como una persona que usualmente adquiere algún tipo de producto o servicios en un establecimiento. Conseguir un cliente fiel a una empresa ofrece varias ventajas que bien evaluadas permiten el desarrollo y crecimiento a gran escala en el mercado. Citamos algunas frases de cliente fiel:

“Un cliente fiel y, por lo tanto, satisfecho, es la mejor fuente de comunicación para la empresa: mucho más creíble y barata que la publicidad en medios masivos”.

"Atender a un cliente fiel supone un ahorro de costes para la empresa, porque en la medida en que se conocen mejor sus caprichos cuesta menos atenderle bien".

"Los clientes fieles son menos sensibles a los precios, asimilan mejor los precios elevados, porque también sienten que perciben valores adicionales en los servicios o en las personas que los prestan".
"Los clientes de toda la vida son la mejor fuente de ideas de nuevos productos o de mejora de los servicios ofrecidos".

Contar con clientes fieles es el reto de toda empresa y organización, principalmente hoy que los mercados están apretados y altamente competitivos. Un cliente fiel es aquel con el cual ya hemos establecido una sólida relación, mantenemos niveles de ventas, están satisfechos con nuestro servicio y principalmente son una referencia ante nuevos o futuros clientes. Aunque ya se ha repetido infinidad de veces, conseguir un nuevo cliente cuesta mucho más que mantener a uno actual[10].

[10] Alarcón, 2009

Ahora bien, el estudio del grupo de tablas que contiene información relevante obtenidas de las bases de datos es un lente de gran alcance para analizar mucho mejor la situación actual. Consecuente con esto, y una vez analizadas las distintas variables involucradas por ejemplo cuántas mujeres y hombres tengo, a que región pertenecen, cuánta gente me compró en épocas pasadas, la frecuencias de tiempo, entre otras, determinamos el perfil de un cliente fiel por ciudad donde existe una tienda departamental de moda y hogar de la siguiente forma.

Para Guayaquil (Región Costa) ver figura 3.8, definimos los perfiles de clientes fieles a:

1. *Femenino, MesesCompra2008 > 5 meses, PromedioCompraCliente > $ 33*
2. *Masculino, Ventas2008 > $ 499, MesesCompra > 16 meses*
3. *Femenino, MesesCompra2008 > 5 meses, PromedioCompraCliente <= $ 33*
4. *Femenino, MesesCompra2008 <= 5 meses, Ventas2008 <= $ 398*
5. *Masculino, Ventas2008 <= $ 346*
6. *Femenino, MesesCompra2008 <= 5 meses, Ventas2008 > $ 398*
7. *Masculino, Ventas2008 > $ 346 y <= $ 499*
8. *Masculino, Ventas2008 > $ 499, MesesCompra <= 16 meses*

Para Quito (Región Sierra) ver figura 3.9, definimos como perfiles de clientes fieles a:

1. *Femenino, Ventas2008 > $ 464*
2. *Masculino, Ventas2008 > $ 472, MesesCompra > 14 meses*
3. *Femenino, Ventas2008 <= $ 330*
4. *Femenino, Ventas2008 > $ 330 y <= $ 464*
5. *Masculino, Ventas2008 <= $ 472, PromedioCompraCliente <= $ 25*
6. *Masculino, Ventas2008 > $ 472, MesesCompra <= 14 meses*
7. *Masculino, Ventas2008 <= $ 472, PromedioCompraCliente > $ 25*

3.4 Verificación y Validación del Modelo

Una vez determinado el perfil, el siguiente pasó y quizás uno de los más interesantes es la verificación y validación del modelo. Siempre será recomendable durante su elaboración y validación la cercanía con los usuarios finales a fin de mitigar alguna desconfianza de los mismos, dado que el resultado final se utilizará en el objetivo de concluir para la situación real de la empresa.

Es muy importante tener claro cuáles son los conceptos y las bases teóricas de los árboles de decisión para garantizar un juicio de valor que el modelo que será

utilizado apropiadamente, por lo cual la verificación, validación y calibración son los términos más apropiados para proporcionar el nivel aceptable de credibilidad requerida.

Así, definimos Verificación como el proceso de determinar si la lógica operacional del modelo (programa de ordenador) se corresponde con la lógica del diseño. En términos más simples, consiste en determinar si hay errores en el programa (Universidad de Jaén). La Validación es el proceso de comparar la salida del modelo con el comportamiento del fenómeno. En otras palabras, es comparar la ejecución del modelo con la realidad (física o otra cualquiera). (ETSII) y Calibración es un conjunto de operaciones (prueba y ajuste) de los medidas existentes, es decir es la relación entre los valores indicados por un instrumento de medición y los valores correspondientes a las magnitudes establecidas por los patrones[11].

El proceso de validación denota en conocer cómo se comporta la estructura del árbol para generalizar los datos disponibles en una muestra más grande. Cabe recordar que el árbol fue creado a raíz de una muestra de entrenamiento del

[11] SENCAMER, http://portal.sencamer.gov.ve

30%, desde la cual se genera el modelo, es decir la muestra de comprobación es del 70%, en la cual se prueba el modelo generado.

Realizamos entonces la comprobación del modelo y tenemos los siguientes resultados a raíz del resumen de ganancias:

Muestra de entrenamiento;

Tabla 3.3 Resumen ganancia (Entrenamiento 30%)

Gain Summary

Target variable: BM Target category: B

	Node-by-Node						Cumulative Statistics					
Nodes	Node: n	Node: %	Gain: n	Gain (%)	Resp: %	Index (%)	Node: n	Node: %	Gain: n	Gain (%)	Resp: %	Index (%)
22	2211	1,9	1923	4,3	87,0	234,6	2211	1,9	1923	4,3	87,0	234,6
18	8337	7,0	7206	16,3	86,4	233,1	10548	8,8	9129	20,6	86,5	233,4
28	898	0,8	724	1,6	80,6	217,5	11446	9,6	9853	22,2	86,1	232,2
12	5127	4,3	3911	8,8	76,3	205,8	16573	13,9	13764	31,1	83,1	224,0
21	1942	1,6	1350	3,0	69,5	187,5	18515	15,5	15114	34,1	81,6	220,2
16	3257	2,7	2101	4,7	64,5	174,0	21772	18,2	17215	38,9	79,1	213,3
27	983	0,8	616	1,4	62,7	169,0	22755	19,0	17831	40,2	78,4	211,4
17	11417	9,6	6643	15,0	58,2	156,9	34172	28,6	24474	55,2	71,6	193,2
20	3387	2,8	1958	4,4	57,8	155,9	37559	31,4	26432	59,7	70,4	189,8
24	3826	3,2	2049	4,6	53,6	144,5	41385	34,6	28481	64,3	68,8	185,6
26	1084	0,9	554	1,3	51,1	137,9	42469	35,5	29035	65,5	68,4	184,4
15	22394	18,7	5993	13,5	26,8	72,2	64863	54,3	35028	79,1	54,0	145,7
19	15737	13,2	3332	7,5	21,2	57,1	80600	67,4	38360	86,6	47,6	128,4
23	24477	20,5	4071	9,2	16,6	44,9	105077	87,9	42431	95,8	40,4	108,9
25	14434	12,1	1876	4,2	13,0	35,1	119511	100,0	44307	100,0	37,1	100,0

In versions prior to AnswerTree 3.0 the Gains column was known as Responses and vice versa.

Fuente: Base de datos de la tienda departamental

Elaborado por: Aguirre, Gallo y Valencia

Podemos observar que la tendencia en ganancia, respuesta e índice se mantiene. Para el modelo generado con la muestra de entrenamiento, obtenemos para el nodo 18 una ganancia del 16,3%, mientras que la ganancia en la muestra de comprobación es de 16,7%.

Si revisamos el nodo 17, *Región costa*, *Femenino, MesesCompra2008 > 5 meses y PromedioCompraCliente <= $ 33;* tenemos una ganancia de 15% vs 14.8% en la muestra de entrenamiento y comprobación respectivamente.

Muestra de Comprobación;

Tabla 3.4 Resumen ganancia (Comprobación 70%)

Gain Summary

Target variable: BM Target category: B

	Node-by-Node						Cumulative Statistics					
Nodes	Node: n	Node: %	Gain: n	Gain (%)	Resp: %	Index (%)	Node: n	Node: %	Gain: n	Gain (%)	Resp: %	Index (%)
22	914	1,8	803	4,2	87,9	235,3	914	1,8	803	4,2	87,9	235,3
18	3692	7,1	3225	16,7	87,4	234,0	4606	8,9	4028	20,9	87,5	234,3
28	365	0,7	302	1,6	82,7	221,6	4971	9,6	4330	22,4	87,1	233,3
12	2259	4,4	1741	9,0	77,1	206,4	7230	14,0	6071	31,5	84,0	224,9
21	901	1,7	642	3,3	71,3	190,9	8131	15,7	6713	34,8	82,6	221,2
16	1438	2,8	964	5,0	67,0	179,6	9569	18,5	7677	39,8	80,2	214,9
27	432	0,8	260	1,3	60,2	161,2	10001	19,4	7937	41,1	79,4	212,6
17	4855	9,4	2862	14,8	58,9	157,9	14856	28,8	10799	56,0	72,7	194,7
20	1482	2,9	869	4,5	58,6	157,1	16338	31,6	11668	60,5	71,4	191,3
24	1604	3,1	852	4,4	53,1	142,3	17942	34,7	12520	64,9	69,8	186,9
26	485	0,9	261	1,4	53,8	144,2	18427	35,7	12781	66,3	69,4	185,8
15	9761	18,9	2506	13,0	25,7	68,8	28188	54,6	15287	79,3	54,2	145,3
19	6808	13,2	1504	7,8	22,1	59,2	34996	67,7	16791	87,1	48,0	128,5
23	10449	20,2	1680	8,7	16,1	43,1	45445	88,0	18471	95,8	40,6	108,9
25	6221	12,0	817	4,2	13,1	35,2	51666	100,0	19288	100,0	37,3	100,0

In versions prior to AnswerTree 3.0 the Gains column was known as Responses and vice versa.

Fuente: Base de datos de la tienda departamental

Elaborado por: Aguirre, Gallo y Valencia

Así también, realizamos otra prueba de la siguiente forma; Tomamos una muestra de entrenamiento del 10%, generamos el árbol y lo probamos en la muestra de comprobación del 90% de los datos. Vemos que las valores de ganancia, respuesta e índice no difieren con la muestra anterior (30%). Ver Tabla 3.5

Tabla 3.5 Resumen ganancia (Comprobación 90%)

Gain Summary

Target variable: BM Target category: B

	Node-by-Node						Cumulative Statistics					
Nodes	Node: n	Node: %	Gain: n	Gain (%)	Resp: %	Index (%)	Node: n	Node: %	Gain: n	Gain (%)	Resp: %	Index (%)
22	2211	1,9	1923	4,3	87,0	234,6	2211	1,9	1923	4,3	87,0	234,6
18	8337	7,0	7206	16,3	86,4	233,1	10548	8,8	9129	20,6	86,5	233,4
28	898	0,8	724	1,6	80,6	217,5	11446	9,6	9853	22,2	86,1	232,2
12	5127	4,3	3911	8,8	76,3	205,8	16573	13,9	13764	31,1	83,1	224,0
21	1942	1,6	1350	3,0	69,5	187,5	18515	15,5	15114	34,1	81,6	220,2
16	3257	2,7	2101	4,7	64,5	174,0	21772	18,2	17215	38,9	79,1	213,3
27	983	0,8	616	1,4	62,7	169,0	22755	19,0	17831	40,2	78,4	211,4
17	11417	9,6	6643	15,0	58,2	156,9	34172	28,6	24474	55,2	71,6	193,2
20	3387	2,8	1958	4,4	57,8	155,9	37559	31,4	26432	59,7	70,4	189,8
24	3826	3,2	2049	4,6	53,6	144,5	41385	34,6	28481	64,3	68,8	185,6
26	1084	0,9	554	1,3	51,1	137,9	42469	35,5	29035	65,5	68,4	184,4
15	22394	18,7	5993	13,5	26,8	72,2	64863	54,3	35028	79,1	54,0	145,7
19	15737	13,2	3332	7,5	21,2	57,1	80600	67,4	38360	86,6	47,6	128,4
23	24477	20,5	4071	9,2	16,6	44,9	105077	87,9	42431	95,8	40,4	108,9
25	14434	12,1	1876	4,2	13,0	35,1	119511	100,0	44307	100,0	37,1	100,0

In versions prior to AnswerTree 3.0 the Gains column was known as Responses and vice versa.

Fuente: Base de datos de la tienda departamental

Elaborado por: Aguirre, Gallo y Valencia

CAPÍTULO 4

CONCLUSIONES Y RECOMENDACIONES

4.1 Conclusiones

Basados en los resultados de los análisis realizados en los capítulos dos y tres, se llega a las siguientes conclusiones:

1) El árbol con menor riesgo de clasificación de segmento tiene un 23% de margen de error.

2) De un total de 119,551 (Muestra de Entrenamiento) clientes el 37.07% son clasificados como clientes buenos.

3) La primera variable discriminante fue región (costa / sierra) con lo cual se determina dos mercados diferentes para cada la ciudad de Guayaquil y Quito con una tasa de clientes buenos del 44.5% y 27.1% respectivamente.

4) Existe una diferencia denotada entre los clientes buenos femeninos y masculinos para cada región. En la costa Femeninos 48.3% y Masculinos 36.7%, mientras que en la Sierra Femeninos 30% y Masculinos 21.6%

5) Los mejores perfiles de cliente fiel por ciudad son los que poseen las siguientes características

   Para Guayaquil (Región Costa):

   Femenino, MesesCompra2008 > 5 meses, PromedioCompraCliente > $ 33

   Masculino, Ventas2008 > $ 499, MesesCompra > 16 meses

   Para Quito (Región Sierra):

   Femenino, Ventas2008 > $ 464

   Masculino, Ventas2008 > $ 472, MesesCompra > 14 meses

6) El perfil que obtuvo la mayor ganancia involucra al género femenino en ambas regiones; para la costa obtuvo una ganancia del 16.3% y para la sierra el 8.8%

7) Se analizaron un total de 171,177 registros, de los cuales el 8.04% son clientes fieles.

4.2 Recomendaciones

1) Se requiere la actualización del cubo de información con fecha de corte al día anterior a la modelación de los arboles de decisión.

2) Ningún departamento necesitará mas registros de clientes buenos o malos que el máximo número de clientes fieles sea determinado mediante árboles de decisión.

3) Es muy importante que se confié en el modelo para garantizar que este será utilizado apropiadamente.

BIBLIOGRAFÍA

1. Alarcon, D. (Abril de 2009). Los Cuatro Puntos Fantásticos En La Fidelización De Clientes.
2. Amaya Amaya, J. (1998). *Tomas de Decisiones Gerenciales - Métodos Cuantitativos.* Bucaramanga, Colombia: Universidad Santo Tomas.
3. Anderson, D. R. (2004). *Metodos Cuanticos para Negocios.* Thomson.
4. Bustamante, J., Rodriguez J., Echeverría F. (2009). *Data Warehousing y su aplicación como apoyo a la toma de decisiones en la ESPOL.* Obtenido de http://www.dspace.espol.edu.ec/bitstream/123456789/18/1/1.pdf
5. *DAEDALUS - DATA, Decisions and Language, S.A.* (1998). Obtenido de http://www.daedalus.es/mineria-de-datos/proceso-de-mineria-de-datos/
6. *DYBOX Information Technology*. (2000). Obtenido de Cubos: http://www.dybox.cl/pdf/cubos.PDF
7. *ETL-Tools.Info*. (2006). Obtenido de Business Intelligence - Almacenes de Datos - ETL: http://etl-tools.info/es/bi.htm
8. *ETSII.* (s.f.). Obtenido de Escuela Superior de Ingeniería Informática: http://jair.lab.fi.uva.es/~pablfue/leng_simulacion/materiales/v_v_0405.pdf

9. Hernández Orallo J., Ramírez Quintana M. y Ferri Ramírez C. (2004). *Introducción a la Minería de Datos.* Madrid, España: Pearson.
10. Lopez Gonzalez, C. (2008). *Mi Tecnologico.* Obtenido de www.MiTecnologico.com
11. (1949). El Jefe y la Jerarquía. En M. Matteucci Zatinni, *"Las Unidades de Abastecimiento Mundial".* (pág. 126). Buenos Aires.
12. Miranda, J. (2007). *Introduccion a la Minería de Datos.* Obtenido de Departamento de Ingeniería Industrial, Universidad de Chile: https://www.u-cursos.cl/ingenieria/2007/1/IN60E/1/material_docente/objeto/125935
13. Portocarrero, L. A. (s.f.). *CLASIFICACIÓN USANDO ÁRBOLES DE DECISION* . UNIVERSIDAD DE PUERTO RICO RECINTO DE MAYAGUEZ.
14. Puerta Goicoechea, A. (2002). *Imputación Basada en Arboles de Clasificacion.* Obtenido de http://www.eustat.es/document/
15. Schiattino Lemus, I., & Silva Zamora, C. (2008). Árboles de Clasificación y Regresión: Modelos Cart. *Ciencia y Trabajo , 10* (30), 161 - 166.
16. *SENCAMER.* (s.f.). Obtenido de http://portal.sencamer.gob.ve
17. *SINNEXUS Business Intelligence + Informática estratégica*. (2007). Obtenido de http://www.sinnexus.com/business_intelligence/datawarehouse.aspx
18. SPSS. (2001). *Guía del usuario de AnswerTree 3.0.* Chicago.

19. Tamayo, M. y Moreno, F. (2006). Análisis del Modelo de Almacenamiento. (U. N. Colombia, Ed.) *Revista Ingeniería e Investigación* , 135-142.
20. *Universidad de Jaén.* (s.f.). Obtenido de Escuela Superior Politecnica: http://wwwdi.ujaen.es/asignaturas/computacionestadistica/pdfs/tema6.pdf
21. Velasco, R. H. (2004). *Almacenes de datos (Datawarehouse)*. Obtenido de http://www.rhernando.net