INTERSEDES
REVISTA ELECTRÓNICA DE LAS SEDES REGIONALES DE LA UNIVERSIDAD DE COSTA RICA

Cortés Amarillo. Oleo de Norma Varela

Modelo de referencia CHISP-DM para el desarrollo de proyectos de minería de datos y CHAID como técnica de análisis para obtener información en minería de datos

Rafael Martínez Villareal

WWW.INTERSEDES.UCR.AC.CR
Vol. XIII, N°25 (2012)
ISSN 2215-2458

Modelo de referencia CHISP-DM para el desarrollo de proyectos de minería de datos y CHAID como técnica de análisis para obtener información en minería de datos

Reference Model CHISP-DM for the development of mining projects and CHAID analysis as a technique for information on data mining

Rafael Martínez Villarreal [1]

Recibido: 16.05.12 **Aprobado: 09.07.12**

Resumen

La minería de datos ha resultado ser una herramienta muy valiosa para las empresas modernas, garantizando la maximización de los recursos y beneficios económicos, utilizando las diferentes técnicas para la obtención de información calificada para tales propósitos, sin embargo, el proceso de estandarización y sistematización de un proyecto de minería de datos es fundamental para llegar a su implementación.

El modelo de referencia CHISP-DM permite un seguimiento paso a paso del proyecto. Uno de los pasos más importantes en éste modelo, es determinar cuál es la mejor técnica para analizar la información.

En este documento se describe al CHISP-DM con todas sus fases, además se presenta una descripción del método CHAID, como técnica de análisis, perteneciente a los métodos basados en árboles de decisión.

Palabras claves

Minería de datos, CHISP-DM, CHAID, Árboles de decisión, comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación, presentación.

Abstract

The data mining has turned out to be a very valuable tool for the modern companies, guaranteeing the maximization of the resources and economic benefits, using the different techniques the obtain qualified information for such purposes, however, the process of standardization and systematization of a project of data mining is fundamental to arrive to its implementation.

The reference model CHISP-DM allows a pursuit step to step of the project. One of the most important steps in this model, is to determine which is the best technique to analyze the information.

In this document the CHISP-DM is described with all its phases, also a description of the method CHAID is presented, as analysis technique, belonging to the methods based on trees of decision.

Key words

Data Mining, Sparkle-DM, CHAID, decision trees, business understanding, data understanding, data preparation, modeling, evaluation, presentation.

[1] Docente e investigador de la Sede Guanacaste. Universidad de Costa Rica. Email: rafael.martinez@ucr.ac.cr

1. Introducción

La tecnología de minería de datos ha hecho posible que se obtenga información calificada con la capacidad de ser predecible a través de procesos que la descubren y cuantifican, para obtener las relaciones predecibles de los datos, que son obtenidas de grandes volúmenes de datos, con la capacidad de incrementar las ventas o servicios de las empresas u organizaciones, con la ayuda de los sistemas de bases de datos operacionales y otros datos externos es posible extraer y transformar los datos en verdadera información valiosa para el negocio [Mapes1]. Para llegar a poner en práctica esta tecnología son necesarias una serie de técnicas y procesos que deben de estar debidamente sistematizados y normalizados, con este propósito es que se deben establecer modelos estandarizados que permitan a las empresas tanto desarrolladoras como usuarias de los sistemas de minería de datos, llevar un control sobre las actividades y etapas de desarrollo de este tipo de sistemas. En este artículo se trata el modelo de desarrollo conocido como CHISP-DM (*CRoss-Industry Standard Process for Data Mining*), este es un modelo de referencia que provee las metodologías y procesos necesarios para culminar con éxito un proyecto de minería de datos, este puede ser usado por principiantes y expertos, esta metodología divide en seis fases básicas el ciclo de vida de este tipo de proyectos: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación, y presentación [Colin01].

Durante el proceso de desarrollo de un proyecto de minería de datos es necesario la utilización de una diversidad de técnicas de análisis de datos que ayudan en el tratamiento y obtención de la información. En el presente artículo se tratará la técnica conocida como CHAID(*Chi Squared Automatic Interaction Detector*) este es un método basado en árboles de decisión, que consiste en un análisis que genera un árbol de decisión para predecir el comportamiento de una variable, a partir de una o más variables predictoras [Mapes1], además, existen otras técnicas utilizadas según la naturaleza del negocio y el tipo de análisis que se realice a los datos en el proyecto de minería de datos: Análisis estadísticos(ANOVA, Regresión, Ji cuadrado, Componentes principales, Análisis Cluster, Análisis discriminante), Métodos basados en árboles de decisión(CHAID), Algoritmos genéricos, Redes neuronales, Lógica difusa y Series temporales. A continuación se describen con mayor detalle el modelo CHISP-DM y la técnica CHAID.

Para la obtención de la información en este articulo se utilizará las referencias de diferentes autores y principalmente la información suministrada por las empresas que se dedican a la actividad comercial de la minería de datos, difundida en la Internet.

2.CHISP-DM Modelo de Referencia.

2.1 Antecedentes

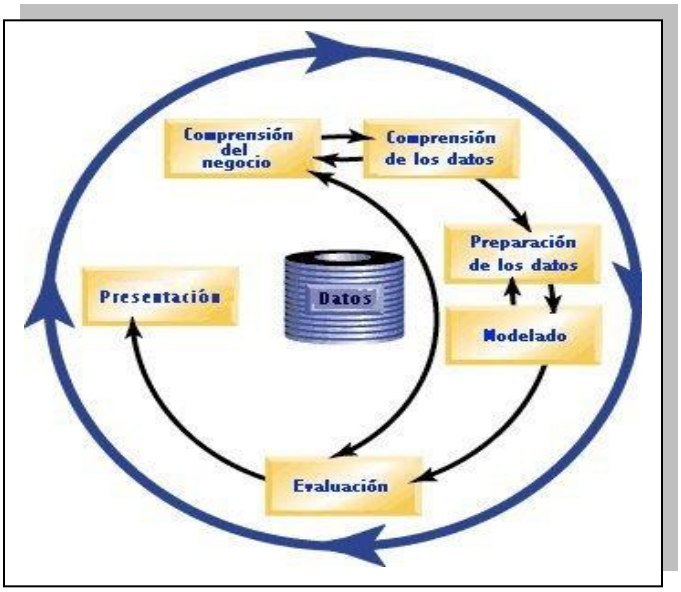
CHISP-DM es un modelo concebido al final del año 1996 por tres grandes empresas veteranas en la que era la nueva e inmadura tecnología de minería de datos: DaimlerChrysler (entoces Daimler-Benz), SPSS (entonces ISL) y NSR, más tarde se unió AHRA, para formar lo que se llamó SIG, por sus siglas en inglés, Grupo de Interés Especial, este se consolidó a mediados de 1999, desarrollando un documento de referencia llamado CHISP-MD versión 1, con base en la experiencia y los errores que se debían de solucionar en el desarrollo de un proyecto de minería de datos [Crisp-dm1], sin embargo la ACM(*Association for Computing Machinery*) organiza cada año lo que han denominado KDD(*Knowledge Discovery in Data and Data Mining*), que se refiere a ciclos de conferencias y discusiones sobre temas de datos y propiamente minería de datos, que ayudan a una retroalimentación de parte de las empresas y usuarios [ACMSIGKDD].

2.2 El Concepto de CHISP-DM

Este es una metodología y un proceso para el desarrollo de un modelo utilizado en la implementación de proyectos de minería de datos. Puede ser utilizado por principiantes y expertos, ofrece una estandarización para los proyectos, permite llevar un seguimiento paso a paso de las directrices, las tareas y objetivos de cada etapa del proyecto. Como norma se divide en seis fases básicas el ciclo de vida del proyecto: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación, y presentación [Spss01].

En la siguiente figura se representan las fases y la interrelación entre ellas.

Figura 1. Fases del modelo de referencia CRISP-DM [Colin01]



Las flechas indican las frecuencias y dependencias más importantes entre las fases, el círculo exterior simboliza la naturaleza cíclica de la minería de datos [Colin01]. La descripción detallada de cada una de las etapas se presenta seguidamente.

2.3. Fase uno, Comprensión del negocio

El entender el negocio, quizás es la fase más importante, en ella se debe determinar los objetivos, desde la perspectiva del negocio, las metas propuestas y sobre todo la definición de los problemas de minería de datos, con esto de antemano es posible diseñar un plan preliminar de desarrollo, estos pasos se pueden subdividir en la siguiente serie de actividades o tareas:

2.3.1 Determinar los objetivos del negocio

El comprender cuáles son las verdaderas metas que persigue el cliente, es un factor crítico para determinar, qué se debe involucrar en el plan de proyecto, es necesario plantearse algunas preguntas que ayuden a hacer un análisis más detallado de los objetivos principales del negocio y así determinar con claridad como ayudaría la minería de datos a la solución de los problemas actuales.

2.3.2 Evaluar la situación

Un paso importante es hacer una evaluación de la situación, que involucre un análisis de los recursos necesarios como hardware, software, personal disponible, fuentes de datos. También se deben establecer todas las condiciones bajo las cuales se dará el proyecto, en este punto se deberá elaborar una lista de los posibles riesgos potenciales, así como de las soluciones a los mismos. Con estos elementos identificados es necesario realizar un análisis costo – beneficio que indique que resultará del proyecto.

2.3.3 Determinar las metas de la minería de datos

Los objetivos determinan cuáles son las metas que se persiguen con la implementación del proyecto de minería de datos con respecto al negocio, el determinar las metas permiten conocer con qué exactitud se debe de predecir la información y cuál es el nivel de éxito que se obtendrá.

2.3.4 Producción del plan del proyecto

El plan del proyecto describe detalladamente lo que se quiere alcanzar según las metas de la minería de datos, involucra un análisis de los riesgos, herramientas y técnicas necesarias para apoyar el proyecto. El plan del proyecto debe de tomar muy en cuenta el tiempo de desarrollo. Las normas internacionales en minería de datos sugieren que un 50 a 70 por ciento es invertido en la fase de preparación de los datos, 20 a 30 por ciento se invierte en la fase de la comprensión de los

datos, solamente el 10 a 20 por ciento en el modelado, evaluación y la fase de compresión del negocio, la fase de planificación de la presentación necesita de un 5 a 10 por ciento [Colin01].

2.4 Fase dos, Comprensión de los datos

La comprensión de los datos debe iniciar con un conjunto de datos, el analista debe familiarizarse con los mismos; de esta forma podrá identificar los posibles problemas que se presenten con estos. Esta etapa o fase, involucra cuatro pasos: la colección de datos iniciales, la descripción de datos, la exploración de datos y la comprobación de calidad de los datos, a continuación se describe cada una de ellas.

2.4.1 Colección de datos

El analista debe cargar e integrar los datos necesarios y determinar las fuentes de origen, debe de considerar que este proceso puede volverse muy extenso, pues se deben recuperar y en algunos casos transformar los datos.

2.4.2 Descripción de los datos

En éste punto el analista realiza un estudio de la naturaleza de los datos, aspectos como el número de tablas, la cantidad de campos por tabla, el dominio de cada campo y cualquier otro aspecto de los campos, aquí se debe determinar si se tienen los datos necesarios para cumplir con los requerimientos del proyecto.

2.4.3 Explorar los datos

En esta tarea es donde se sustenta la minería de datos, aquí se decide cuales son las preguntas a resolver, las cuales pueden deducirse por medio de una hipótesis, explorando la generación gráfica de la información o simplemente interpretando reportes.

2.4.4 Calidad de los datos

El analista verifica la calidad de los datos, en este punto se debe de examinar si los datos obtenidos son suficientes, si tienen el valor adecuado, si hay campos nulos, si hay campos repetidos con nombres diferentes o si tiene el mismo nombre pero con valores diferentes, se deben identificar estos posibles conflictos y dar soluciones utilizando el sentido común [Colin01].

2.5 Fase tres: Preparación de los datos

Aquí se deben de desarrollar todas las actividades necesarias para construir los datos que serán colocados en la(s) herramienta(s) de modelado, se debe de trabajar en la selección de tablas, registros y atributos, así como en la limpieza y transformación de los datos con la ayuda de

herramientas de modelado. Esta fase se divide en cinco pasos: la limpieza de datos, la construcción de datos, la integración de datos y el dar formato a los datos.

2.5.1 Selección de datos

Este es un paso decisivo pues aquí se decide cuales datos deben ser analizados y los criterios pueden ser varios, inclusive tener presente las metas por las cuales se realizará minería de datos en la empresa, la calidad de los datos, las restricciones técnicas como el volumen máximo y tipos de datos permitidos.

2.5.2 Limpieza de los datos

Si no se limpian los datos, el hacer minería de datos no pasaría de ser un gran número de consultas sin llegar a obtener resultados, aquí se implementan técnicas bastante ambiciosas que permitan estimar datos desconocidos. Para realizar la limpieza se debe de asegurar la verificación de la calidad de los datos (punto 2.4.4).

2.5.2 Construcción de los datos

Pasado el proceso de limpieza de datos, el analista debe de construir nuevos datos derivándolos de los ya obtenidos, aquí se crean nuevas tablas y nuevos registros, en ocasiones es necesario transformar valores simbólicos a numéricos para evitar ambigüedades en los datos, otra práctica es obtener un dato a partir de otros.

2.5.3 Integración de los datos

Este es un proceso donde se integran tablas o valores nuevos a partir muchas de las tablas o valores obtenidos llamadas tablas bases o valores base, se construyen nuevas tablas con información relacionada de diferentes tablas o diferentes bases de datos, estas tablas o valores nuevos pueden contener agregados (valores calculados) provenientes de otros valores o tablas diferentes.

2.5.4 Formateo de los datos

El analista debe de dar el formato adecuado a los datos para su procesamiento, esto debido a que es necesario adaptarlos a la herramienta de modelado utilizada. En muchos casos en este proceso los cambios en los datos se pueden volver muy complejos y en ocasiones resultan muy sencillos.

2.6 Fase cuatro, Modelado

En esta fase se seleccionan las diferentes técnicas de modelado que se ofrecen para la minería de datos, algunas de estas técnicas exigen requerimientos específicos en cuanto a la forma de los datos, por ello es necesario adaptarlos, en ocasiones es necesario retroceder a la fase de preparación de los datos (fase anterior). Se necesitan 4 pasos: seleccionar la técnica de modelado, elaborar un plan de prueba, construir el modelo y evaluar el modelo.

2.6.1 Selección de la técnica de modelado

La selección de las técnicas de modelado de datos específica para el proyecto, es una tarea que involucra un poco de conocimiento de parte del analista, las opciones más comunes son: redes neuronales artificiales, árboles de decisión entre ellas (CART, CHAID, esta última será tratada en detalle más adelante en este artículo), algoritmos genéricos, método del vecino más cercano y reglas de inducción [Presser].

2.6.2 Elaboración del plan de prueba

Cuando ya se tiene el modelo de datos a usar, es necesario realizar algunas pruebas para comprobar la calidad y la validez del modelo, comprobando empíricamente si el modelo puede predecir los datos en función de los datos históricos y además que cumpla con las expectativas en la empresa.

2.6.3 Construcción del modelo

Después de realizadas las pruebas, el analista debe de "correr" el modelo en la herramienta modeladora a utilizar, generando uno o más modelos.

2.6.4 Evaluación del modelo

En este punto el analista en minería de datos utiliza su conocimiento técnico así como su experiencia teniendo en cuenta el resultado de las pruebas de los datos y el modelo, junto con los expertos en el negocio, utiliza su criterio para determinar si el modelo cumple con las expectativas planteadas para dar solución a los problemas que soluciona la minería de datos en el negocio.

2.7 Fase cinco, Evaluación

El proceso de evaluación permite al analista dar una revisión al proceso de modelado, aquí debe de determinar y estar seguro si se cumplieron los objetivos fijados, básicamente esta es una fase de retroalimentación, la cual debe considerar: el evaluar los resultados, repasar y determinar los próximos pasos.

2.7.1 Evaluación de resultados

Las evaluaciones anteriores permiten tener claro la exactitud del modelo en términos generales. En este paso el analista debe de evaluar el modelo de acuerdo a la aplicación en el mundo real, aquí determina aspectos reales como las limitaciones de presupuesto, además puede determinar cuanto se puede proyectar el modelo en el futuro y cuáles son los posibles cambios que pueden experimentar los datos en adelante.

2.7.2 Proceso de repaso

El proceso de repaso permite al analista repasar los anteriores proceso y determinar si se pasó por alto un proceso o tarea importante para realizar minería de datos en el negocio, permite también revisar la calidad con la que se han realizado los anteriores procesos o pasos.

2.7.3 Determinar los próximos pasos

Este es el último paso de la evaluación, en este punto el director de proyecto tiene la suficiente información para decidir dar fin al proyecto y continuar con el proceso de presentación o si es necesario una iteración entre los anteriores pasos, también puede decidir empezar con un nuevo proyecto de minería de datos.

2.8 Fase seis, La presentación

La tarea de crear el modelo no es la última que se realiza en un proyecto de minería de datos, todo el conocimiento adquirido y la información debe de ser organizado y presentado de forma que el cliente la entienda y la utilice para la toma de decisiones. La presentación del proyecto puede ser un proceso tan sencillo como generar reportes, o tan complejo, que permita generar información para retroalimentar el proceso de minería de datos dentro de la empresa, por supuesto, la complejidad depende de los requerimientos y objetivos planteados en el proyecto. En esta fase es importante saber que los usuarios o clientes del proyecto debe entender cuáles son y cómo se obtienen los datos para dar con el modelo, aquí son necesarios los siguientes pasos: plan de la presentación, plan de supervisión y mantenimiento, elaboración del reporte final y repaso del proyecto.

2.8.1 Plan de la presentación

Para presentar los resultados del proyecto de minería de datos hay que evaluar primero los resultados y seguidamente establecer una estrategia para determinar cómo va a ser presentado.

2.8.2 Plan de supervisión y mantenimiento

El desarrollo de un plan para supervisar y dar el mantenimiento adecuado al proyecto de minería de datos evita en gran medida el uso incorrecto de los resultados de la minería de datos en la empresa.

2.8.3 Elaboración del reporte final

En este punto el director de proyecto en conjunto con su equipo de trabajo deben elaborar un reporte final, en donde incluyen un resumen de los resultados del proyecto, haciéndose mención de la experiencia obtenida, una exposición conclusiva y comprensiva del proyecto de minería de datos en forma verbal. Para realizar dicha exposición se organiza una reunión final con el cliente.

2.8.4 Revisión del proyecto

La revisión del proyecto incluye entrevistas con los participantes más importantes del proyecto, aquí se debe de documentar los fracasos y los éxitos que se dieron en el desarrollo del mismo, así como las recomendaciones para el desarrollo de proyectos futuros, en el documento debe ir escrito el informe de cada uno de los miembros del equipo de desarrollo, por cada una de las etapas y tareas realizadas [Colin01]. Con la fase revisión del proyecto, se concluye la exposición de una visión general de cada una de las fases y sus respectivos pasos del modelo de referencia CRISP-DM, a continuación se expondrá la técnica de análisis y predicción de información conocida como CHAID para brindar una muestra de cómo se obtiene la información.

3. CHAID (*Chi Squared Automatic Interaction Detector*)

3.1 Generalidades

El fin global de la minería de datos es encontrar información predecible procedente de grandes volúmenes de datos, para conseguir esto no solo basta con tener almacenados y organizados los datos o tener hardware con un alto desempeño en procesamiento, hacen falta las técnicas y métodos algorítmicos capaces de procesar los datos y hacer proyecciones validas de información, que ayuden a la empresa a predecir el comportamiento futuro sobre sus servicios y permitirles a partir de los resultados obtenidos, tomar sus decisiones.

Dentro de estas técnicas existe gran diversidad de tipos, utilizadas según la naturaleza del negocio y el tratamiento de los datos, entre ellos se establecen las siguientes categorías:

Análisis estadísticos (ANOVA, Regresión, Ji cuadrado, Componentes principales, Análisis Cluster, Análisis discriminante), **Métodos basados en árboles de decisión**(Árboles de Clasificación y Regresión CART, "*Classification And Regression Tree"*) y Detección de Interacción Automática de Chi Cuadrado CHAID, "*Chi Square Automatic Interaction Detection*), **Algoritmos genéricos**, **Redes neuronales**, **Lógica difusa** y **Series temporales**[Presser], a continuación se describe la técnica CHAID para ilustrar un poco la implementación de las mismas.

3.2 Funcionamiento de CHAID

El CHAID es un análisis que genera un árbol de decisión utilizado para predecir cómo se debe comportar una variable a partir de una o varias variables predictoras, formándose un árbol en el que los conjuntos de una misma rama y mismo nivel son disyuntivos. Esta técnica es utilizada cuando se requiere dividir una población en varios segmentos, teniendo claro un criterio de decisión.

El árbol se construye particionando los datos en dos o más subconjuntos de observación a partir de los valores que toman las variables predictoras, a cada subconjunto se le aplican nuevamente el algoritmo hasta que no existan diferencias significativas en la influencia de las variables de predicción de uno de estos grupos hacia el valor de la variable de respuesta.

El árbol entonces está constituido en su raíz por los datos íntegros y sus ramas por los conjuntos y subconjuntos. En una partición pueden existir dos o más subconjuntos que están determinados por el número de los distintos valores que toma la variable utilizada para realizar la segmentación, la variable de predicción utilizada para crear una partición, es la que tiene mayor relación significativa con la variable de respuesta, en concordancia con la prueba de independencia de la chi cuadrado sobre una tabla de contingencia.

Esta técnica resulta apropiada para predecir información con variables predictoras, permitiéndose realizar cálculos complejos y además en la mayoría de sus implementaciones generan el árbol de forma gráfica, ayudando a una interpretación rápida y fácil de los resultados, también se ofrece la forma tabular de los resultados [SmartDrill].

Para ilustrar la técnica se expondrá un ejemplo con propósitos ilustrativos, considere el diagrama de la figura 2, en el cual se presenta un árbol, en su raíz presenta la muestra entera de los datos a analizar, que precisamente son 81.040 casas en las que se vendió vía correo, en el mismo cuadro que simboliza la raíz se incluye la ganancia promedio por casa que es de $0. 75. El número de casas es identificado por CHAID como el mejor predictor con el cual se debe de empezar a segmentar el mercado. En el diagrama del árbol se puede ver que cuando las casas son de 2 a 4

personas, se obtiene una ganancia promedio de $1.64, que es dos veces la ganancia generada por una casa de una persona($0.82) y casi siete veces la ganancia generada por una casa de 5 a 6 personas($0.24 * 7 = $1.68).

Entonces CHAID presenta en el segmento de 2 a 4 personas por casa, si se cuenta con una tarjeta de un banco, las ganancias suben a $3.58, si no tienen tarjeta solamente se tiene un porcentaje de ventas de $1.29. Sin embargo si dentro de este segmento, la cabeza de familia trabaja en una oficina, la rentabilidad sube a $2.25.

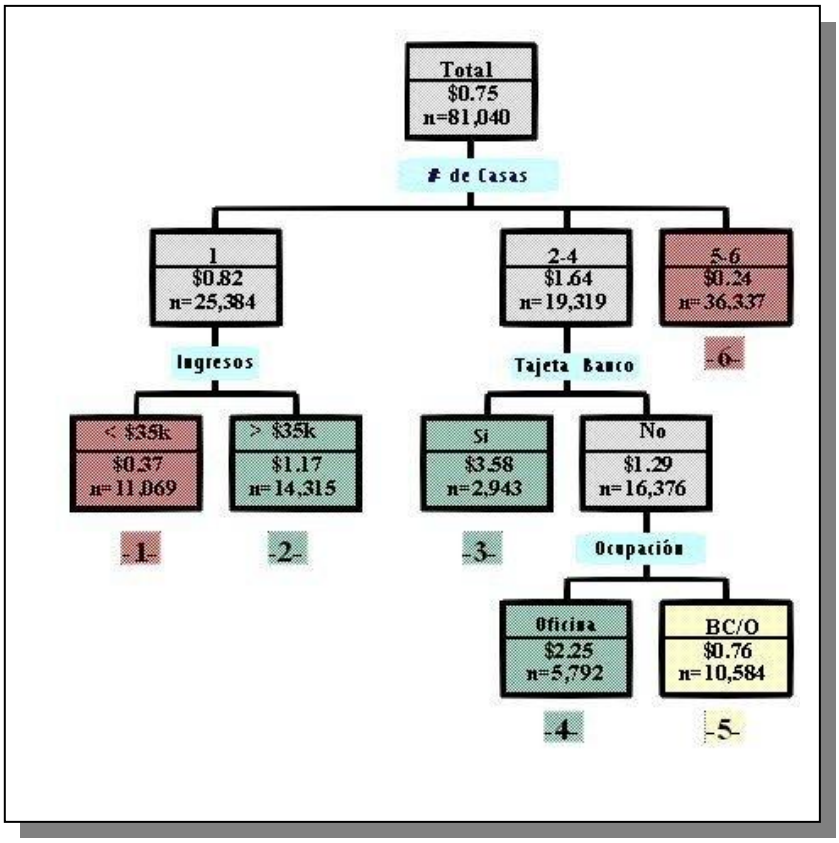


Figura 2. Árbol de decisión CHAID [SmartDrill]

Con propósitos ilustrativos en la tabla 1 (Tabla de ganancias) y el diagrama de árbol, los segmentos son coloreados de forma diferente: promedio de segmento en verde, promedio de segmento en amarillo y el último promedio de segmento en rojo. De igual manera se numeraron los segmentos del 1 al 6, además en la tabla1, se representa otra información adicional considerada importante, para todos los segmentos.

La primera columna representa el número de identificación del segmento, la segunda presenta el número de casas por segmento, en la tercera se muestra el porcentaje del total por segmento de casas(2943*100/81040 = 3.6), para el segmento 3, la columna cuatro presenta el porcentaje de ganancias por casa en cada segmento, la columna cinco presenta las ganancias como un índice relativo para el porcentaje del modelo entero, puesto a 100 ($0.76 * 100 / $0.75 = 101), para el segmento 5, con esta medida se genera una ganancia de 4.76 veces el promedio del modelo en la muestra total(476 * $0.75 = 3.58), y más de 15 veces la rentabilidad del peor segmento(476 / 31 = 15), así se puede determinar que el segmento que presenta el mejor índice es el número 3, en las columnas de la 6 a la 9 se presentan datos acumulativos por segmento, de las columnas de la 2 a 5, otra información que presenta la tabla de ganancias son los mejores segmentos, el 3, 4 y 2 representando el 28.4% ((2943+5792+14315) * 100 / 81040 = 28.4) de la totalidad del ejemplo, teniendo un porcentaje de ganancias del $1.75 por casa(según CHAID), por consiguiente es 2.32 veces aprovechables que el porcentaje de la muestra del ejemplo(1.75 / 0,75 = 2.32).

CHAID es particularmente útil para generar modelo de segmentos de mercado, permitiendo predecir a las empresas cuales segmentos del mercado deben dirigirse, según el cálculo de ganancias, es muy utilizado en los medios de comunicación. Los resultados pueden ser colocados fácilmente en una base de datos y ser procesados para la toma de decisiones [SmartDrill].

Segmen. ID	Casas/Segm.	Porc. del Total	Porc. $ por casa	Indice Segm.	Acumula. casa por segm.	Acumula. Porc. total	Acumu $ Porc. casas	Acum. Index Segm.
3	2,943	3.6	3.58	476	2,943	3.6	3.58	476
4	5,792	7.1	2.25	298	8,735	10.8	2.70	358
2	14,315	17.7	1.17	155	23,050	28.4	1.75	232
5	10,584	13.1	0.76	101	33,634	41.5	1.44	191
1	11,069	13.7	0.37	49	44,703	55.2	1.17	156
6	36,337	44.8	0.24	31	81,040	100.0	0.75	100

Tabla1. Tabla de Ganancias [SmartDrill].

4. Conclusión

El tener claro cuál es el camino para llevar a cabo un proyecto de minería de datos depende sin lugar a duda de la sistematización de todo el ciclo de vida, el modelo de referencia CRISP-DM permite dar un seguimiento ordenado a cada una de las fases y tareas que involucran el desarrollo del proyecto, dentro de las fases que deben de realizarse figuran: Comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y presentación del proyecto. Es necesario tener presente que todo el modelo siempre debe de ir en concordancia con los objetivos de la empresa y llevar un control sobre el cumplimiento de los requisitos establecidos. Dentro de la fase cuatro "Modelado", es necesario escoger la técnica de análisis, que más se adapte a las necesidades y naturaleza del proyecto. Hoy en día existe un gran número de técnicas de análisis de datos, entre ellas se mencionan: Análisis estadísticos, Métodos basados en árboles de decisión (Árboles de Clasificación y Regresión CART, "*Classification And Regression Tree"*) y Detección de Interacción Automática de Chi Cuadrado CHAID, "*Chi Square Automatic Interaction Detection*), Algoritmos genéricos, Redes neuronales, Lógica difusa y Series temporales. En particular el objetivo global de la minería de datos es poder realizar predicciones sobre el comportamiento de diversas variables, alimentándose con grandes cantidades de información actualizada y veraz, una técnica muy utilizada para esto es el CHAID, este es un método de análisis que genera un árbol de decisión utilizado para predecir cómo se debe comportar una variable a partir de una o varias variables predictoras, con esta información se debe segmentar el mercado, permitiéndose tomar decisiones, basándose en el retorno de la inversión o las mejores opciones de rentabilidad.

5. Bibliografía

[Mapes1] "***Data Mining* o Minería de Datos**". Home page del Ministerio de Administración Pública. URL: http:/www.map.es/csi/silice/DW2253.html. Madrid, España, 1999.

[Colin01] Colin Shearer. "***The CRISP-DM Model: The New Blueprint for Data Mining***" En *Journal of Data Warehousing*, Volumen 5, Número 4, EEUU, 2000.

[Spss01] "***What is CRIDP-DM***". Home Page Spss. URL: http:/www.spss.com/datamine/approach.html. EEUU. 2001.

[ACMSIGKDD] "***ACM Special Interest Group on Knowledge Discovery in Data and Data Mining***". Home page de ACM. URL: http://www.acm.org/sigkdd/learn.html. EEUU. 2001.

[Crisp-dm1] "***Crisp-dm Proyect Overview***". Home page de Crisp-dm.org. URL: http/www.crisp-dm.org/home.html.

[Presser] Presser Carne Cynthia "***Data Mining***". Home page Monografías.com. URL: http/www.monografías.com/trabajos/dataminig/datamining.html. 2001.

[SmartDrill] "***Analistc Tecnicques***". Home page SmartDrill.com. URL: http/www.smartdrill/process4.html. EEUU, 2001.