Universidad Austral de Chile

Facultad de Ciencias de la Ingeniería

Escuela de Ingeniería Civil en Informática

DISEÑO E IMPLEMENTACIÓN DE UN MODELO PREDICTIVO PARA DETECTAR PATRONES DE FUGA EN LOS SERVICIOS DE TELEFÓNICA DEL SUR

Tesis para optar al Título de:
Ingeniero Civil en Informática

Profesor Patrocinante:
Sr. José Loyola Abello
Analista de Sistemas
Ingeniero en Administración Empresas

Profesor Co-Patrocinante:
Sr. Wladimir Ríos Martínez
Ingeniero Ejecución Eléctrico,
Ingeniero Civil Industrial,
Magister en Administración de Empresas (MBA)
Doctor (C) en Telemática

Profesor Informante:
Sr. Juan Pablo Salazar Fernández
Ingeniero Civil en Informática
Magíster en Administración de Empresas

JOSUÉ MAXIMILIANO ALVARADO BUSTOS

VALDIVIA - CHILE

2011

Valdivia, 6 de Julio de 2011

De: José Loyola Abello
Patrocinante

A: Juan Pablo Salazar Fernández
Director
Escuela de Ingeniería Civil en Informática

Ref: Calificación proyecto de título

De mi consideración:

Habiendo revisado el trabajo de titulación **"Diseño e Implementación de un Modelo Predictivo para Detectar Patrones de Fuga en los Servicios de Telefónica del Sur"**, presentado por el alumno Sr. Josué Maximiliano Alvarado Bustos, mi evaluación del mismo es la siguiente:

Nota: **6.8** (Seis coma ocho).

Fundamento de la nota:

El alumno se desempeño en forma eficiente y eficaz en las reuniones realizadas con los usuarios claves, liderando éstas de acuerdo a lo esperado. En relación a la búsqueda de la información y data relevante para las métricas solicitadas fue superior a lo esperado aportando sus conocimientos para obtener las variables adecuadas y eficientar el modelo. En relación a los resultados obtenidos a través del modelo, se hace evidente que el aporte del proyecto impacta fuertemente en la precisión para detectar en forma temprana la fuga de los clientes y tomar acciones tendientes a revertir tal situación focalizando la atención en el conjunto bajo predicción.

Aspecto	**Evaluación**
Cumplimiento de objetivos	7.0
Satisfacción de alguna necesidad	7.0
Aplicación de metodologías pertinentes	7.0
Interpretación de los datos y obtención de conclusiones	6.8
Originalidad	6.4
Aplicación de criterios de análisis y diseño	7.0
Perspectivas del trabajo	6.9
Coherencia y rigurosidad lógica	6.5
Precisión del lenguaje técnico	6.8

Sin otro particular, saluda atentamente a usted,

José Loyola Abello
Data Warehouse e Informes Gestión
TELEFÓNICA DEL SUR

Valdivia, 11 agosto de 2011

De: Wladimir Rios Martinez, Co-patrocinante Proyecto Tesis

A: Juan Pablo Salazar Fernández, Director
Escuela de Ingeniería Civil en Informática

Ref: Calificación proyecto de título

De mi consideración:

Habiendo revisado el trabajo de titulación "**Diseño e Implementación de un Modelo Predictivo para Detectar Patrones de Fuga en los Servicios de Telefónica del Sur**", presentado por el alumn Sr. **JOSUÉ MAXIMILIANO ALVARADO BUSTOS**, mi evaluación del mismo es la siguiente:

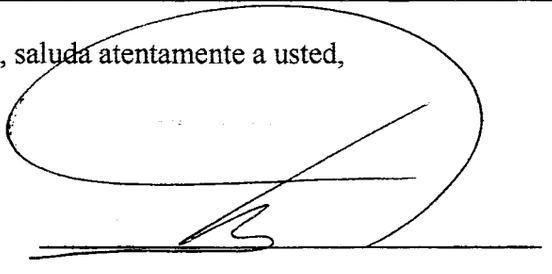
Nota: **7,0** Siete coma cero

Fundamento de la nota:

Tesis ha sido desarrollada sobre una base tecnológica existente en la Compañía y tomado en cuenta esas restricciones, el desarrollo está hecho con un alto nivel técnico.
El desarrollo de este proyecto da lugar a muchos otros posibles proyectos en esta misma línea y en cada salto tecnológico en que invierta la Compañía.
Proyecto ha sido elaborado en forma muy dedicada y cuidando los detalles

Aspecto	Evaluación
Cumplimiento de objetivos	7.0
Satisfacción de alguna necesidad	7.0
Aplicación de metodologías pertinentes	7.0
Interpretación de los datos y obtención de conclusiones	7.0
Originalidad	7.0
Aplicación de criterios de análisis y diseño	7.0
Perspectivas del trabajo	7.0
Coherencia y rigurosidad lógica	7.0
Precisión del lenguaje técnico	7.0

Sin otro particular, saluda atentamente a usted,



Wladimir Rios Martinez
Académico
Instituto de Informática, UACH

Valdivia, 04 de octubre de 2011

De: Juan Pablo Salazar Fernández
Profesor Informante

Ref: Calificación proyecto de título

De mi consideración:

Habiendo revisado el trabajo de titulación "**Diseño e Implementación de un Modelo Predictivo para Detectar Patrones de Fuga en los Servicios de Telefónica del Sur**", presentado por el estudiante sr. **Josué Maximiliano Alvarado Alarcón**, mi evaluación del mismo es la siguiente:

Nota: **6,4** (seis, coma cuatro).

Fundamento de la nota:

Aspecto	**Evaluación**
Cumplimiento de objetivos	6,2
Satisfacción de alguna necesidad	7,0
Aplicación de metodologías adecuadas	6,5
Interpretación de los datos y obtención de conclusiones	6,0
Originalidad	6,5
Aplicación de criterios de análisis y diseño	6,0
Perspectivas del trabajo	6,5
Coherencia y rigurosidad lógica	6,5
Precisión del lenguaje técnico	6,5

Sin otro particular, saluda atentamente a usted,

Juan Pablo Salazar Fernández
Profesor Auxiliar
Instituto de Informática
Universidad Austral de Chile

AGRADECIMIENTOS

Quiero agradecer a todas las personas que hicieron posible de una u otra forma la realización de este trabajo de tesis, en especial a Carolina Cárcamo Lange por su preocupación, amor y compañía.

De manera especial agradezco a mis padres, y hermanos por el apoyo incondicional que me dieron a lo largo de esta carrera.

Esta tesis está dedicada a mis padres con mucho cariño.

ÍNDICE

ÍNDICE DE FIGURAS

Figura **Página**

ÍNDICE DE TABLAS

Tabla **Página**

RESUMEN

Telefónica del Sur es una empresa de telecomunicaciones que cuenta con una gran cantidad de información concerniente a los comportamientos y factores sociodemográficos de sus clientes. Esta se encuentra almacenada en un Data Warehouse Corporativo, siendo una fuente de datos propicia para ser analizada.

Este proyecto consiste en elaborar un modelo predictivo que permita descubrir de forma anticipada, el comportamiento normal que manifiestan los clientes cuando consideran la opción de rescindir sus servicios contratados, posiblemente buscando una mejor oferta en la competencia. Este proceso fue realizado mediante la aplicación de técnicas de Minería de Datos, que consiste en la extracción de forma automática de información relevante, útil y no evidente contenida en los datos.

La información utilizada para la realización del proyecto considera todos los datos históricos del año dos mil nueve. Para el descubrimiento de patrones y tendencias que existen en esta información, se emplearon algoritmos de clasificación como, redes neuronales, árboles de decisión y redes bayesianas. La utilización de estos algoritmos en conjunto, fue el mecanismo por el cual se creó el modelo de minería de datos.

El proceso fue desarrollado por medio del software de gestión de bases de datos relacionales SQL Server 2005, quien ofrece un entorno integrado para crear y trabajar con modelos de minería de datos denominado Business Intelligence Development Studio. El entorno incluye algoritmos y herramientas de minería de datos que facilitan la generación de una solución completa para diversos proyectos.

El efecto de este proyecto es la maximización de la rentabilidad de la empresa, mediante la detección precoz de las posibles fugas, con el objetivo de realizar ofertas y mejoras personalizadas a los servicios prestados, aumentando la tasa de retención de clientes. Además se espera un acrecentamiento en los niveles de satisfacción, producto de la eficiente relación que existirá entre la empresa y el cliente.

ABSTRACT

Telefónica del Sur is an telecommunications company that counts with a large amount of information about the sociodemographic, factors and behavior of their customers. This information is stored in a Corporative Data Warehouse that is a proper data source to be analyzed.

This Project consists in develop a predictive model that allows to anticipate the normal behavior that customers manifest when they consider the option of rescind the contracted services, looking for a better offer in other companies. This process was developed by the application of data mining techniques that consist in the extraction in an automatic way of the relevant, useful and not evident information present in the data.
The information used to develop of this project considers all the historic data of the year two thousand nine. The discovering of the patterns and trends presents in the information was made using classification algorithms such as neural networks, decisions trees and Bayesian networks. The use of this set of algorithms was the mechanism used to create the data mining model.

This process was developed using the relational database management software SQL Server 2005, that offers an integrated environment to create and develop data mining models called Business Intelligence Development Studio. The environment includes algorithms and data mining tools that makes easier the generation of a complete solution for diverse projects.

The effect of this project is the maximization of the company profitability by the early detection of possible leaks, with the objective of make offers or personalized improvements to the contracted services, to increase the customer retention rate. Also an increase of the satisfaction levels is expected, result of the efficient relationship between the company and the customer.

1 INTRODUCCIÓN

1.1 Información de índole general sobre el problema a tratar

La alta competitividad que se da en las empresas de telecomunicaciones genera una maquinaria gigantesca en términos publicitarios y de marketing, lo que redunda en que el costo de adquisición de los clientes sea relevante y como consecuencia de ello, la rotación de los clientes se transforma en un hecho fundamental dentro de la industria. Es sabido que el costo de adquirir un cliente nuevo es entre cinco y quince veces mayor que el costo de mantenerlo, por ello es de vital importancia la incorporación de nuevas tecnologías que faciliten la generación del conocimiento en base al tratamiento y análisis de la información.

En la actualidad Chile está viviendo un proceso de convergencia tecnológica semejante al que viven otros países desarrollados. El nivel de penetración de la telefonía móvil es alrededor de 90%, con más de 14 millones de abonados a nivel nacional. El incremento de las conexiones a internet entre el año 2005 y 2009 fue duplicada, llegando aproximadamente a los 2 millones de conexiones a nivel nacional. La TV de pago, por su parte, cuenta con 1,47 millones de suscripciones a nivel nacional. La telefonía fija no se ha quedado estancada, a pesar de que las líneas fijas a nivel nacional se han mantenido en 3,5 millones [DOM09].

En este escenario, donde el mercado de las telecomunicaciones tiende a alcanzar su mercado potencial, existe un creciente aumento de la rivalidad competitiva entre las empresas de telecomunicaciones [SUB], lo que podría provocar una disminución en la cantidad de consumidores nuevos a Telefónica del Sur. Por este motivo es momento de mirar con mayor interés el problema de abandono o fuga de clientes, denominado *Churn*[1], y centrar la mirada en las nuevas herramientas tecnológicas que ofrece el mercado.

En la actualidad una de las herramientas que permite potenciar el conocimiento del mercado y de los clientes es la Minería de Datos. Esta herramienta es pieza fundamental en la Business Intelligence, permite obtener información anticipada de los patrones de comportamiento que conducen a un abandono, dando la posibilidad de realizar

[1] Churn: Tasa de clientes que dan de baja un servicio.
URL: http://www.tecnologiahechapalabra.com/tecnologia/glosario_tecnico/articulo.asp?i=3370

campañas de retención apuntadas a aquellos clientes con riesgo real de fuga, focalizando los esfuerzos en esos clientes y no en la realización de campañas de retención masivas.

El objetivo central de este proceso fue potenciar la explotación de la información almacenada, mediante un uso más sofisticado de ella. Para lograrlo fue de vital importancia caracterizar a cada cliente fugado mensualmente, considerando las áreas de negocio de Internet Banda Ancha (ADSL), Telefonía Post Pago (TPOS), Televisión Digital Inalámbrica (WITV) y Telefonía Local Inalámbrica (PHS). Analizando cada una de forma separada, buscando descubrir los patrones que desencadenan la fuga de un servicio por parte del cliente.

La información que se utilizó para la consecución del objetivo planificado se clasificó en cuatro tipos de variables, quienes abarcan todos los ámbitos de información que se necesitan conocer del cliente, ellas son:

- ***Variables de Comportamiento:*** Focalizan la atención en indicadores que permiten identificar cambios en el comportamiento transaccional de los clientes.
- ***Variables Socio – Demográficas:*** Conocer aspectos generales del cliente.
- ***Variables de Entorno:*** Indicadores que permiten cuantificar los posibles efectos que ejerce el medio en el que se desenvuelve la cartera de clientes (información de inteligencia del mercado).
- ***Variables de Calidad de Servicio:*** Corresponde a indicadores que apuntan a ponderar la diferencia entre las expectativas de sus clientes y la percepción que en realidad tienen de la empresa.

1.2 Antecedentes existentes al respecto

Telefónica del Sur posee una gran cantidad de información almacenada en un Data Warehouse Corporativo, de esta fuente de datos se generan una serie de informes, reportes y documentos que permiten conocer cuál es el comportamiento y el rendimiento de la empresa en muchos aspectos. Entre ellos existen informes de gestión con información estática de estadísticas de ventas, bajas de clientes, abandono de clientes e ingresos medios por cliente (*Arpu*[2]), etc.

En cuanto a proyectos desarrollados con el fin de investigar las base de datos de Telefónica del Sur, buscando información escondida o patrones de comportamiento de sus clientes, existe un trabajo de titulación llamado, "Diseño de un modelo parametrizado, usando Data Mining, que permita a Telefónica del Sur segmentar por valor a los clientes de prepago". Este sistema permite entender con mayor precisión y profundidad las distintas variables relevantes que impactan en la venta de tarjetas de prepago, logrando realizar una mejor gestión de la cartera de clientes pertenecientes a este segmento de negocio [JAR05].

En la actualidad este sistema no se encuentra operativo debido a que es desarrollado exclusivamente para investigar el comportamiento de los clientes de telefonía prepago, segmento de negocio que ya no es prioridad en Telefónica del Sur.

Para el desarrollo de éste proyecto se utilizará la herramienta SQL Server 2005, en conjunto con la metodología de trabajo específica para minería de datos CRISP-DM[3].

CRISP-DM es un consorcio de empresas europeas y estadounidense creada a finales de 1996 por tres importantes actores en proyectos de minería de datos, que son SPSS, NCR y DaimlerChrysler [DAT00]. Esta metodología trata de desarrollar los proyectos de minería de datos bajo un proceso estandarizado de definición y validación, de tal forma que se desarrollen proyectos con un costo razonable y con un alto impacto en el negocio. Para ello ésta metodología consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos,

[2] Arpu: Ingresos medios por clientes.
URL: http://www.tecnologiahechapalabra.com/tecnologia/glosario_tecnico/articulo.asp?i=2578
[3] CRISP-DM: http://www.crisp-dm.org/

estructurando el ciclo de vida del proyecto en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto [CHA99].

Esta metodología fue fielmente desarrollada por medio de la herramienta Business Intelligence Development Studio, que incluye algoritmos y herramientas de minería de datos que facilitan la generación de una solución para el modelo predictivo. A través de la figura 1 se explica gráficamente el desarrollo que tendrá el proyecto en su construcción, siguiendo cada una de las fases de la metodología junto con el proceso que la desarrollará.

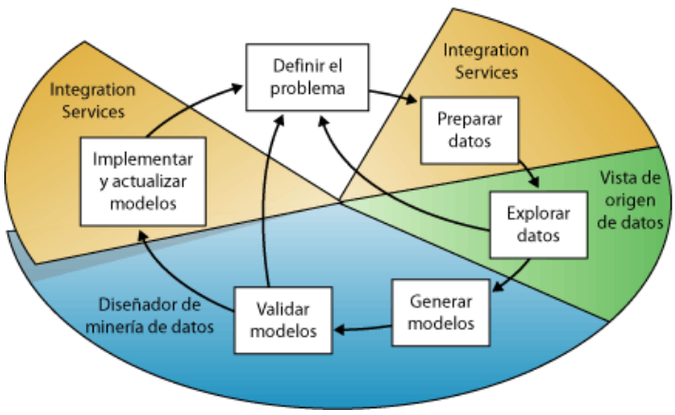


Figura 1: Relaciones entre cada fase del proceso y las tecnologías de Microsoft SQL Server 2005 [MDN08].

El algoritmo de minería de datos es el mecanismo por el cual se crearon los modelos de minería de datos. Para crear un modelo, el algoritmo analizó primero el conjunto de datos de aprendizaje correspondiente a cada área de negocio, estos datos se encuentran en tablas de escenarios que contienen todas las variables que permiten caracterizar al cliente, buscando patrones y tendencias específicas.

Dentro de los algoritmos que posee la herramienta y que fueron utilizados para implementar la solución del proyecto se encuentran:

- ***Algoritmo de árboles de decisión:*** Es un algoritmo de clasificación y regresión que proporciona herramientas para el modelado predictivo de atributos discretos y continuos.

- ***Algoritmo Bayes naive:*** Es un algoritmo de clasificación para el modelado de predicción. Este algoritmo calcula la probabilidad condicional entre columnas de entrada y de predicción y supone que las columnas son independientes.
- ***Algoritmo de Red Neuronal:*** Crea modelos de minería de datos de clasificación y regresión mediante la generación de una red de neuronas de tipo perceptrón multicapa. De forma similar al algoritmo de árboles de decisión de Microsoft, el algoritmo de red neuronal de Microsoft calcula las probabilidades para cada posible estado del atributo de entrada cuando se da cada estado del atributo de predicción. Posteriormente, puede utilizar estas probabilidades para predecir un resultado del atributo predicho basado en los atributos de entrada.

1.3 Importancia y naturaleza del estudio

El desarrollo exitoso de este proyecto puede ser muy beneficioso para cualquier empresa que necesite conocer y entender mejor a sus clientes, pues esta eficiente relación traerá como consecuencia innumerables beneficios económicos, debido a que la información que se obtiene al aplicar técnicas de minería de datos al data warehouse corporativo, permite ejecutar una acción predictiva y no reactiva, es decir, realizar una acción de retención cuando un cliente está considerando dejar el servicio en los próximos meses y no cuando ya ha tomado la decisión de dejarlo. Esta acción favorece indiscutiblemente al incremento de los índices de retención de clientes, pues permite orientar los esfuerzos hacia nichos específicos, mediante ofertas acordes a sus necesidades, concentrando los presupuestos disponibles en clientes con riesgo real de fuga y no en campañas de retención masiva.

Por otra parte, se obtiene un mayor nivel de automatización, dado que el ciclo del proceso, desde la extracción y carga de la información, la creación del modelo predictivo y la generación de reportes para el análisis de los resultados, son administrados por la misma herramienta.

1.4 Objetivos

1.4.1 Objetivo General

Desarrollar un modelo predictivo; utilizando técnicas de minería de datos; para detectar los patrones de fuga existentes en el comportamiento de los clientes de Telefónica del Sur con el objetivo de buscar la retención por medio de ofertas personalizadas.

1.4.2 Objetivos Específicos

- Analizar el problema desde una perspectiva empresarial y estudiar el conjunto de técnicas que permiten descubrir el conocimiento del negocio almacenado en las bases de datos y la extracción de patrones contenidos en ellas.

- Conocer y estudiar en profundidad el entorno de trabajo que posee la herramienta, determinar las variables predictivas y seleccionar los algoritmos que se utilizarán para el desarrollo del proyecto.

- Modelar e implementar un prototipo funcional del modelo predictivo.

- Evaluar la calidad de predicción del prototipo y corrección de los posibles errores del modelo.

2 INTRODUCCIÓN A LA MINERÍA DE DATOS

2.1 Conceptos generales

La minería de datos es un miembro clave en la familia de productos de Business Intelligence (BI), junto con el procesamiento analítico en línea (OLAP), los informes empresariales y ETL (Cargas, transformación y extracción de datos).
La minería de datos trata de analizar los datos y la búsqueda de patrones ocultos utilizando métodos automáticos o semiautomáticos. Durante la última década, grandes volúmenes de datos se han almacenado en las bases de datos, gran parte de estos datos proviene de software de negocios, tales como Aplicaciones Financieras, Planificación de Recursos Empresariales (ERP), Gestión de la Relación con Clientes (CRM), y Registros web. El resultado de este proceso ha convertido a las organizaciones ricas en datos e información pero pobres en conocimientos, llegando a alcanzar colecciones de datos tan grandes que el uso práctico de estos almacenes se ha convertido en limitada. El objetivo principal de la Minería de Datos es extraer patrones ocultos a partir de estos datos, aumentando su valor intrínseco y la transferencia de los datos al conocimiento.

La minería de datos proporciona a las empresas una gran valorización del negocio. Por ejemplo, el aumento de la competencia como resultado del marketing moderno y canales de distribución como Internet y las telecomunicaciones, han llevado a que las empresas se enfrenten a competencias en todo el mundo, siendo la clave para el éxito del negocio la capacidad de retener a los clientes existentes y adquirir nuevos. La minería de datos contiene las tecnologías que permiten a las empresas analizar los factores que afectan a estas materias.

Por otra parte, las tecnologías de minería de datos que anteriormente sólo existían en el ámbito académico, son tecnologías que han madurado y están listas para ser aplicadas en la industria. Los algoritmos son más precisos, más eficientes y pueden manejar datos cada vez más complicados. Además, las interfaces de Minería de Datos para la programación de aplicaciones (API) se han estandarizado, permitiendo a los desarrolladores construir mejores aplicaciones [Tan05].

2.2 Aplicaciones de Negocio para Minería de Datos

Las técnicas de minería de datos se pueden utilizar en muchas aplicaciones, en respuesta a diversos tipos de necesidades empresariales. La siguiente lista muestra algunos típicos problemas que pueden ser resueltos mediante la minería de datos:

- **Patrones de Fuga:** ¿Qué clientes tienen más probabilidades de cambiarse a un competidor? Las industrias de telecomunicaciones, banca y seguros se enfrentan a una intensa competencia en estos días. En promedio, cada nuevo suscriptor de telefonía móvil tiene un costo para las compañías de teléfono de más de 200 dólares de inversión en marketing. Por este motivo todas las empresas desearían mantener a tantos clientes como sea posible. El análisis del abandono, en base al descubrimiento de los patrones de fuga, puede ayudar a los gerentes de marketing a comprender los motivos de la rotación de clientes, mejorando la relación y comunicación con ellos.

- **Venta cruzada:** ¿Qué productos los clientes están dispuestos a comprar? Al comprar en una librería en línea se observa que el sitio web da una serie de recomendaciones sobre libros relacionados, estas recomendaciones podrían derivar del análisis de la minería de datos. Cross-Selling o ventas cruzadas es un reto empresarial importante para las empresas de retail. Muchos comerciantes, especialmente los minoristas en línea, utilizan esta función para incrementar sus ventas [Tan05].

- **Detección del fraude:** ¿Este seguro tendrá reclamos fraudulentos? Las compañías de seguros reciben miles de reclamaciones al día. Es imposible para ellos investigar cada caso. La minería de datos puede ayudar a identificar los reclamos que tengan mayor probabilidad de ser falso [Tan05].

- **La gestión del riesgo:** ¿Debería ser aprobado el préstamo a este cliente? Esta es la pregunta más común en el escenario de la banca. Las técnicas de minería de datos pueden ser utilizadas para marcar el nivel de riesgo del cliente, ayudando a los administradores a tomar una decisión adecuada para cada aplicación [Tan05].

- **La segmentación de clientes:** ¿Quiénes son mis clientes? La segmentación de clientes ayuda a los gerentes de marketing a entender los diferentes perfiles de sus

clientes y a tomar las medidas adecuadas de comercialización sobre la base de los segmentos [Tan05].

- **Los anuncios orientados:** ¿Qué banner se debe mostrar a un visitante específico? Los distribuidores de Internet y portales intentan personalizar su contenido para sus clientes web. La navegación de los clientes o los patrones de compra en línea, pueden utilizar soluciones de minería de datos para mostrar anuncios dirigidos a sus clientes [Tan05].

2.3 Tareas de Minería de Datos

La minería de datos se puede utilizar para resolver cientos de problemas de negocios. Sobre la base de la naturaleza de estos problemas, se pueden agrupar en las siguientes tareas.

2.3.1 Clasificación

La clasificación es una de las tareas de minería de datos más populares. Aplicaciones de negocios como el análisis de la rotación de clientes, gestión de riesgos y orientación de anuncios por lo general involucran clasificación. La clasificación se refiere a la asignación de casos en categorías basadas en un atributo predecible. Cada caso contiene un conjunto de atributos, uno de los cuales es el atributo de clase (atributo de predicción). La tarea requiere encontrar un modelo que describe el atributo de predicción en función de los atributos de entrada. Los típicos algoritmos de clasificación incluyen Arboles de Decisión, Redes Neuronales, y Redes Bayesianas [Tan05].

2.3.2 Agrupamiento

El agrupamiento o también llamado segmentación. Se utiliza para identificar agrupaciones naturales de los casos, sobre la base de un conjunto de atributos. Estos casos dentro del mismo grupo tienen valores similares de los atributos.
La segmentación es una tarea de minería de datos sin supervisión, es decir sin la guía de ninguna variable en particular. Todos los atributos de entrada son tratados por igual y el éxito consiste en agrupar a los individuos en segmentos que resulten significativos para el objetivo del proyecto. La mayoría de los algoritmos de agrupación construyen

el modelo a través de un número de iteraciones y se detienen cuando el modelo converge, es decir cuando los límites de estos segmentos son estabilizados [Tan05].

2.3.3 Asociación

Asociación es otra de las tareas de minería de datos más populares. También se suele llamar análisis del carrito de compras. Un problema típico de la asociación en aplicaciones empresariales es analizar una tabla de transacción de ventas e identificar los productos que a menudo se venden en la misma canasta de la tienda. El uso común de Asociación es identificar los grupos comunes de elementos y las reglas con el fin de conocer la venta cruzada. En términos de asociación, cada producto, o más generalmente, cada atributo / par de valor se considera como un elemento. La tarea de asociación tiene dos objetivos, encontrar conjuntos frecuentes de elementos y encontrar reglas de asociación [Tan05].

2.3.4 Regresión

La tarea de regresión es similar a la clasificación. La diferencia principal es que el atributo de predicción es una serie continua. Las técnicas de regresión se han estudiado ampliamente desde hace siglos en el campo de las estadísticas. Regresión lineal y regresión logística son los métodos de regresión más populares. Otras técnicas de regresión incluyen árboles de regresión y redes neuronales [Tan05].

Las tareas de regresión pueden resolver muchos problemas de negocios. Por ejemplo, pueden utilizarse para predecir las tasas de descuento sobre la base de la redención del valor nominal, método de distribución y volumen de distribución o para predecir la velocidad del viento sobre la base de temperatura, presión atmosférica y humedad [Tan05].

2.3.5 Pronostico

El pronóstico o forecasting es otra tarea de minería de datos importante. Es la proyección de una tendencia con respecto al tiempo u otra variable, consiste en la estimación y el análisis de la demanda futura para un producto en particular, componente o servicio, utilizando como entrada datos históricos de venta, informaciones de marketing o información promocional.

2.3.6 Análisis de la Desviación

Esta tarea sirve para encontrar los casos que se comportan muy diferentes de los demás. También se le llama detección de valores, que se refiere a la detección de cambios significativos del comportamiento observado previamente. El análisis de desviaciones puede ser utilizado en muchas aplicaciones. El más común es la detección de fraudes de tarjetas de crédito. Para identificar los casos anormales de millones de transacciones es una muy difícil. Otras aplicaciones incluyen la detección de intrusiones en la red, análisis de generación de errores, y así sucesivamente [Tan05].

2.4 Técnicas de Minería de Datos

Aunque la minería de datos como un término es relativamente nueva, la mayoría de las técnicas de minería de datos han existido por años. En las raíces de los algoritmos de minería de datos, se encontró que deriva principalmente de tres campos: las estadísticas, el aprendizaje automático, y las bases de datos.

La mayoría de las tareas de minería de datos que figuran en la sección anterior se han tratados en la comunidad estadística. Una serie de algoritmos de minería de datos, incluyendo la regresión, series de tiempo, y árboles de decisión, fueron inventados por los estadísticos. Las técnicas de regresión han existido por siglos. Los algoritmos de serie de tiempo se han estudiado durante décadas. El algoritmo de árbol de decisiones es uno de las técnicas más recientes, que datan de mediados de la década de 1980.

La minería de datos se centra en el descubrimiento automático o semiautomático de patrones. Varios algoritmos de aprendizaje automático se han aplicado a la minería de datos. Las redes neuronales son una de estas técnicas y son excelentes para la clasificación y la regresión, sobre todo cuando las relaciones de atributo no son lineales.

Los algoritmos de genética son una nueva técnica de aprendizaje automático, simula la evolución natural del proceso, trabajando con un conjunto de candidatos y una función de supervivencia. La función de supervivencia repetidamente selecciona a los candidatos más adecuados para la próxima generación. Los algoritmos genéticos se pueden utilizar para tareas de clasificación y agrupación, también se puede utilizar en combinación con otros algoritmos, por ejemplo, ayudar a una red neuronal para encontrar el mejor conjunto de pesos entre las neuronas.

La estadística tradicional asume que todos los datos pueden ser cargados en la memoria para el análisis estadístico, desafortunadamente esto no siempre es posible en el mundo moderno, sino son los expertos en bases de datos quienes saben cómo manejar grandes cantidades de datos que no caben en la memoria, por ejemplo, la búsqueda de reglas de asociación en una tabla de hechos que contienen millones de transacciones de ventas. Como cuestión de hecho, los algoritmos de asociación más eficientes provienen de la comunidad de investigación de bases de datos. Hay también algunas versiones escalables de algoritmos de clasificación y agrupamiento que utilizan técnicas de base de datos [Tan05].

2.5 Tipos de Columnas y Atributos

Una columna en un modelo de minería de datos es similar a la columna de una tabla relacional, ambas son llamadas variable o atributo en la terminología estadística. En función del uso, un modelo de minería de datos pueden tener tres tipos de columnas: clave, de entrada, de predicción o una columna que es de entrada y predicción a la vez. La columna de predicción es el objetivo del modelo de minería. La mayoría de los modelos de minería de datos utiliza el conjunto de columnas de entrada para predecir una columna de salida, aunque algunos algoritmos, como la agrupación, no requieren columnas de predicción, en este caso, el modelo de minería puede contener sólo las columnas de entrada.

Hay dos tipos de atributos: continuos y discretos. Los atributos continuos son aquellos atributos por lo general numéricos, tales como 23.45, 23.4555, 87. Por ejemplo, un atributo de datos continuos puede contener información como valor del sueldo actual, edad o distancia del trabajo al hogar. Los atributos discretos son aquellos datos categóricos tales como alto, bajo, varón o mujer. Generalmente se afirma que un atributo es discreto cuando tiene una cantidad determinada y acotada de posibles valores a seleccionar. Por ejemplo, una columna puede contener información salarial en rangos de sueldo codificados como 1 = < $250.000; 2 = de $250.000 a $500.000 y 3 > $500.000 o la edad codificada en los valores: niño, adolecente, adulto y mayor [Tan05].

2.6 El modelo de Minería de Datos

Un modelo de minería de datos, o modelo de minería, se puede considerar como una tabla relacional. Contiene las columnas clave, las columnas de entrada y las columnas de predicción, asociando cada modelo con un algoritmo de minería de datos con el que se entrena el modelo. El entrenamiento de un modelo significa encontrar patrones en el conjunto de datos de formación, utilizando determinados algoritmos de minería de datos con los parámetros del algoritmo adecuado.

Después del entrenamiento, el modelo de minería de datos almacena los patrones que el algoritmo de minería de datos descubre sobre el conjunto de datos. Mientras que una tabla relacional es un contenedor de archivos, un modelo de minería de datos es un recipiente de los patrones de comportamiento [Tan05].

2.6.1 Creación del Modelo

El concepto de creación de modelos, simplemente se refiere a la creación de un modelo de minería de datos vacía, similar a la manera de crear una nueva tabla.

2.6.2 Entrenamiento del Modelo

El entrenamiento del modelo también se conoce como procesamiento del modelo. Se utiliza para invocar al algoritmo de minería de datos y descubrir el conocimiento sobre el conjunto de datos de aprendizaje. Después del entrenamiento, los patrones se almacenan en el modelo de minería.

2.6.3 Predicción del Modelo

La predicción del modelo se utiliza para la aplicación de los patrones del modelo de minería de datos a un nuevo conjunto de datos y predecir el valor potencial de las columnas de predicción de cada nuevo caso.

2.7 Algoritmos de Minería de Datos

El algoritmo de minería de datos es el mecanismo que crea modelos de minería de datos. Para crear un modelo, un algoritmo analiza primero un conjunto de datos, buscando patrones y tendencias específicos. Después, el algoritmo utiliza los resultados de este análisis para definir los parámetros del modelo de minería de datos [MDN08].

El modelo de minería de datos creado por un algoritmo, puede tomar diversas formas incluyendo aplicaciones particulares como:

- Un conjunto de reglas de asociación que describen cómo se agrupan los productos en una transacción.
- Un árbol de decisión que predice si un cliente determinado comprará un producto.
- Un modelo matemático que predice las ventas.
- Un conjunto de clústeres que describe cómo se relacionan los escenarios de un conjunto de datos.

La elección del algoritmo apropiado para una tarea empresarial específica puede ser un trabajo difícil. Aunque puede utilizar diferentes algoritmos para realizar la misma tarea, cada uno de ellos genera un resultado diferente, y algunos pueden generar más de un tipo de resultado. Por ejemplo, puede usar el algoritmo Árboles de Decisión no sólo para la predicción, sino también como una forma de reducir el número de columnas de un conjunto de datos, ya que el árbol de decisión puede identificar las columnas que no afectan al modelo de minería de datos final.

Tampoco es necesario usar los algoritmos de modo independiente: en una solución de minería de datos puede utilizar algunos algoritmos para examinar los datos y, después, usar otros para predecir un resultado específico basándose en esos datos. Por ejemplo, se puede utilizar un algoritmo de clústeres, que reconoce patrones, para dividir los datos en grupos que sean más o menos homogéneos, y luego usar los resultados para crear un mejor modelo de árbol de decisión. También se podría utilizar varios algoritmos dentro de una solución para realizar tareas independientes, por ejemplo, usar un algoritmo de árbol de regresión para obtener información de previsiones financieras y un algoritmo basado en reglas para llevar a cabo un análisis del carro de compra [MDN08].

2.7.1 Algoritmo de Arboles de Decisión

El algoritmo de Árboles de Decisión es un algoritmo de clasificación y regresión para el modelado de predicción de atributos discretos y continuos [MDN08].

Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos. Utiliza los valores, o estados, de estas columnas para predecir los estados de una columna que se designa como elemento de predicción. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción. Por ejemplo, en un escenario para predecir qué clientes van a adquirir probablemente una bicicleta, si nueve de diez clientes jóvenes compran una bicicleta, pero sólo lo hacen dos de diez clientes de edad mayor, el algoritmo infiere que la edad es un buen elemento de predicción en la compra de bicicletas. El árbol de decisión realiza predicciones basándose en la tendencia hacia un resultado concreto.

Para los atributos continuos, el algoritmo usa la regresión lineal para determinar dónde se divide un árbol de decisión. Si se define más de una columna como elemento de predicción, o si los datos de entrada contienen una tabla anidada que se haya establecido como elemento de predicción, el algoritmo genera un árbol de decisión independiente para cada columna de predicción.

2.7.1.1 Funcionamiento del Algoritmo

El algoritmo de árboles de decisión genera un modelo de minería de datos mediante la creación de una serie de divisiones (denominadas nodos) en el árbol. El algoritmo agrega un nodo al modelo cada vez que una columna de entrada está correlacionada de forma significativa con la columna de predicción. La forma en que el algoritmo determina una división varía en función de si predice una columna continua o una columna discreta.

La forma en que el algoritmo de árboles de decisión genera un árbol para una columna de predicción discreta puede mostrarse mediante un histograma. La Figura 2 muestra un histograma que traza una columna de predicción, “Comprador Bicicletas”, con una columna de entrada, “Edad”. El histograma muestra que la edad de una persona ayuda a distinguir si esa persona comprará una bicicleta.

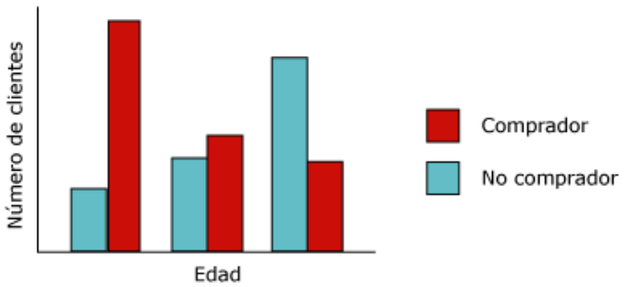


Figura 2: Relación entre la columna de predicción comprador y la columna de entrada Edad.

La correlación que aparece en el diagrama hará que el algoritmo de árboles de decisión cree un nuevo nodo en el modelo. En la Figura 3, se puede apreciar que a medida que el algoritmo agrega nuevos nodos a un modelo, se forma una estructura en el árbol. El nodo superior del árbol describe el desglose de la columna de predicción para la población global de clientes. A medida que el modelo crece, el algoritmo considera todas las columnas.

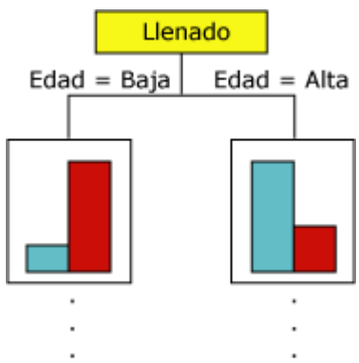


Figura 3: Formación de la estructura del árbol de decisión.

Cuando el algoritmo de árboles de decisión genera un árbol basándose en una columna de predicción continua, cada nodo contiene una fórmula de regresión. Se produce una división en un punto de no linealidad de la fórmula de regresión. Como se observa en la figura 4, el diagrama contiene los datos que pueden modelarse utilizando una sola línea o dos líneas conectadas. Sin embargo, una sola línea realizará un pobre trabajo en la representación de los datos. En su lugar, si se usan dos líneas, el modelo hará un mejor trabajo en la aproximación a los datos. El punto donde las dos líneas se unen es el punto de no linealidad y donde se dividiría un nodo de un modelo de árbol de decisión. Por ejemplo, el nodo que corresponde al punto de no linealidad del diagrama (a), podría

representarse mediante el diagrama (b). Las dos ecuaciones representan las ecuaciones de regresión de las dos líneas.

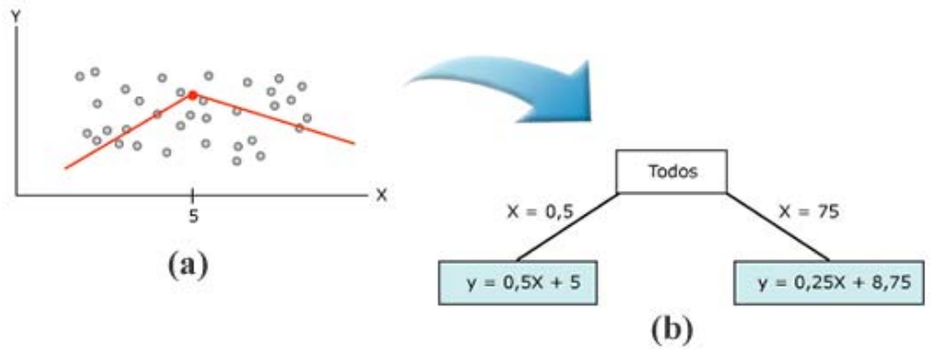


Figura 4: Formación de la estructura de un árbol de decisión con atributos de predicción continua.

2.7.2 Algoritmo de Bayes Naive

El algoritmo Bayes naive es un algoritmo de clasificación utilizado para el modelado de predicción. Este algoritmo calcula la probabilidad condicional entre columnas de entrada y de predicción y supone que las columnas son independientes. Esta suposición de independencia implica, de manera un tanto ingenua en ocasiones, que este algoritmo no tiene en cuenta las dependencias que puedan existir [MDN08].

Desde el punto de vista computacional, el algoritmo es menos complejo que otros, por lo tanto resulta útil para generar rápidamente modelos de minería de datos para descubrir relaciones entre columnas de entrada y columnas de predicción. Se puede utilizar este algoritmo para realizar exploraciones iniciales de datos y, más adelante, aplicar los resultados para crear modelos de minería de datos adicionales con otros algoritmos más complejos y precisos desde el punto de vista computacional.

El Algoritmo Bayes naive calcula la probabilidad de cada estado de cada columna de entrada, dado cada posible estado de la columna de predicción. El modelo debe contener una columna de clave, columnas de entrada y una columna de predicción, además todas las columnas deben ser discretas o discretizadas.

2.7.3 Algoritmo de Red Neuronal

Las redes neuronales realizan tareas de clasificación y regresión. Al igual que los árboles de decisión, las redes neuronales pueden encontrar relaciones no lineales entre los atributos de entrada y los atributos de predicción. Una desventaja importante de las Redes Neuronales es la dificultad de interpretar los resultados, debido a que no contiene más que un conjunto de pesos para la red, dificultando ver las relaciones en el modelo y por qué son válidas [Tan05].

Las redes neuronales soportan salidas discretas y continuas. Cuando las salidas son continuas, la tarea es de regresión. De hecho, las técnicas de regresión clásica, como la regresión logística, se puede representar como casos especiales de las redes neuronales. Aunque se utilizan normalmente para la clasificación y regresión, las redes neuronales *feed-forward*[4] también se puede aplicar a la segmentación, cuando se utiliza con una configuración de cuello de botella (capa oculta pequeños).

2.7.3.1 Estructura de una Red Neuronal

Las Redes Neuronales son más sofisticadas que los Árboles de Decisión y las Bayes Naive. La figura 5 muestra un par de ejemplos. Una red neuronal contiene un conjunto de nodos (neuronas) y los bordes que forman una red. Hay tres tipos de nodos: de entrada, ocultas y salida. Cada arista une dos nodos con un peso asociado. La dirección de un borde representa el flujo de datos durante el proceso de predicción. Cada nodo es una unidad de procesamiento, los nodos de entrada forman la primera capa de la red. En la mayoría de las redes neuronales, a cada nodo de la entrada se asigna un atributo de entrada como la edad, el género o los ingresos. El valor original de un atributo de entrada tiene que ser convertido a un número de punto flotante en la misma escala (a menudo entre -1 a 1) antes de procesar.

[4] Feed-forward o proalimentación describe un tipo de sistema que reacciona a los cambios en su entorno, normalmente para mantener algún estado concreto del sistema.

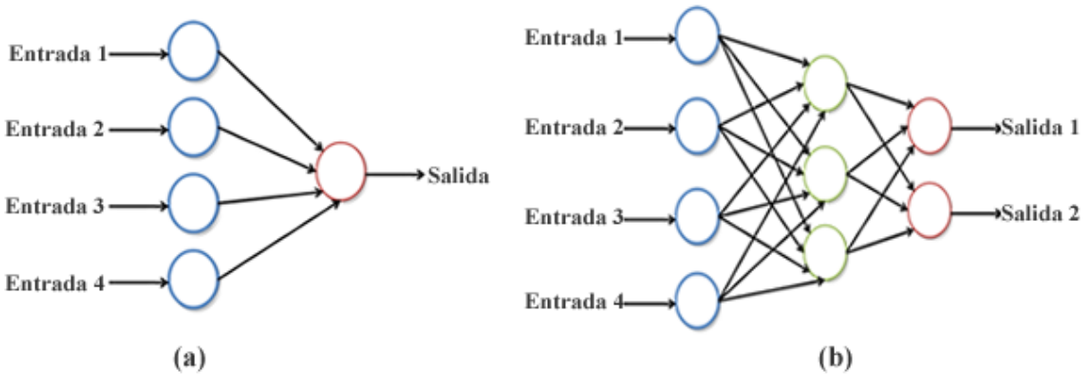


Figura 5: Ejemplo de una Red Neuronal

Los nodos ocultos son los nodos de las capas intermedias. Un nodo oculto recibe información de los nodos de las capas de entrada o de la capa oculta precedente. Esta combina todas las entradas basadas en el peso de los bordes asociados, procesa algunos cálculos, y emite un valor de resultado para el procesamiento en la siguiente capa.

Los nodos de salida por lo general representan los atributos de predicción. Una red neuronal puede tener varios atributos de salida, como se muestra en la figura 5.b. Es posible separando los nodos de salida en varias redes diferentes. Pero la mayoría de casos, se reduce el tiempo de procesamiento cuando se combinan en estas redes que pueden compartir los costes comunes de la exploración de los datos de origen. El resultado del nodo de salida es a menudo un número de punto flotante entre 0 y 1.

La predicción de la red neuronal es directa, los valores de atributos de entrada se normalizan y se asignan a las neuronas de la capa de entrada, a continuación cada nodo de la capa oculta procesa las entradas y desencadena una salida para las capas que siguen. Al final, las neuronas de salida comienzan un proceso y generan un valor de salida. Este valor se asigna a la escala original (en términos del atributo continuo) o la categoría original (en términos del atributo discreto).

Mientras que el procesamiento de una red neuronal toma mucho tiempo, hacer predicciones sobre una red neuronal entrenada es bastante eficiente. Como se muestra en la Figura 5, las topologías de las redes neuronales pueden variar, la figura 5 (a) muestra una red muy simple. Tiene un atributo de salida sin una capa oculta y todas las neuronas de entrada conectadas a la neurona de salida directamente, esta red neuronal es exactamente igual que la regresión logística. La figura 5 (b), es una red de tres capas: de entrada, ocultas y de salida. Hay tres neuronas en la capa oculta, cada neurona de la capa

oculta está plenamente conectada a las entradas de la capa precedente. La capa oculta es un aspecto muy importante de la red neuronal. Permite a la red aprender relaciones no lineales. Después que la topología de una red neuronal está configurada, es decir, el número de nodos ocultos se especifica, el proceso de formación consiste en encontrar el mejor conjunto de pesos para los bordes de la red, siendo esta una tarea que consume tiempo. En un principio, los pesos son asignados al azar, durante cada iteración de entrenamiento, la red procesa los casos de entrenamiento para generar predicciones sobre la capa de salida basándose en las configuraciones de red actual. A continuación, calcula el error de las salidas y sobre la base de estos errores, se ajusta los pesos de la red utilizando propagación hacia atrás.

3 METODOLOGÍA DE TRABAJO

Antes de comenzar a desarrollar cualquier proyecto que integre los nombres de minería de datos e inteligencia de negocios, se debe seleccionar una metodología de trabajo. Esta metodología se debe seguir paso a paso, con el objetivo de comprender cada una de sus fases, una metodología que explique cuando se debe hacer cada actividad y su razón.

Para este proyecto en particular el objetivo principal de la correcta aplicación de una metodología es alcanzar información relevante y precisa para los directores de Telefónica del Sur, así como para las personas encargadas del proceso de toma de decisiones.

Tomando en cuenta la gran cantidad de información que se maneja dentro de las empresas de telecomunicaciones y el progreso tecnológico de la última década, se han creado diferentes metodologías para realizar un análisis utilizando minería de datos, estableciendo ciertos parámetros que se deben cumplir, dependiendo de la información que se tenga y de lo que concretamente se desea, aunque los resultados sean obtenidos en un plazo no muy corto ya que una metodología consta de ciertas fases sucesivas que hay que seguir respetando el orden jerárquico.

Ciertas empresas han desarrollado metodologías para que el usuario pueda seguir utilizando al máximo la información. SAS[5] por ejemplo, puso a disposición de los usuarios la metodología SEMMA por sus siglas en ingles (Sample, Explore, Modify, Model, Assess). Un grupo de empresa de países europeos creo la metodología CRISP-DM por sus siglas en ingles (Cross-Industry Standard Process for Data Mining).

Las metodologías SEMMA y CRISP-DM comparten la misma esencia, estructurando el proyecto de Data Mining en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de minería de datos en un proceso iterativo e interactivo. La metodología SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología CRISP-DM, mantiene una perspectiva más amplia respecto a los objetivos empresariales del proyecto. Esta diferencia se establece ya desde la primera fase del proyecto, donde la metodología SEMMA comienza realizando un

[5] SAS: http://www.sas.com/
SAS es uno de los líderes en Business Analytics y servicios de software, y uno de los mayores proveedores independientes de Inteligencia de Negocio del mercado.

muestreo de datos, mientras que la metodología CRISP-DM comienza realizando un análisis del problema empresarial, para su transformación en un problema técnico.

Desde ese punto de vista más global se puede considerar que la metodología CRISP-DM está más cercana al concepto real de proyecto, pudiendo ser integrada con una Metodología de Gestión de Proyectos específica que completaría las tareas administrativas y técnicas.

Otra diferencia significativa entre la metodología SEMMA y la metodología CRISP-DM radica en su relación con herramientas comerciales. La metodología SEMMA sólo es abierta en sus aspectos generales ya que está muy ligada a los productos SAS donde se encuentra implementada. Por su parte la metodología CRISP-DM ha sido diseñada como una metodología neutra respecto a la herramienta que se utilice para el desarrollo del proyecto, siendo su distribución libre y gratuita [Fer06].

3.1 Elección de la Metodología

Se escogió a CRISP-DM como la metodología más apropiada para el desarrollo del trabajo de tesis por considerarla más completa que SEMMA, principalmente porque posee una fase de desarrollo dedicada íntegramente al entendimiento del negocio, y por su flexibilidad, al permitir trabajar con cualquier herramienta de explotación de datos.

A continuación se ampliaran las definiciones hechas sobre la metodología CRISP-DM.

3.2 Metodología CRISP-DM

3.2.1 Introducción

La Metodología CRISP-DM, aunque se desarrolló para llevar adelante grandes proyectos, es suficientemente amplia y flexible para aplicarla a proyectos de cualquier tamaño. En la figura 6, se esquematiza el ciclo de vida de un proyecto desarrollado con la metodología.

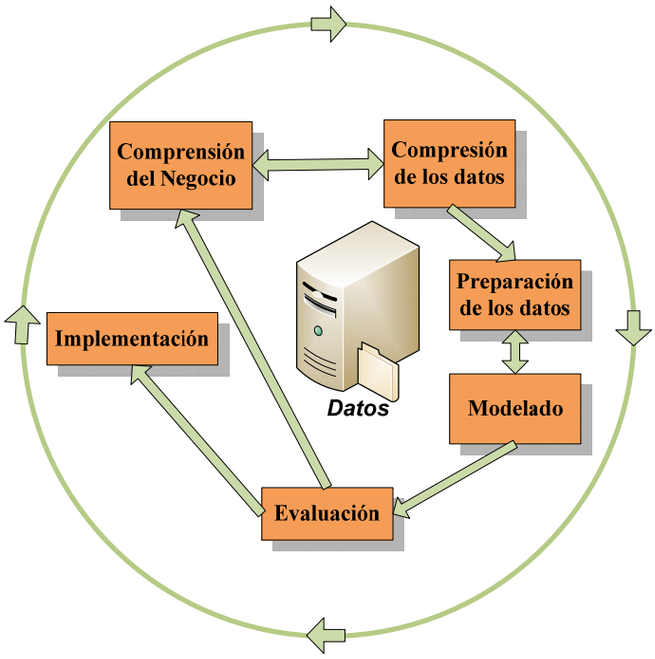


Figura 6: Fases de la Metodología CRISP-DM [CRI].

El ciclo de vida de la metodología consiste en seis fases, cuya sucesión no es rígida, y se puede mover entre ellas siempre que se requiera. Las flechas indican la dependencia más importante y frecuente entre las fases. El círculo exterior simboliza la naturaleza cíclica de los proyectos de minería de datos.

La metodología se presenta en términos de un proceso jerárquico. Consiste en un juego de tareas descriptas en niveles de abstracción (de lo general a lo específico): la fase, la tarea genérica o subfase, la tarea especializada y el caso del proceso.

El contexto de CRISP-DM se maneja entre lo genérico y el nivel especializado, dentro del cual se distinguen cuatro dimensiones diferentes:

- ***Dominio de la aplicación:*** Especifica el área en que el proyecto de minería de datos tiene lugar.
- ***Tipo de problema:*** Describe la clase y objetivos del proyecto.
- ***Aspecto técnico:*** Cubre procesos específicos de la minería de datos, describe diferentes desafíos que normalmente ocurren.
- ***Herramienta técnica:*** Especifica que se aplica durante el proceso de minería de datos.

3.2.2 Análisis de las fases de CRISP-DM

A continuación se detalla como se compone cada una de las fases metodológicas de CRISP-DM.

3.2.2.1 Comprensión del Negocio

Esta fase inicial se focaliza en el entendimiento de los objetivos y requerimientos desde una perspectiva de negocios. Este conocimiento se convierte en una definición de problema de minería de datos y en un plan preliminar diseñado para llevar a cabo los objetivos.

El primer objetivo es comprender a fondo, desde una perspectiva de negocio, lo que el cliente realmente quiere lograr. A menudo el cliente tiene muchos objetivos y restricciones que compiten, los cuales deben ser correctamente equilibrados. El objetivo del analista debe destapar factores importantes en el principio del proyecto esto puede influir en el resultado final. Una consecuencia probable de descuidar este paso debe ser a expensas de hacer un gran esfuerzo de producir las respuestas correctas a las preguntas incorrectas.

Finalmente esta fase debe terminar con la descripción del plan intencionado para alcanzar los objetivos de minería de datos y así alcanzar los objetivos de negocio. El plan debería especificar los pasos para ser realizados durante el resto del proyecto, incluyendo la selección inicial de herramientas y técnicas.

3.2.2.2 Comprensión de los Datos

La fase de comprensión de los datos, comienza con una colección inicial y continua con actividades tendientes a familiarizarse con ellos, para identificar los problemas de calidad, descubrir primeras vistas internas o detectar interesantes subconjuntos para formular hipótesis sobre la información oculta.

Esta colección inicial incluye carga de datos, si es necesario, para la comprensión de los datos, significando un esfuerzo que posiblemente conduce a los pasos iniciales de la preparación de datos.

3.2.2.3 Preparación de los Datos

La fase de preparación de datos cubre todas las actividades para construir el conjunto de datos finales a partir de los datos iniciales en bruto. Este conjunto de datos llamado datos de aprendizaje conforman las tablas de escenarios que se utilizaran para entrenar los modelos de minería de datos.

Las tareas de preparación de datos se suelen desarrollar múltiples veces y no tienen un orden prescrito. Las tareas incluyen la selección de tablas, registros y atributos como también la transformación y limpieza de los datos por herramientas de modelamiento.

En algunas ocasiones se desarrollan operaciones de preparación de datos, tales como la producción de atributos derivados, el ingreso de nuevos registros o la transformación de valores para atributos existentes.

3.2.2.4 Modelado

En esta fase se seleccionan y aplican varias técnicas de modelado, calibrando sus parámetros a valores óptimos. Típicamente hay algunas técnicas para los mismos tipos de problemas de minería de datos, aunque algunas técnicas tienen requerimientos específicos en la forma de los datos, y muchas veces es necesario volver a la fase de preparación de datos.

Como primer paso, se selecciona la técnica de modelado real que se utilizará. Aunque la selección de una herramienta fue realizada durante la fase de comprensión del negocio, esta tarea se refiere a la técnica de modelado específico. Si se aplican múltiples técnicas, cada tarea se realizará separadamente para cada técnica.

El ingeniero de minería de datos interpreta los modelos según su conocimiento de dominio, los criterios exitosos de minería de datos, y el diseño de prueba deseado. El ingeniero de minería de datos juzga el éxito de la aplicación del modelado y descubre nuevas técnicas más eficientes; él se pone en contacto con analistas del negocio para hablar de los resultados de la minería de datos en el contexto de negocio.

3.2.2.5 Evaluación

En esta fase, ya se ha construido el modelo (o modelos) que parecen tener alta calidad desde una perspectiva de análisis de datos. Antes de proceder a la implementación final del modelo, se debe evaluar de forma más exhaustiva el modelo y revisar los pasos ejecutados para construirlo y asegurarse que interprete de forma adecuada los objetivos del negocio. Un objetivo clave es determinar si hay algún aspecto importante del negocio que no haya sido suficientemente considerado. Al final de esta fase, se debería haber alcanzado alguna decisión en el uso de los resultados de minería de datos.

Los pasos de la evaluación trata factores como la exactitud y la generalidad del modelo. Este paso evalúa el grado al que el modelo responde (encuentra) los objetivos de negocio y procura determinar si hay alguna decisión de negocio por el que este modelo es deficiente. Otra opción de evaluación es probar el/los modelo/s sobre aplicaciones de prueba en la aplicación real, si el tiempo y las restricciones de presupuesto lo permiten.

3.2.2.6 Implementación

La creación del modelo no es, por lo general, el final del proyecto. Aún si el propósito del modelo es incrementar el conocimiento de los datos, el conocimiento obtenido necesitara ser reorganizado y presentado de una manera que el cliente pueda utilizarlo. A menudo, la aplicación involucra modelos “vivos” dentro de los procesos de toma de decisiones de la organización, por ejemplo, en la personalización de páginas web en tiempo real. Sin embargo dependiendo de los requerimientos, la fase de implementación puede ser tan simple como generar un reporte, o tan complejo como la implementación de un proceso de minería de datos a lo ancho de toda la empresa.

En muchos casos es el cliente, no el analista de datos, quien lleva a cabo los pasos de esta implementación. Sin embargo, aún si el analista no lo lleva a cabo es importante para el cliente entender que acciones necesitan ser llevadas a cabo en orden de hacer un uso práctico de los modelos creados.

4 HERRAMIENTA DE DESARROLLO

En el mercado actual las herramientas informáticas y software de Data Mining han visto incrementada su oferta de forma exponencial. Numerosas empresas dedicadas al análisis y creación de bases de datos, en búsqueda de información valiosa para el cliente, han desarrollado su propio software de minería de datos para responder a las necesidades de información y búsquedas de asociaciones en los datos estudiados [San05].

La mayoría de las herramientas que aplican técnicas de minería de datos para el análisis y búsqueda de tendencias y asociaciones entre los datos disponibles, emplean ciertos procedimientos estadísticos tales como árboles de decisión, redes neuronales, series temporales, criterios bayesianos, etc. El empleo de estos métodos es indispensable para la clasificación, reducción, simulación, asociación y predicción en los datos. Ello no implica, sin embargo, que los usuarios de las herramientas de minería de datos existentes necesiten disponer de los conocimientos teóricos y estadísticos asociados a dichos procedimientos. Los entornos en que se desarrollan estos programas suelen presentar una visualización intuitiva y enlaces a través de iconos y ediciones que permiten su uso de forma rápida y sencilla.

Existen muchas herramientas de software para el desarrollo de modelos de minería de datos tanto libres como comerciales, por ejemplo, SQL Server Analysis Services, SPSS Clementine, RapidMiner, Orange, Weka, KNIME, etc. Estas aplicaciones de software sobre minería de datos se pueden encontrar en la actualidad en el mercado, constituyen una amplia variedad de programas con características muy versátiles y distintas relaciones calidad-precio.

El desarrollo de este proyecto fue realizado con la herramienta de la empresa Microsoft llamada SQL Server 2005 Analysis Services (SSAS). Esta plataforma utiliza componentes de servidor y de cliente para proporcionar la funcionalidad de procesamiento analítico en línea (OLAP) y de minería de datos para aplicaciones de Business Intelligence.

4.1 Introducción a SQL Server 2005

SQL Server 2005 es una plataforma completa, de extremo a extremo para la inteligencia de negocio (BI), incluyendo el Data Warehause, el procesamiento análisis en línea (OLAP), la extracción, transformación y carga de datos (ETL), Minería de Datos, y presentación de informes. Las herramientas para diseñar y desarrollar soluciones, así como el manejo y la operación con ellos también están incluidas [Han05].

Para lograr lo anterior, SQL Server 2005 se compone de una serie de componentes integrados, como se muestra en la Figura 7. En ella se puede apreciar los componentes relativos a una solución de BI, pero SQL Server también incluye todos los servicios necesarios para la construcción de todo tipo de aplicaciones centradas en datos seguros, confiables y robustos.

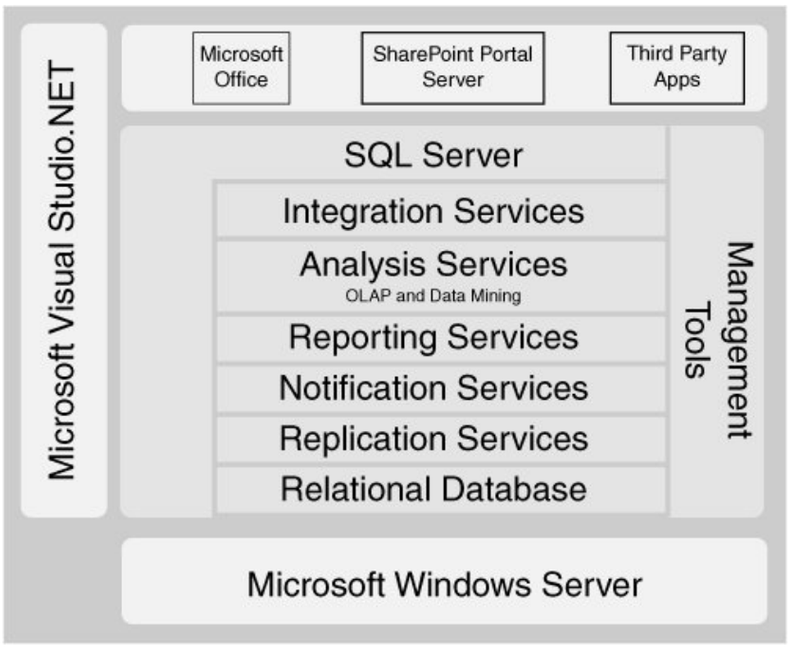


Figura 7: Componentes de SQL Server 2005 [Han05].

4.2 Desarrollo y Herramientas de Gestión

SQL Server incluye dos ambientes complementarios para el desarrollo y gestión Soluciones de Inteligencia de Negocios. **SQL Server Management Studio** que permite administrar todos los aspectos de las soluciones en un solo ambiente de gestión. Los administradores pueden gestionar varios servidores diferentes en el mismo lugar,

incluyendo el Motor de Base de Datos, Servicios de Análisis, Servicios de Integración y los Servicios de Reportes.

Para el desarrollo de soluciones de BI se utiliza la herramienta **Business Intelligence Development Studio.** Este es un ambiente único, rico para el análisis de la construcción de cubos y estructuras de minería de datos, paquetes de Integration Services, y para el diseño de informes. BI Development Studio es construido en la cima de la tecnología de Visual Estudio, por lo que se integra bien con las herramientas existentes, como fuente de repositorios de control.

En la Figura 8 se muestra cómo cada una de estas piezas se une para crear un ambiente de apoyo a la solución de Inteligencia de Negocio.

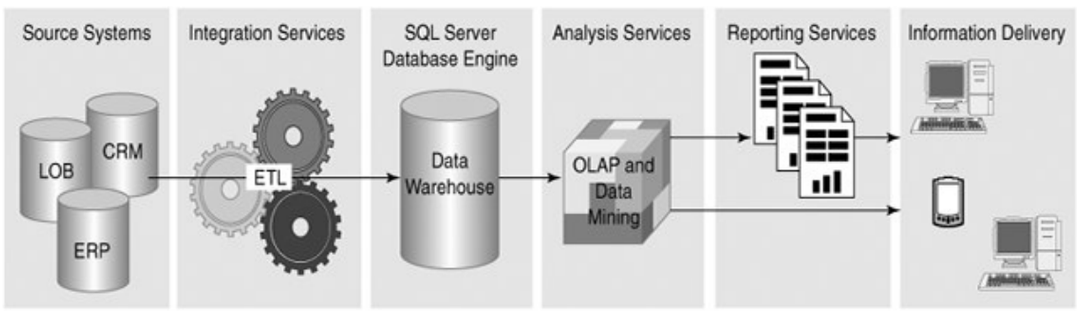


Figura 8: Arquitectura de una solución de Inteligencia de Negocio en SQL Server [Han05].

4.3 Arquitectura de Minería de Datos

La minería de datos en SQL Server 2005 está integrado en el motor de Analysis Services, como se muestra en la Figura 9. La información derivada de la minería de datos puede estar disponible como parte de los cubos de Analysis Services e informes de Reporting Services, de esta forma los usuarios puedan aplicar las agrupaciones y las predicciones de minería de datos para los datos existentes.

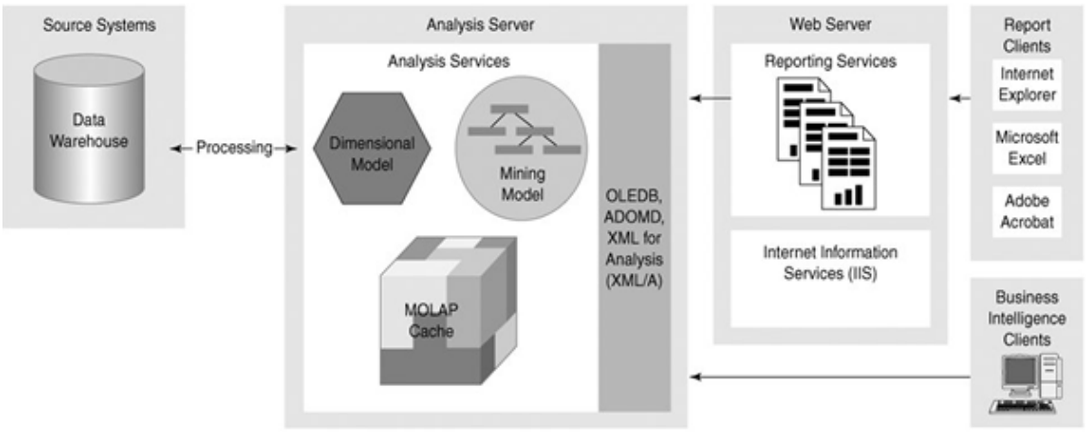


Figura 9: Arquitectura de una solución de Minería de Datos en SQL Server [Han05].

4.4 Usar las herramientas de Minería de Datos

Microsoft SQL Server 2005 Analysis Services (SSAS) proporciona las herramientas que se utilizan para crear soluciones de minería de datos que permiten resolver problemas empresariales concretos.

Business Intelligence Development Studio, el Asistente para minería de datos, facilita la creación de estructuras y de modelos de minería de datos basados en orígenes de datos OLAP. Puede utilizar el asistente para definir estructuras y modelos que utilicen técnicas de minería de datos específicas para analizar datos. Puede utilizar el Diseñador de minería de datos para perfeccionar la definición de modelos de minería datos y explorar y trabajar con los resultados del modelo.

SQL Server Management Studio proporciona herramientas que pueden utilizarse para administrar y explorar los modelos de minería de datos. SQL Server 2005 Integration Services (SSIS) contiene herramientas útiles para limpiar datos, automatizar tareas como la creación de predicciones o actualización de modelos y para crear soluciones de minería de datos de texto.

4.4.1 Asistente para la Minería de Datos

El Asistente para minería de datos de Microsoft SQL Server 2005 Analysis Services (SSAS) se ejecuta cada vez que se agrega una nueva estructura de minería de datos a un proyecto de minería de datos. El asistente define nuevas estructuras y también define el modelo de minería de datos inicial para cada estructura. La estructura del modelo inicial, incluyendo las tablas y columnas, se deriva de una vista de origen de datos o un cubo existentes.

Cuando se crea un modelo de minería de datos desde un origen de datos relacional, previamente se especifica en el asistente para minería de datos que se desea usar una base de datos relacional para definir la estructura del modelo. Después, se especifica la técnica de minería de datos que se va a utilizar, seleccionando el algoritmo más apropiado para el tipo de análisis de minería de datos que se desea.

El algoritmo de minería de datos es el mecanismo que crea modelos de minería de datos. Para crear un modelo, un algoritmo analiza primero un conjunto de datos, buscando patrones y tendencias específicos. Después, el algoritmo utiliza los resultados de este análisis para definir los parámetros del modelo de minería de datos.

Microsoft SQL Server 2005 Analysis Services (SSAS) proporciona varios algoritmos que puede usar en las soluciones de minería de datos. Estos algoritmos son un subconjunto de todos los algoritmos que pueden utilizarse en la minería de datos. También puede utilizar algoritmos de minería de datos desarrollados por terceros que cumplan la especificación OLE DB[6] para minería de datos.

Analysis Services incluye los siguientes algoritmos:

- Algoritmo de asociación de Microsoft.
- Algoritmo de clústeres de Microsoft.
- Algoritmo de árboles de decisión de Microsoft.
- Algoritmo Bayes naive de Microsoft.
- Algoritmo de red neuronal de Microsoft.
- Algoritmo de clústeres de secuencia de Microsoft.
- Algoritmo de serie temporal de Microsoft.

[6] OLE DB: Object Linking and Embedding for Databases ("Enlace e incrustación de objetos para bases de datos"). http://es.wikipedia.org/wiki/OLE_DB

- Algoritmo de regresión lineal de Microsoft.
- Algoritmo de regresión logística de Microsoft.

4.4.2 Diseñador de Minería de Datos

El Diseñador de minería de datos es el entorno principal en el que se trabaja con modelos de minería de datos en Microsoft SQL Server 2005 Analysis Services (SSAS). Puede obtener acceso al diseñador seleccionando una estructura de minería de datos existente o utilizando el asistente para minería de datos para crear una nueva estructura y un nuevo modelo de minería de datos. Puede usar el diseñador de minería de datos para realizar las tareas siguientes:

- Modificar la estructura y el modelo de minería de datos que se crearon inicialmente con el asistente para minería de datos.
- Crear nuevos modelos basados en una estructura de minería de datos existente.
- Entrenar y examinar modelos de minería de datos.
- Comparar modelos mediante gráficos de precisión.
- Crear consultas de predicción basadas en modelos de minería de datos.

Un elemento de estructura de minería de datos contiene una única estructura de minería de datos y todos sus modelos de minería de datos asociados. Cada modelo de minería de datos puede tener diferentes tipos de algoritmos, configuraciones de parámetros y columnas de los que se incluyen en la estructura de minería de datos. Debido a que todos los modelos asociados están contenidos en una única estructura, puede comparar el rendimiento de los modelos utilizando visores y gráficos de precisión.

4.4.3 Minería de Datos en SQL Server Management Studio

SQL Server Management Studio proporciona un entorno para administrar y trabajar con modelos de minería de datos que ya existen en una base de datos de Analysis Services. Con Management Studio puede realizar las siguientes tareas:

- Procesar objetos de modelo de minería de datos.
- Examinar modelos de minería de datos.
- Crear secuencias de comandos de objetos de minería de datos.
- Crear consultas de predicción.
- Eliminar objetos de minería de datos de la base de datos

4.4.3.1 Procesar Objetos de Minería de Datos

El procesamiento de una estructura o un modelo de minería de datos es diferente a los procesamientos de objetos OLAP, en el caso multidimensional. En el procesamiento OLAP se crean cubos con agregaciones, en tanto que en el procesamiento de minería de datos se crean datos de entrenamiento y se ejecutan algoritmos de minería en esos datos.

Los modelos de minería de datos deben ser procesados siempre que se cambia la estructura del modelo de minería de datos, se actualizan los datos de aprendizaje, se cambia el modelo de minería de datos existente o se agrega un nuevo modelo de minería de datos a la estructura.

4.4.3.2 Examinar los Modelos de Minería de Datos

Se puede obtener acceso a los visores de modelos de minería de datos de Management Studio desde una estructura de minería de datos o desde un modelo de minería de datos. Esta herramienta utiliza los mismos visores disponibles en Business Intelligence Development Studio.

La exploración de un modelo de minería de datos permite comprender el comportamiento del modelo antes de implementarlo. Cada algoritmo que utilice para generar un modelo devuelve un tipo de resultados diferente. Por tanto, Analysis Services proporciona un visor independiente para cada algoritmo. Cuando se examina un modelo de minería de datos en Business Intelligence Development Studio, el modelo se muestra utilizando el visor adecuado para el modelo.

El visor para cada tipo algoritmo también contiene el visor de contenido genérico. Este visor genérico muestra el contenido común a todos los algoritmos. El visor genérico contiene información sobre los patrones y estadísticas que Analysis Services captura durante el análisis, las características y probabilidades para los nodos individuales, las fórmulas de regresión y otros detalles técnicos.

4.4.3.3 Crear Consultas de Predicción

SQL Server Management Studio proporciona un conjunto completo de características, incluyendo un editor de consultas y un generador de consultas, que puede usarse para

generar y ejecutar consultas de extensiones de minería de datos (DMX). El editor de consultas de predicción contiene herramientas que simplifican el proceso de creación y modificación de consultas.

Este editor de consultas, puede escribir una consulta propia o utilizar una plantilla para tareas habituales como la creación de un nuevo modelo de minería de datos.

4.4.4 Transformaciones y Tareas de Minería de Datos en Integration Services

SQL Server Integration Services proporciona herramientas que se puede utilizar para automatizar tareas comunes de minería de datos, como procesar un modelo de minería de datos y crear consultas de predicción. Por ejemplo, si dispone de un modelo de minería de datos generado a partir de un conjunto de datos de posibles clientes, puede crear un paquete de Integration Services que actualice automáticamente el modelo cada vez que el conjunto de datos se actualice con nuevos clientes. A continuación podría utilizar el paquete para crear una predicción, separando los clientes potenciales en dos tablas. Una tabla contendría los clientes probables y la otra los clientes que posiblemente no adquirirán ningún producto.

5 DESARROLLO DEL MODELO PREDICTIVO

5.1 Comprensión del Negocio

El primer objetivo en la etapa de analista de datos, fue entender desde una perspectiva del negocio, lo que Telefónica del Sur realmente quería lograr con el proyecto.

Como primera etapa se efectuaron reuniones con el personal del Área Comercial de Telefónica del Sur, quienes están encargados de los sistemas abocados a la fidelización de clientes y la disminución de los índices de abandono. El objetivo de las reuniones fue dialogar sobre la potencialidad de las aplicaciones de minería de datos en las empresas de telecomunicaciones, y de que forma la aplicación de modelos predictivos que descubren los patrones de fuga de clientes, podía potenciar sus sistemas de fidelización.

El resultado de estas reuniones fue fijar los objetivos del proyecto expuestos en el capítulo 1 y comenzar con una serie de reuniones, enfocándose en la investigación detallada de todos los recursos, restricciones y necesidades que debían ser considerados en la construcción del plan del proyecto.

En las reuniones de trabajo se vieron y trabajaron los siguientes temas:

- Conocer de qué forma trabaja la empresa en cuanto a la generación de información que permita conocer más y mejor al cliente, en función de disminuir los índices de abandono.
- Cuáles eran los sistemas que generaban reportes para evaluar la situación de la empresa en estas áreas.
- Que herramientas de software ocupan para el desarrollo de estos reportes.
- Cuáles son las variables que consideran estos sistemas para generar reportes.

Una vez realizada estas reuniones se definió claramente el ámbito del problema y el objetivo final del proyecto de minería de datos, quedando claro qué se estaba buscando y qué atributo del conjunto de datos se desea predecir.

El análisis de mayor relevancia fue el de enfocar el proyecto en las áreas de negocio y no en los clientes, es decir, se construiría un modelo de minería de datos para las cuatro áreas de negocio que a Telefónica del Sur le interesa analizar. Estos servicios son,

Telefonía Post Pago (TPOS), Internet Banda Ancha (ADSL), Televisión Digital Inalámbrica (WITV) y Telefonía Local Inalámbrica (PHS). Para generar estos modelos de minería de datos, la fuente de datos será una tabla bidimensional llamada tabla de escenarios, que contendrá la información de los clientes que poseen ese determinado servicio o que en el pasado lo poseyeron.

Cada registro de la tabla de escenario asociada a la estructura y al modelo de minería de datos será un cliente del área de negocio que se está analizando, en donde un atributo del registro dirá si cada cliente, es un cliente actual de la empresa o un cliente que se fugó. Este atributo, será el único atributo de predicción que tendrá el modelo.

5.2 Comprensión de los Datos

En esta fase de comprensión de los datos se inicia la colección de datos inicial, siguiendo con actividades que permiten familiarizarnos con los datos, identificar los problemas de calidad, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

Para el proceso de colección de datos inicial, como primera instancia, se realizo una reunión con la Subgerencia de Planificación, Seguimiento y Fidelización. El objetivo de esta reunión fue solicitar el conjunto de variables o atributos que según su experiencia, en las empresas de telecomunicaciones, son relevantes para poder identificar comportamientos de satisfacción o insatisfacción de los clientes.

Una vez obtenida estas variables, principalmente de comportamiento y de calidad de servicio, como se pueden apreciar en la Tabla 1, se dio inicio a la construcción de las tablas que servirán como fuente de datos para la creación del modelo.

Tabla 1: Variables para el modelo predictivo y relación con el área de negocio en la que influye.

Variable	TPOS	BA	WITV	PHS
Q Visitas Oficinas Comerciales	✓	✓	✓	✓
Q de Llamadas a la competencia en los últimos X mes(es)	✓			✓
Llamadas a Call Center	✓	✓	✓	✓
Promedio Tráficos entrantes y salientes de los 3 meses anteriores y el promedio del último mes	✓			✓
Reclamos Comerciales (evaluar Q y por cuanto tiempo, últimos 3 meses=? Ultimo mes?	✓	✓	✓	✓
Reclamos Técnicos (evaluar Q y por cuanto tiempo, últimos 3 meses=? Ultimo mes?	✓	✓	✓	✓
Impugnaciones	✓	✓	✓	✓
No pago facturas (evolución conducta de pago)	✓	✓	✓	✓
Contratación de PPV			✓	
Modificaciones en el servicio los últimos 3 meses (cambios de velocidad, planes, baja de planes de canales, etc.)	✓	✓	✓	✓
Antigüedad de los servicios	✓	✓	✓	✓

En la Tabla 1, se cuenta con un conjunto de variables que poseen información relevante sobre el comportamiento de los clientes en Telefónica del Sur. En la columna de la izquierda, llamada variables, se encuentra una pequeña definición del objetivo de cada una de ellas. En las columnas de la derecha se encuentran las áreas de negocio en las que estas variables tienen relevancia e influencia, como se explico anteriormente las áreas de negocios son: Telefonía Post Pago (TPOS), Internet Banda Ancha (ADSL), Televisión Digital Inalámbrica (WITV) y Telefonía Local Inalámbrica (PHS).

La forma en que se procesó cada una de estas variables entregadas por el área de fidelización de clientes de Telefónica del Sur es la siguiente.

- ***Q Visitas a Oficinas Comerciales:*** Esta atributo se consideró como una variable continua. El objetivo es conocer la cantidad de veces que un cliente concurre a las oficinas comerciales para manifestar alguna inquietud, ya sea por una contratación, atención general, servicio técnico o contención, esto es, cuando un cliente se acerca a devolver su servicio contratado o solicita alguna oferta.

Esta variable es considerada en las tablas de escenarios de las cuatro áreas de negocio investigadas.

- ***Q de Llamadas a la competencia:*** Esta atributo se consideró como una variable continua. Consiste en conocer la cantidad de llamadas que realizan los clientes a los números de la competencia, los números considerados son aquellos que poseen las empresas de la competencia para la venta de sus productos, por lo general, números 800 y 600. Este atributo se analiza en las áreas de negocio de TPOS y PHS, conformando además dos variables, una en la que se analizan las llamadas el mes anterior al mes de análisis y otra para las llamadas realizadas en los últimos tres meses anteriores al mes de análisis.

- ***Llamadas a Call Center:*** Esta atributo se consideró como una variable continua. El objetivo era conocer la cantidad de veces que un cliente llama a Call Center, para ello se pretendía agregar a cada tabla de escenario de las cuatro áreas, los clientes que solicitaban la atención de Call Center, identificando en cada llamada el servicio que generaba la consulta.

  Posteriormente en la etapa de preparación de los datos se determinó que esta variable no será considerada por el sistema, ya que la empresa al procesar la llamada, no especifica claramente los datos para relacionar la mayoría de las llamadas a un cliente y por sobre todo, al área de negocio que provoca la necesidad de llamar al Call Center.

- ***Promedio del Tráfico de llamadas entrantes y salientes:*** Esta atributo se consideró como una variable discreta, en donde, existen cinco rangos de llamadas para clasificar la cantidad de llamadas realizadas o recibidas por un cliente. Al igual que el caso anterior, en la etapa de preparación de los datos, por un tema de disponibilidad de la información por parte de la empresa, se consideró para el análisis solamente el tráfico de llamadas salientes de los clientes. Este tráfico de llamadas es estimado para las áreas de negocio de TPOS y PHS, siendo además, contabilizado el tráfico en los últimos tres meses anteriores al mes de análisis.

- ***Reclamos Comerciales:*** Este atributo al igual que en casos anteriores, es una variable continua, tiene como objetivo contabilizar la cantidad de reclamos de tipo comercial que realizan los clientes de Telefónica del Sur. Este atributo es considerada en las tablas de análisis de las cuatro áreas de negocio investigadas, generando a su vez dos variables, uno con los reclamos comerciales realizado el mes anterior y otro con el historial de reclamos realizados los últimos tres meses anteriores al mes analizado.

- ***Reclamos Técnicos:*** Este atributo es muy similar a la anterior, es continuo, se considera en las tablas de análisis de las cuatro áreas de negocio investigadas y se generan dos variables, uno con los reclamos técnicos realizados el mes anterior y otro con los reclamos realizados los últimos tres meses anteriores. La diferencia es que este atributo contabiliza todos los reclamos de tipo técnico, como por ejemplo, lentitud al navegar por internet, imposibilidad de realizar llamadas, imagen en negro sin guía de canales, etc.

- ***Impugnaciones:*** Es un tipo de reclamo distinto, es generados en una etapa posterior a los reclamos anteriores, se genera a partir de un rechazo por parte del cliente a un cobro indebido o mal realizado, asociado a una boleta o factura emitida. Se trata de una variable continua, considerada en las tablas de escenarios de las cuatro áreas de negocio investigadas y se descompone en una variable para las impugnaciones ocurridas el mes anterior y en otra variable para las impugnaciones ocurridas los últimos tres meses anteriores del mes analizado.

- ***No Pago de Facturas:*** Este atributo tiene como objetivo conocer la conducta de pago de los clientes, para ello asigna un código numérico a la variable según la situación de pago del mes analizado. Esta es una variable discreta y se asignan valores cuando se dan situaciones como: cancelación antes de la fecha de vencimiento, antes del corte, antes del vencimiento del documento siguiente, cliente sin facturación, etc.

- ***Contrataciones de pay per view (PPV):*** Las contrataciones de servicios de pay per view o también conocidas como televisión a la carta, crean una variable discreta. El atributo almacena en rangos la cantidad de veces que un cliente de Televisión Digital Inalámbrica contrata uno de estos productos. Para un análisis

más detallado, se desprendieron dos variables, una con las contrataciones del mes anterior y otra para las contrataciones de los últimos tres meses anteriores.

- ***Modificaciones de Servicios**:* Esta es una variable discreta, que tiene por objetivo almacenar todas las modificaciones que los clientes hacen en sus servicios, como por ejemplo cambios de velocidad, planes, baja de planes de canales, cambios de equipos, etc. Esta variable se considera para las cuatro áreas de negocio analizadas y al igual que en atributos anteriores, se desprende una variable con las modificaciones hechas el mes anterior y otra variable con las modificaciones de los últimos tres meses anteriores.

- ***Antigüedad de los Servicios:*** Esta es una variable discreta que almacena la cantidad de tiempo en meses que un cliente posee un servicio. Se considera en las cuatro áreas de negocio que se analizaron en el proyecto.

Una vez adquirida y analizadas estas variables, que sirvieron como inicio al proceso de colección de datos, se comenzó con el proceso de la búsqueda de ellas, con el fin de realizar una exploración inicial de los datos junto con una verificación de la calidad.

Para ello, se generaron ticket de requerimiento a la subgerencia de informática de Telefónica del Sur, al área de sistemas y administración del data warehause. Conforme se encontraba la información se realizaban dos tareas, primero se almacenaba la ruta de acceso a ellos, para una posterior carga y extracción de los datos a la herramienta de desarrollo (SQL Server 2005) y segundo se descargaban los datos en archivos de tipo excel para realizar las tareas de exploración y verificación de la calidad.

Por otra parte, como se explico anteriormente este conjunto de variables o atributos facilitados por el área de fidelización de clientes de Telefónica del Sur, forman solamente, una parte del conjunto total de variables analizadas por el software. Este conjunto de variables son principalmente variables de comportamiento, de calidad de servicio y de entorno.

Para poder completar el conjunto total de datos, se debe incorporar otro grupo de variables que tienen una importancia significativa para este tipo de proyectos, estas son las variables socio demográficas, quienes intentan retratar aspectos personales del clientes, siendo fundamental a la hora de conocer a la persona que está detrás del

clientes, con sus características, formas y condiciones, muchas veces muy distinta unas de otras.

Después de un largo proceso de búsqueda e indagación de los datos, en donde se hizo gran hincapié en la verificación de la calidad, evitando considerar datos incompletos, se llegó a la conclusión que las variables socio demográficas que posee Telefónica del Sur de sus clientes, y que pueden aportar al desarrollo del software son:

- Zona de Residencia
- Localidad de Residencia
- Edad

Junto con estos atributos, se incorporó una nueva variable que tiene por objetivo conocer que tan ligado esta un cliente a la empresa en cuanto a la contratación de los servicios de un mismo tipo, la variable es:

- Cantidad de productos que un cliente posee de un servicio

Con este atributo se espera saber que tan probable es la fuga de un cliente que solo posee un producto a otro cliente que posee tres o más productos de un mismo servicio.

Terminado este proceso, se recolectó el conjunto de datos inicial, el cual sirvie como base para dar comienzo a la etapa de preparación de los datos. Este conjunto de datos es la materia prima en la construcción de los modelos de minerías de datos.

5.3 Preparación de los Datos

En esta fase de preparación de los datos, se cubrieron todas las actividades necesarias para construir el conjunto de datos final o datos de aprendizaje. Estas tareas de preparación de datos fueron realizadas muchas veces y no en cualquier orden prescripto, es decir, muchas de las tareas explicadas aquí, se ejecutaran luego de una iteración entre las etapas de modelado y esta etapa de preparación de datos. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

La primera tarea ejecutada, fue finalizar el proceso de selección de los datos y la conformación de las tablas de escenarios. Para ello se considera junto a los datos o atributos investigados en la etapa anterior, un conjunto de datos que permiten identificar claramente al cliente. Esto será de vital importancia para la gestión y construcción de reportes con los resultados de la predicción.

Los datos considerados en las tablas son los siguientes:

- Rut del cliente
- Nombre
- Área de negocio que se está analizando (TPOS, ADSL, WITV, PHS)
- Fono asociado al servicio
- Mes del análisis
- Año del análisis
- Estado del Clientes (Cliente vigente ó Fugado)

El origen de estos datos se encuentran en dos tablas que serán fundamentales a la hora de construir las tablas de escenarios, estas son la tabla que contiene la cartera total de clientes y la tabla que almacena a todos los clientes que se han fugado de la empresa.

La tabla cartera de clientes es una tabla generada por una vista que tiene como objetivo entregar todos los clientes vigentes que posee la empresa en la fecha de ejecución. Como se ve en la figura 10, estos datos servirán para analizar el comportamiento de los clientes que permanecen en la empresa hasta un mes específico. Junto con esto, en la figura 10 también se explica el funcionamiento de la tabla de fugados, tabla que también es generada por una vista que reconoce a todos los clientes que han abandonado la empresa hasta la fecha de ejecución. Al igual que la tabla anterior, estos datos permiten analizar el comportamiento de los clientes fugados en los meses previos al mes de la fuga.

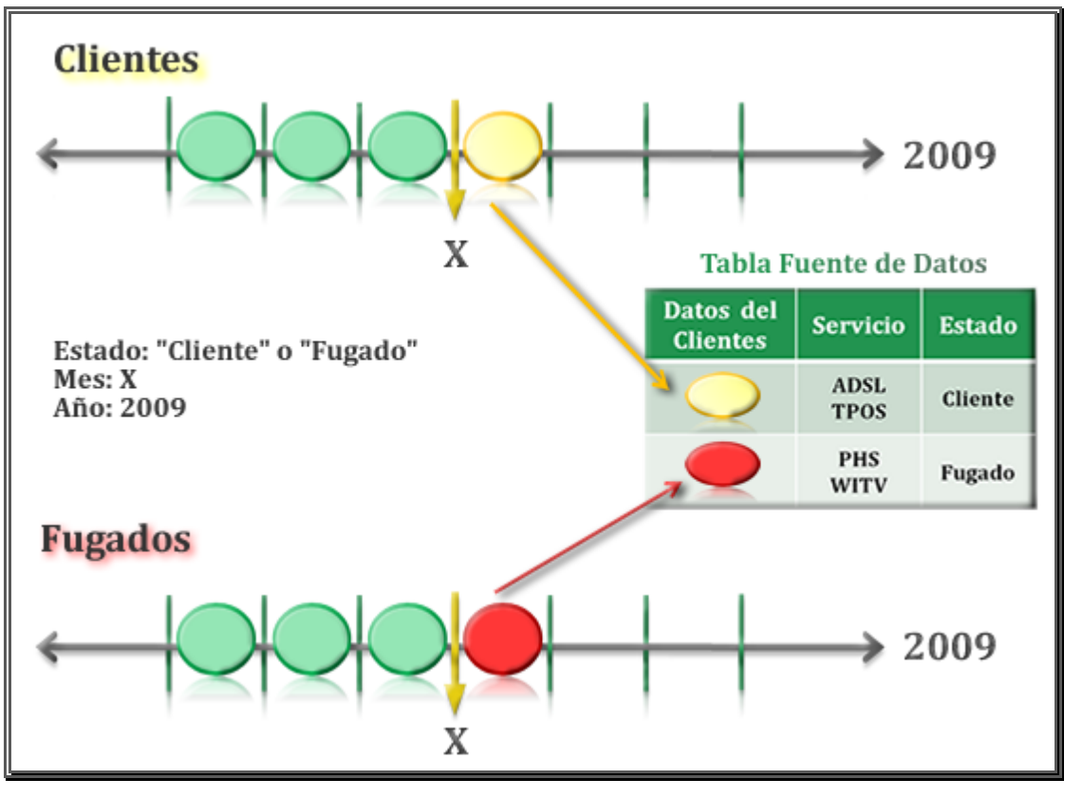


Figura 10: Creación de las tablas de Minería de Datos.

En la figura 10 se observa la política de construcción de las tablas de minería de datos, se consta de dos tablas, la tabla de clientes vigentes y la tabla de clientes fugados, se considera un mes X para el análisis, se seleccionan todos los clientes que pertenecen a la empresa hasta el mes X del año 2009 y a todos los clientes que se fugaron en el mes X del año 2009.

Una vez obtenido los datos de los clientes y los fugados, se separan por área de negocio (ADSL, TPOS, PHS, WITV), si un cliente vigente posee más de un área de negocio, será incluido en todas la tabla que sea cliente, es decir, si un clientes tiene servicios contratados en las cuatro área de negocio analizadas, sus datos se encontrará presente en todas las tablas de escenario. A continuación, como se muestra en las tablas 2,3,4 y 5 se ingresan los datos junto con los atributos de identificación del cliente, estos son Rut del cliente, nombre y fono asociado al área de negocio, además del mes del análisis X y el año del análisis, que para este software, solamente se consideró el año 2009. Finalmente se ingresa el Estado del Cliente, información que determinará si un cliente es vigente o fugado, y será el atributo de mayor relevancia en las tablas de escenarios, pues corresponde a la única variable predictiva del software.

Tabla 2: Datos inicial para la tabla de Telefonía Post Pago (TPOS).

Nombre	Rut	Fono	Área de negocio	Mes	Año	Estado
Cliente 1	x.x.y-z	212121	TPOS	X	2009	Cliente
Cliente 2	x.x.y-z	482120	TPOS	X	2009	Fugado

Tabla 3: Datos inicial para la tabla de Internet Banda Ancha (ADSL).

Nombre	Rut	Fono	Área de negocio	Mes	Año	Estado
Cliente 1	x.x.y-z	DM000054453	ADSL	X	2009	Cliente
Cliente 2	x.x.y-z	DM000100931	ADSL	X	2009	Fugado

Tabla 4: Datos inicial para la tabla de Televisión Digital Inalámbrica (WITV).

Nombre	Rut	Fono	Área de negocio	Mes	Año	Estado
Cliente 1	x.x.y-z	TV00028815	WITV	X	2009	Cliente
Cliente 2	x.x.y-z	TV00015300	WITV	X	2009	Fugado

Tabla 5: Datos inicial para la tabla de Telefonía Local Inalámbrica (PHS).

Nombre	Rut	Fono	Área de negocio	Mes	Año	Estado
Cliente 1	x.x.y-z	572232	PHS	X	2009	Cliente
Cliente 2	x.x.y-z	521520	PHS	X	2009	Fugado

Una vez que se realizó esta tarea, se incorporaron las variables o atributos explicados anteriormente, piezas claves en el resultado o eficiencia de los modelos de minería de datos. Para agregar esta información en las tablas, limpiar los datos y modificar la información, transformándola en variables continuas o discretas, según sea el caso, se crearon un conjunto de procedimientos almacenados en SQL Server 2005.

Un procedimiento almacenado es un elemento de base de datos reutilizable que realiza alguna operación en la base de datos, contiene código SQL que puede, entre otras cosas, insertar, actualizar o eliminar registros de una tabla.

Este conjunto de procedimientos, que forman en su totalidad 116 procedimientos, se programan para ser ejecutados todos de una vez, la función de este proceso es agregar un mes de información a las tablas de minería de datos. La información que ingresa a las tablas corresponde al conjunto total de variables o atributos que serán analizados por el modelo para obtener los patrones de fuga en las áreas de negocio. Toda esta información histórica que comprende el mes actual, últimamente ingresado, y todos los meses anteriormente analizados se obtienen de un grupo de tablas que almacenan la información necesaria para conocer el comportamiento de los clientes, conocer también los clientes vigentes, los clientes fugados y los datos de identificación del cliente, estas tablas son actualizadas mensualmente por medio de la carga automática de los datos utilizando la herramienta Integration Services de SQL Server 2005.

Los procedimientos almacenados que fueron creados para generar las tablas de escenarios son explicados en el anexo 1: “Procedimientos almacenados”.

La larga lista de procedimientos, que cubre prácticamente la totalidad de las tareas realizadas en la etapa de preparación de los datos, tiene la funcionalidad de agregar un mes más de información a la información histórica que poseen las tablas de escenarios. Estas tablas son la materia prima para crear los modelos de minería de datos que descubrirán los patrones de fuga en los servicios analizados.

5.3.1 Problemas encontrados en los Datos

Durante la fase de preparación de los datos que serán provistos en las herramientas de modelado, se encontraron una serie de inconvenientes en los datos brutos iniciales, estos problemas detectados que afectan la calidad de los mismos, fueron manejados de tal forma que permitan sacar el mayor potencial de ellos y limitar lo menos posible la calidad del modelo.

Los inconvenientes encontrados durante la presente fase son los siguientes:

- ***Visitas a Oficinas Comerciales:*** En esta tabla la visita registrada no está asociada al servicio específico por el cual el cliente acude a las oficinas comerciales, sino al conjunto total de servicios que posee el cliente. Este problema disminuye la calidad de los datos, porque en muchos casos el cliente posee más de un servicio, no siendo todos ellos motivo de conflicto. Por otra parte, en ocasiones el cliente se dirige a las oficinas comerciales debido a un descontento general de sus servicios, en cuanto a costos o calidad de estos. Por estos motivos, la forma en que fue trabajada esta variable es agregando la visita del cliente en cada una de las tablas de escenario asociadas al área de negocio que el cliente posea.

- ***Llamadas a la Call Center:*** En este caso, el problema se genera debido a que los datos que registran una llamadas a call center, solamente conocen el número de origen de la llamada (Tpos o Phs), por lo tanto, no se tiene certeza por el área de negocio al cual estaría asociada la llamada del cliente, además una gran parte de las llamadas es desde un celular externo a la compañía, desconociendo completamente al cliente que realiza la llamada. En base a esta situación que afecta gravemente la calidad del dato no pudo ser incluida esta variable en las tablas de escenarios.

- ***Edad de los Clientes:*** Esta variable socio – demográfica es muy importante considerarla, pues el comportamiento de los clientes puede variar en etapas distintas de sus vidas, por ejemplo, un cliente de mayor edad puede tener mayor tolerancias a un servicio que presenta fallas que un cliente joven. El problema en este dato surge al no encontrarse almacenada la fecha de nacimiento del cliente, siendo imposible determinar su edad, para solucionarlo se estima la edad de acuerdo a la numeración de su Rut, teniendo esta medida algún grado de error que puede afectar la calidad del dato.

5.4 Modelado

En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos, principalmente en la utilización de variables discretas o continuas, por tanto fuen necesario volver a la fase de preparación de datos en innumerables ocasiones. Debido a esto, muchos resultados expuestos en la fase anterior fueron obtenidos luego de iterar de la presente fase.

Para el desarrollo del modelo, comenzó a tomar mayor importancia la herramienta utilizada. Para este proyecto, como se explico anteriormente, se utilizó Microsoft SQL Server 2005, quien ofrece cuatro herramientas principales, ellas son Analysis Service, Motor de Base de Datos, Integration Services y Reporting Service.

De estas herramientas, SQL Server Analysis Services fue la que tuvo mayor protagonismo en el desarrollo del modelo de minería de datos, provee un entorno integrado para crear modelos y trabajar con ellos. En este entorno de desarrollo se encuentra SQL Server Management Studio, herramienta principal para administrar el motor de bases de datos y escribir código de Transact-SQL y SQL Server Business Intelligence Development Studio, que es el Microsoft Visual Studio con tipos de proyecto adicionales específicos de Business Intelligence.

En resumidas cuentas, en este entorno integrado de desarrollo se construyó el origen de datos, la estructura de minería de datos, la elección del algoritmo, la determinación de los parámetros, el entrenamiento del modelo y las consultas de predicción a los modelos creados.

A continuación se explicaran detalladamente los pasos seguidos para la creación de un modelo de minería de datos y la utilización de las herramientas de desarrollo en cada etapa del proceso.

5.4.1 Selección de las Técnicas de Modelado

Como primer paso en esta fase de modelado, se realizó la selección de las técnicas de modelado y la función de cada una de ellas. Unas de las primeras necesidades fue saber cuáles son las variables que tienen mayor influencia en la variable predictiva, para ello se aplicó el algoritmo de Bayes Naive de Microsoft, este algoritmo como se explico anteriormente, es un algoritmo de clasificación que calcula la probabilidad condicional entre las columnas de entradas y la columna de predicción, descubriendo relaciones entre ellas. Este algoritmo se utilizó para realizar una exploración inicial de los datos y, más adelante, aplicar los resultados obtenidos para crear modelos de minería de datos con otros algoritmos más complejos y precisos desde el punto de vista computacional, como son el algoritmo de Redes Neuronales de Microsoft y el algoritmo de Árboles de Decisión de Microsoft.

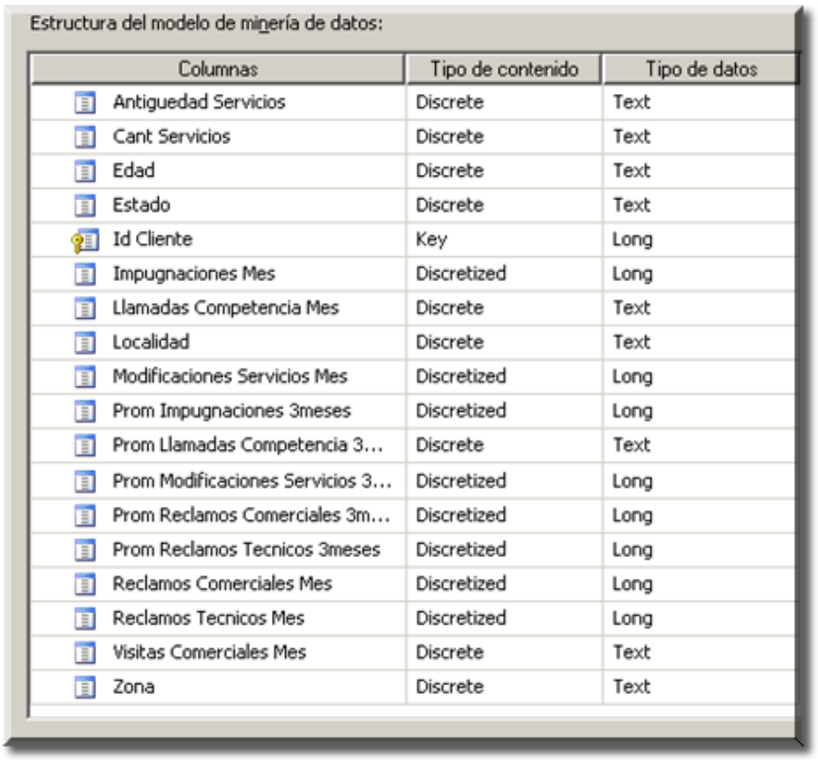
5.4.2 Selección de las Variables Influyentes

Como se acaba de explicar, para seleccionar las variables de entrada que tienen mayor influencia en la variable predictiva, variable “Estado” (Cliente o Fugado), se aplica el algoritmo de Bayes Naives de Microsoft. Este algoritmo actúa calculando la probabilidad de cada estado de cada columna de entrada, dado cada posible estado de la columna de predicción.

Este método es importante no sólo porque ofrece un análisis cualitativo de los atributos y valores que pueden intervenir en el problema, sino porque da cuenta también de la importancia cuantitativa de esos atributos. En el aspecto cualitativo se puede representar cómo se relacionan esos atributos ya sea en una forma causal, o señalando simplemente la correlación que existe entre esos atributos. Cuantitativamente (y ésta es la gran aportación de los métodos bayesianos), da una medida probabilística de la importancia de esas variables en el problema (y por lo tanto una probabilidad explícita de las hipótesis que se formulan). Esta es quizá una de las diferencias fundamentales que ofrecen las redes bayesianas con respecto a otros métodos -como puedan ser los árboles de decisión y las redes neuronales, que no dan una medida cuantitativa de esa clasificación [Mal03].

Como requisito, un modelo Bayes naive debe contener una columna de clave, columnas de entrada y una columna de predicción, en donde, todas las columnas deben ser

discretas. Esta tarea se realiza durante el proceso de construcción de la estructura de minería de datos. Como se aprecia en la figura 11, al momento de seleccionar el conjunto de variables con el que se entrenó el modelo, la herramienta discretiza las variables continuas asignándole estados a cada una de ellas, de modo que el algoritmo pueda obtener las relaciones entre estos estados y los estados de la variable predictiva.


Estructura del modelo de minería de datos:

Columnas	Tipo de contenido	Tipo de datos
Antiguedad Servicios	Discrete	Text
Cant Servicios	Discrete	Text
Edad	Discrete	Text
Estado	Discrete	Text
Id Cliente	Key	Long
Impugnaciones Mes	Discretized	Long
Llamadas Competencia Mes	Discrete	Text
Localidad	Discrete	Text
Modificaciones Servicios Mes	Discretized	Long
Prom Impugnaciones 3meses	Discretized	Long
Prom Llamadas Competencia 3...	Discrete	Text
Prom Modificaciones Servicios 3...	Discretized	Long
Prom Reclamos Comerciales 3m...	Discretized	Long
Prom Reclamos Tecnicos 3meses	Discretized	Long
Reclamos Comerciales Mes	Discretized	Long
Reclamos Tecnicos Mes	Discretized	Long
Visitas Comerciales Mes	Discrete	Text
Zona	Discrete	Text

Figura 11: Especificación del contenido y el tipo de datos de las columnas utilizadas en la estructura de minería de datos para el algoritmo Bayes Naive.

En la columna tipo de contenido, se puede apreciar como las variables continuas, específicamente las que poseen tipo de dato Long, son dejadas en un estado Discretized, esto quiere decir que antes de iniciar el entrenamiento de los datos, los valores continuos serán discretizados, cumpliéndose así el requisito del algoritmo.

Ahora bien, una vez creada la estructura de minerías de datos se deben establecer los parámetros del algoritmo. En la figura 12, se observa con claridad como SQL Server Business Intelligence Development Studio ofrece un visor con los parámetros del algoritmo, permitiendo modificar los valores por defecto y agregar nuevos parámetros según la necesidad de la solución.

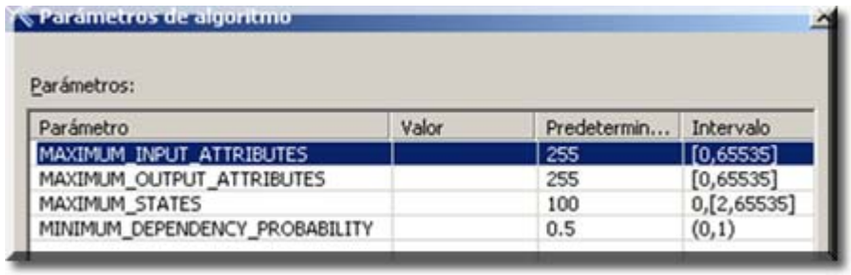


Figura 12: Visor para establecer o modificar los parámetros del algoritmo Bayes Naive.

Cada uno de los parámetros influye en el rendimiento y la precisión del modelo de minería de datos resultante. A continuación, la tabla 6 describe la funcionalidad de los parámetros utilizados.

Tabla 6: Descripción de los parámetros utilizados en la aplicación del algoritmo de Bayes Naives.

Parámetro	**Descripción**
MAXIMUM_INPUT_ATTRIBUTES	Especifica el número máximo de atributos de entrada que puede administrar el algoritmo antes de invocar la selección de características. Si este valor se establece en 0, se deshabilita la selección de características para atributos de entrada. El valor predeterminado es 255.
MAXIMUM_OUTPUT_ATTRIBUTES	Especifica el número máximo de atributos de salida que puede administrar el algoritmo antes de invocar la selección de características. Si este valor se establece en 0, se deshabilita la selección de características para atributos de salida. El valor predeterminado es 255.
MINIMUM_DEPENDENCY_PROBABILITY	Especifica la probabilidad de dependencia mínima entre los atributos de entrada y salida. Este valor se utiliza para limitar el tamaño del contenido generado por el algoritmo. Puede establecerse un valor de 0 a 1 para esta propiedad. Si se aumenta el valor, se reduce el número de atributos del modelo. El valor predeterminado es 0,5.
MAXIMUM_STATES	Especifica el número máximo de estados de atributo que admite el algoritmo. Si el número de estados que tiene un atributo es mayor que el número máximo de estados, el algoritmo utiliza los estados más conocidos del atributo e interpreta que faltan los estados restantes. El valor predeterminado es 100.

Una vez que la estructura y los parámetros para los modelos de minería de datos se completaron, se entrenó el modelo de minería de datos. En la figura 13 se aprecia el modelo creado, el análisis se realizó utilizando la ficha llamada Visor de modelos de minería de datos del Diseñador de minería de datos.

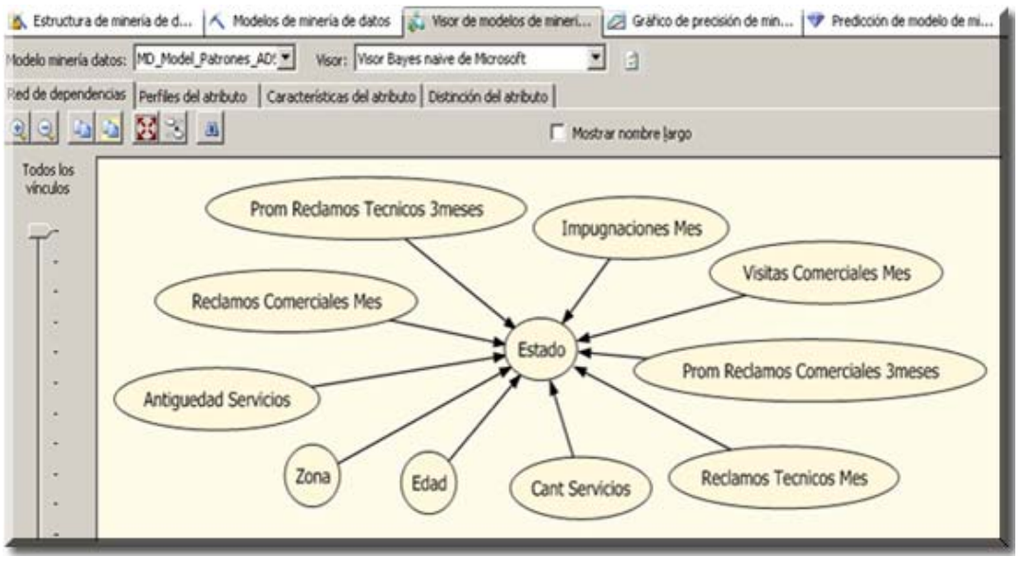


Figura 13: Ficha Red de dependencia del Visor de Bayes Naive para el modelo de ADSL.

El Visor de Bayes Naive proporciona varios métodos para mostrar la interacción entre los atributos de predicción y los atributos de entrada de una tabla de escenarios.

Como se apreció en la figura 13, una de las fichas que dispone el visor es la Red de dependencia, en ella se observan claramente las dependencias entre los atributos de entrada y los atributos de predicción de un modelo. El control deslizante de la izquierda del visor se comporta como un filtro que está asociado a la importancia de las dependencias. Si desplaza el control deslizante hacia abajo, sólo se verán los vínculos más similares.

Cuando se selecciona un nodo, el visor resalta las dependencias específicas de dicho nodo. Por ejemplo, si elige un nodo de predicción, el visor también resalta cada uno de los nodos que ayudan a predecir el nodo de predicción.

Otra ficha importante que ofrece el visor es Perfiles del atributo. En la figura 14, se observan las columnas de entrada del conjunto de datos y el cálculo de la probabilidad de cada estado de las columnas de entrada, dado cada posible estado de la columna de

predicción, indicando por medio de histogramas, cómo se distribuyen los estados de cada columna, dado cada estado de la columna de predicción. Al posicionarse sobre un histograma se puede observar una leyenda que indica el porcentaje de influencia que tiene cada estado de la variable de entrada sobre el estado de la variable de predicción.

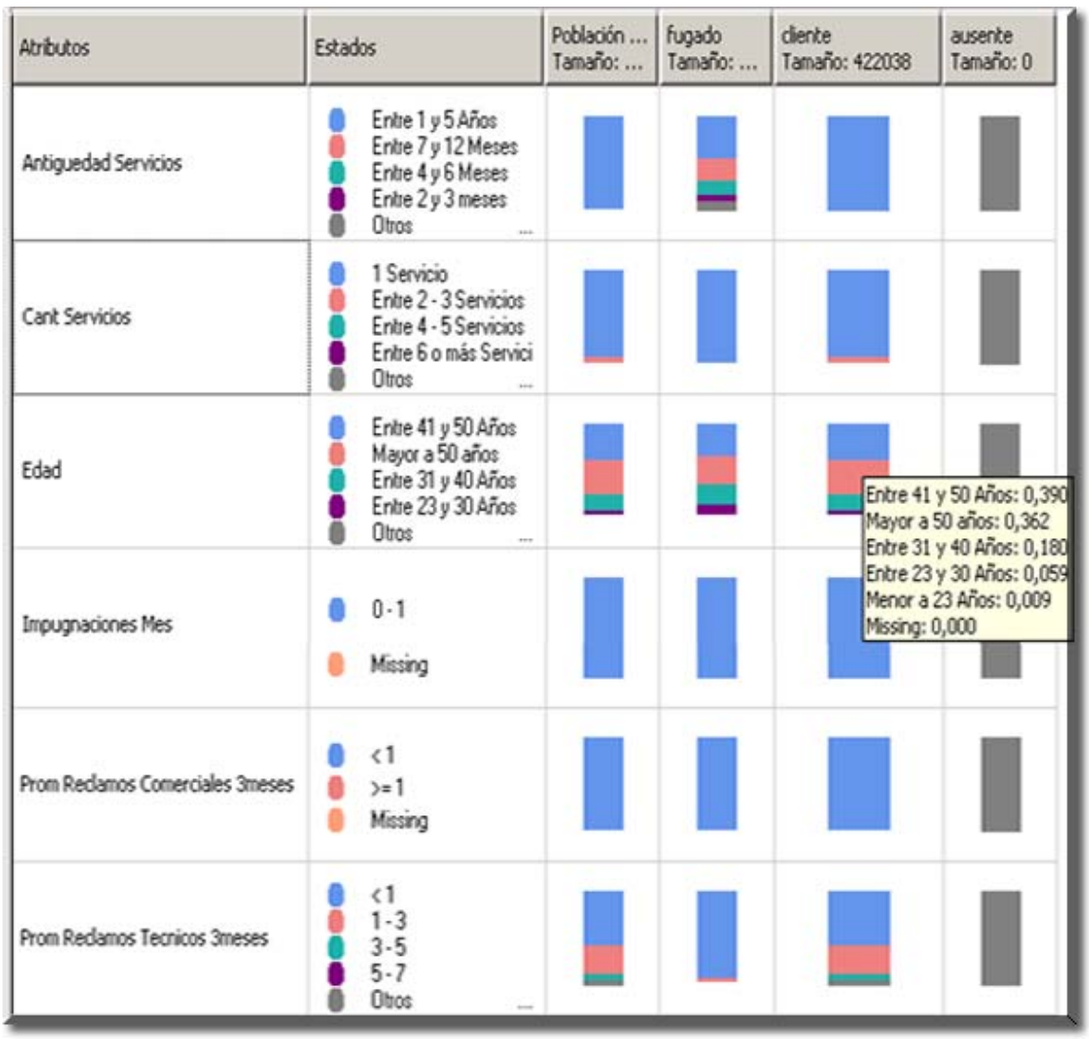


Figura 14: Perfiles de los atributos del Visor de Bayes Naive para el modelo de ADSL.

Esta vista, junto con la red de dependencia, se utilizó para identificar las columnas de entradas que son importantes a la hora de generar los modelos de minería de datos con los algoritmos de árboles de decisión y redes neuronales.

En las figuras 15, 16 y 17 se muestran las redes de dependencias obtenidas al crear los modelos de minería de datos para TPOS, WITV y PHS respectivamente.

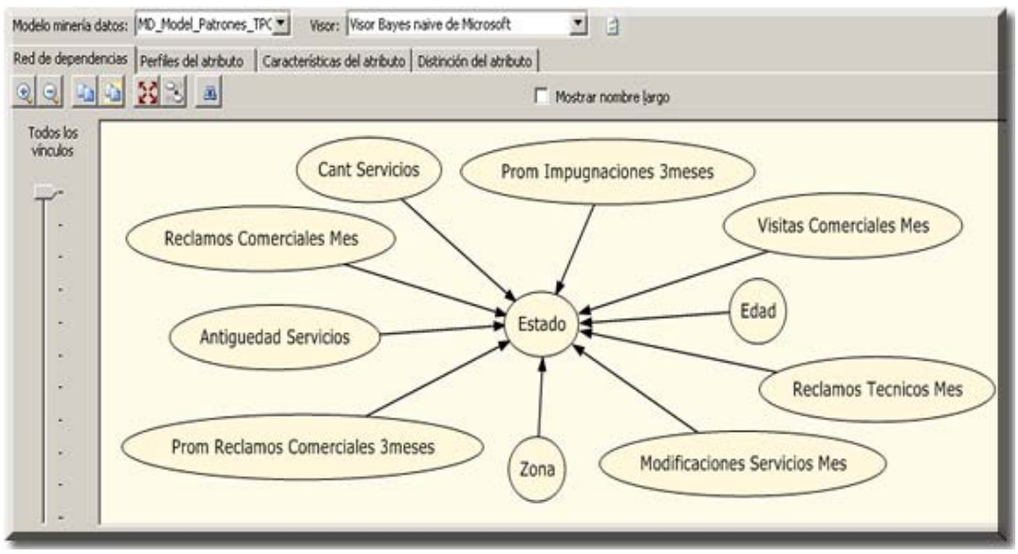


Figura 15: Ficha Red de dependencia del Visor de Bayes Naives para el modelo de TPOS.

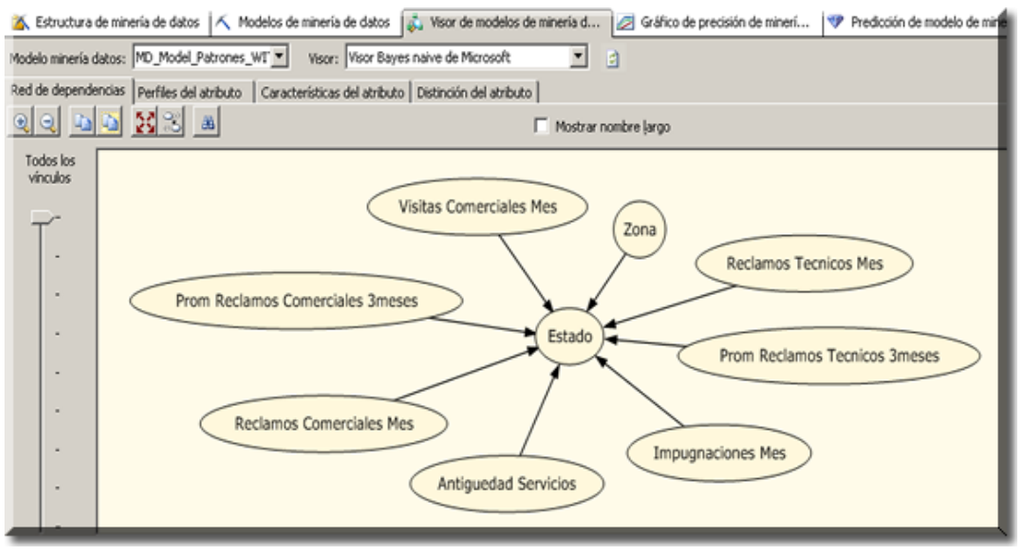


Figura 16: Ficha Red de dependencia del Visor de Bayes Naives para el modelo de WITV.

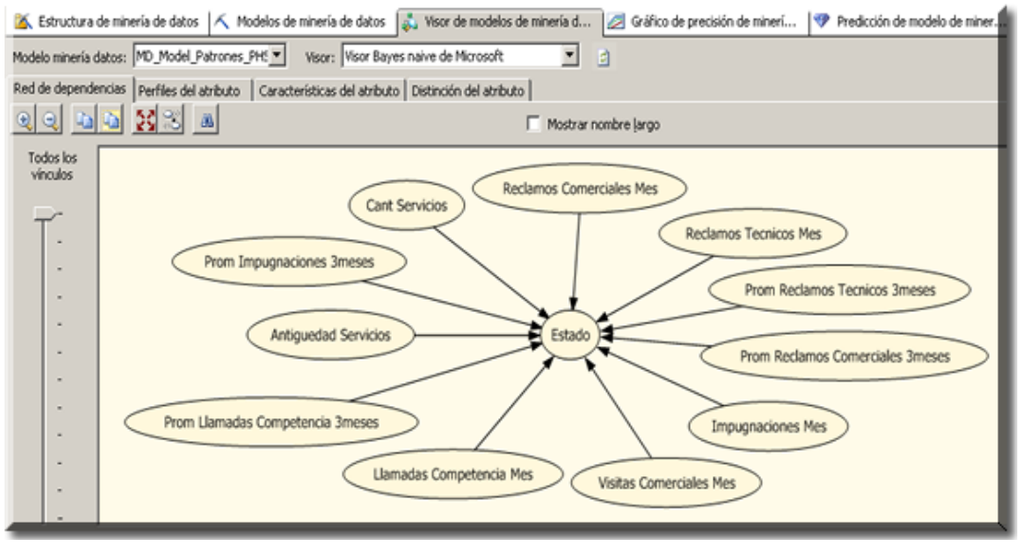


Figura 17: Ficha Red de dependencia del Visor de Bayes Naives para el modelo de PHS.

5.4.3 Construcción del Modelo

Una vez que se realizó la exploración inicial de los datos, conociendo las variables de entrada que poseen mayor influencia en la variable predictiva de cada servicio, fue el momento de ejecutar la herramienta de modelado sobre el conjunto de los datos preparados para crear los modelos. De los cuatro conjuntos de datos distintos que se crearon, uno para cada servicio analizado, se construyeron dos modelos de minería de datos, unos con la utilización del algoritmo de Árboles de Decisión de Microsoft y otro con el algoritmo de Redes Neuronales de Microsoft, generando en total ocho modelos, dos para cada servicio, modelos que posteriormente serán analizados y comparados para conocer cual manifiesta mejores resultados.

Como primera tarea se ejecuta la herramienta Business Intelligence Developmen Studio, en ella se encuentra el asistente para minería de datos de Microsoft, este se ejecuta cada vez que se agrega una nueva estructura de minería de datos a un proyecto de minería de datos. El asistente define nuevas estructuras y también define el modelo de minería de datos inicial para cada estructura.

5.4.3.1 Modelado con el algoritmo de Árboles de Decisión

El algoritmo de árboles de decisión de Microsoft es un algoritmo de clasificación que proporciona SQL Server Analysis Services para modelar la predicción de atributos discretos y continuos.

Como es sabido, este proyecto posee un único atributo de predicción de tipo discreto, por lo tanto, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada para predecir los estados del atributo de predicción.

Para iniciar la construcción de este modelo se selecciona un origen de datos, es decir, se realiza una conexión con la base de datos en donde se encuentran las tablas de escenario, tablas que incluyen los datos de aprendizaje que se utilizaran para entrenar los modelos de minería de datos. Una vez realizada esta conexión se configura la vista de origen de datos, aquí se escoge la tabla de escenario que se utiliza para la creación de la estructura de minería de datos.

Los pasos necesarios para crear la estructura de de minería de datos son los siguientes:

1. Seleccionar un tipo de origen de datos, en este caso una base de datos relacional existente.
2. Seleccionar el algoritmo o la técnica de minería de datos.
3. Seleccionar la vista del origen de datos para proporcionar los datos que la estructura necesita.
4. Seleccionar la tabla de escenarios.
5. Especificar las columnas que se utilizaran para el análisis: columnas de entrada, de predicción o de clave.
6. Especificar el contenido y el tipo de datos de las columnas de la estructura de minería de datos.
7. Asignar nombre y guardar la nueva estructura y el modelo de minería de datos asociado.

En la figura 18 se puede apreciar el punto 5 del proceso, aquí se especifican las columnas que se utilizaran para el análisis. En este ejemplo, se muestra la construcción de la estructura de minería de datos para el servicios de Telefonía de Post Pago, utilizando la técnica de árboles de decisión.

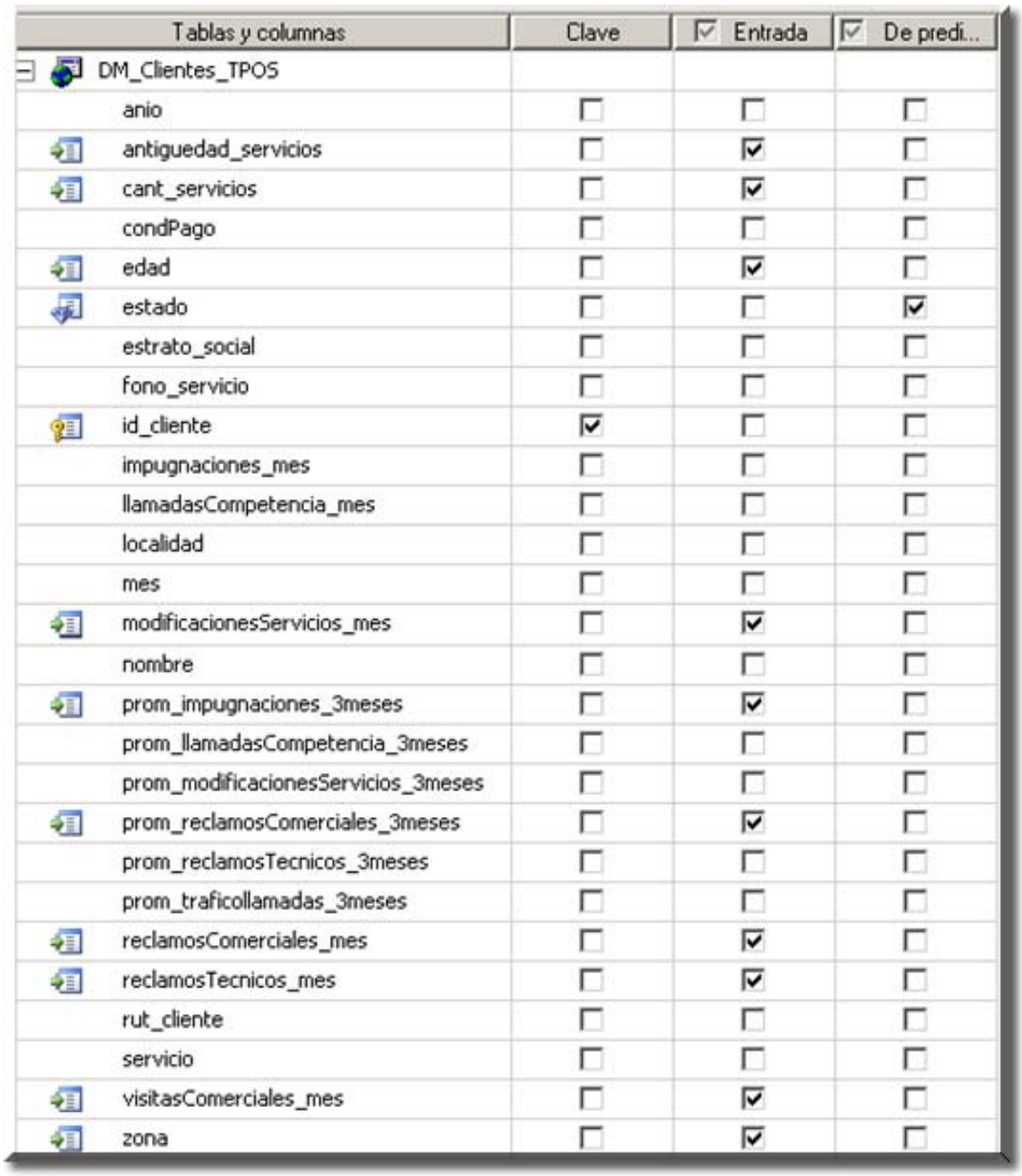


Figura 18: Especificación de las columnas que se utilizan para la creación de la estructura de TPOS.

Una vez finalizado el asistente con el que se crea la estructura de minería de datos, en el explorador de soluciones de la herramienta Business Intelligence Development Studio, como se ve en la figura 19, se almacena el proyecto de Analysis Services que contiene el origen de datos, la vista del origen de datos y la estructura de minería de datos recientemente creada.

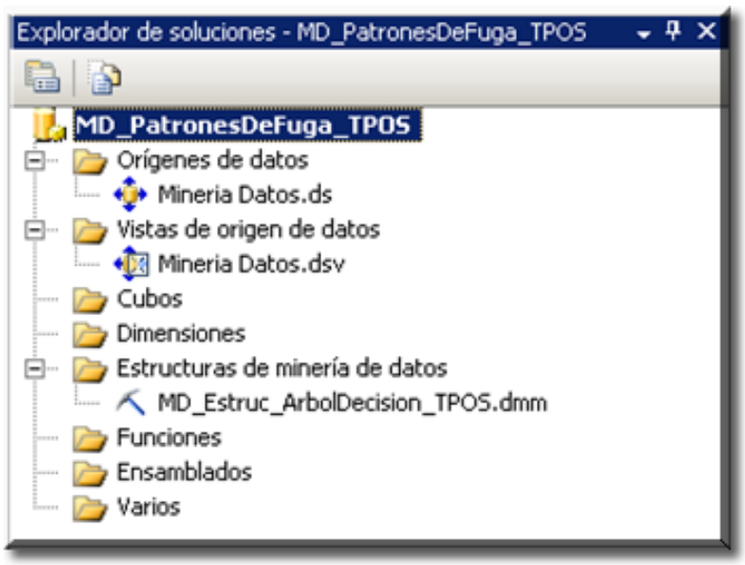
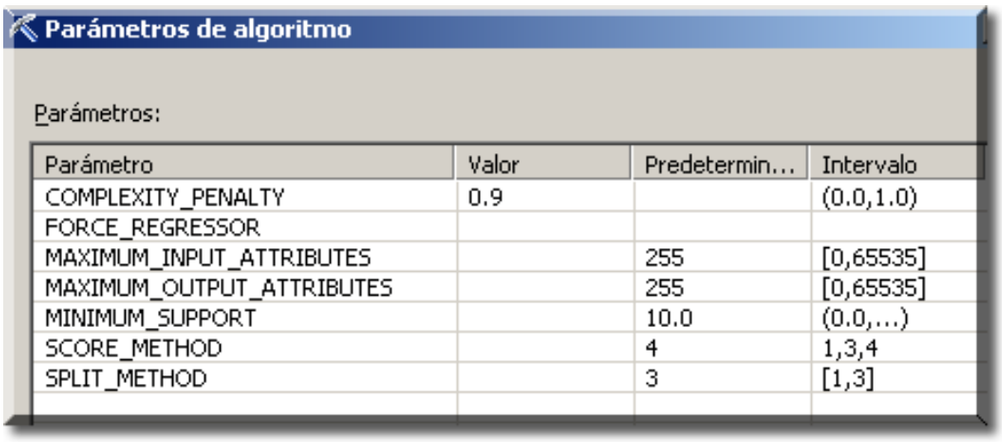


Figura 19: Explorador de soluciones con el cubo que contiene la solución para TPOS.

Antes de iniciar el entrenamiento del modelo se deben configurar los parámetros del algoritmo, quienes afectan al rendimiento y la precisión del modelo de minería de datos. En la figura 20, se aprecia el visor que facilitó la herramienta Business Intelligence Development Studio para especificar los parámetros.



Parámetro	Valor	Predetermin...	Intervalo
COMPLEXITY_PENALTY	0.9		(0.0,1.0)
FORCE_REGRESSOR			
MAXIMUM_INPUT_ATTRIBUTES		255	[0,65535]
MAXIMUM_OUTPUT_ATTRIBUTES		255	[0,65535]
MINIMUM_SUPPORT		10.0	(0.0,...)
SCORE_METHOD		4	1,3,4
SPLIT_METHOD		3	[1,3]

Figura 20: Determinación de los parámetros del algoritmo de arboles de decisión para el modelo de TPOS.

La tabla 7, describe la funcionalidad de cada parámetro que se configuró para la construcción del modelo de minería de datos utilizando la técnica de arboles de decisión.

Tabla 7: Descripción de los parámetro utilizados para el algoritmo de arboles de decisión.

Parámetro	Descripción
MAXIMUM_INPUT_ATTRIBUTES	Define el número de atributos de entrada que el algoritmo puede controlar antes de invocar la selección de características. Establezca este valor en 0 para desactivar la selección de características. El valor predeterminado es 255.
MAXIMUM_OUTPUT_ATTRIBUTES	Define el número de atributos de salida que el algoritmo puede controlar antes de invocar la selección de características. Establezca este valor en 0 para desactivar la selección de características. El valor predeterminado es 255.
SCORE_METHOD	Determina el método utilizado para calcular el resultado de la división. Opciones disponibles: Entropía (1), Bayesiano con prioridad K2 (2) o Equivalente Dirichlet bayesiano (BDE) con prioridad (3). El valor predeterminado es 3.
SPLIT_METHOD	Determina el método utilizado para dividir el nodo. Opciones disponibles: binario (1), completo (2) o ambos (3). El valor predeterminado es 3.
MINIMUM_SUPPORT	Determina el número mínimo de escenarios de hoja necesarios para generar una división en el árbol de decisión. El valor predeterminado es 10.
COMPLEXITY_PENALTY	Controla el crecimiento del árbol de decisión. Un valor bajo aumenta el número de divisiones y un valor alto lo reduce. El valor predeterminado se basa en el número de atributos de un modelo concreto, como se describe en la lista siguiente: • De 1 a 9 atributos, el valor predeterminado es 0,5. • De 10 a 99 atributos, el valor predeterminado es 0,9. • Para 100 o más atributos, el valor predeterminado es 0,99.

Una vez que se entrenó y generó el modelo de minería de datos, se utiliza el Visor de Arboles de Microsoft para examinar el modelo creado. Un árbol de decisión está compuesto por una serie de divisiones, con la división más importante que determina el algoritmo a la izquierda del visor, en el nodo *Todo.* Como se ve en la figura 21 y 22, figuras que en conjunto muestran el árbol completo para el modelo de minería de datos para Telefonía Post Pago, la división del nodo *Todos* es la más importante porque

contiene la condición más determinante de división del conjunto de datos y, por tanto, la que ocasiona la primera división.

Para mostrar u ocultar las divisiones que ocurren en cada nodo y así examinar el árbol, se pueden expandir o contraer los nodos individuales. También puede utilizar las opciones de la ficha Árbol de decisión para influir en el modo en que se muestra el árbol.

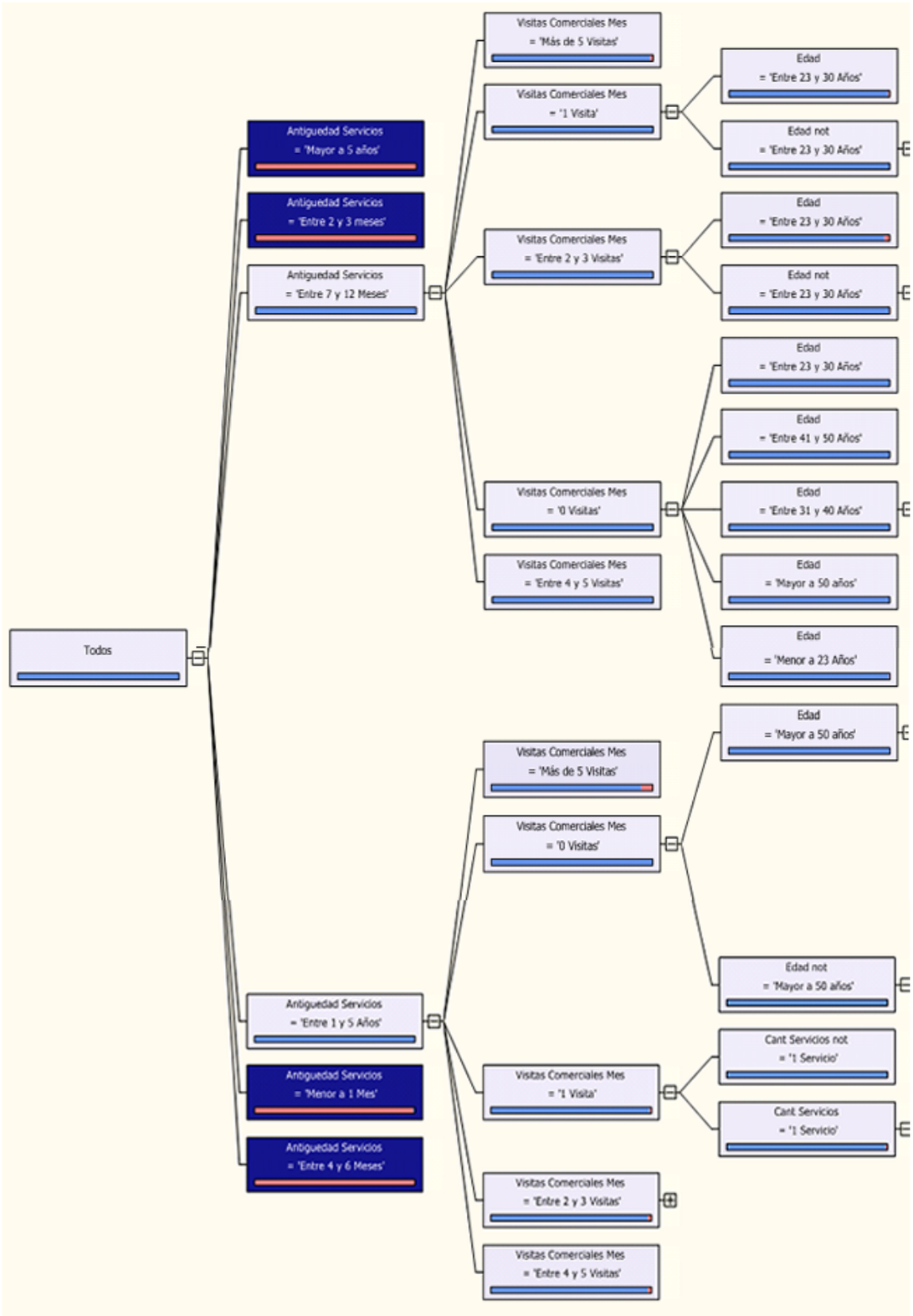


Figura 21: Primera parte del Árbol de decisión del modelo de TPOS.

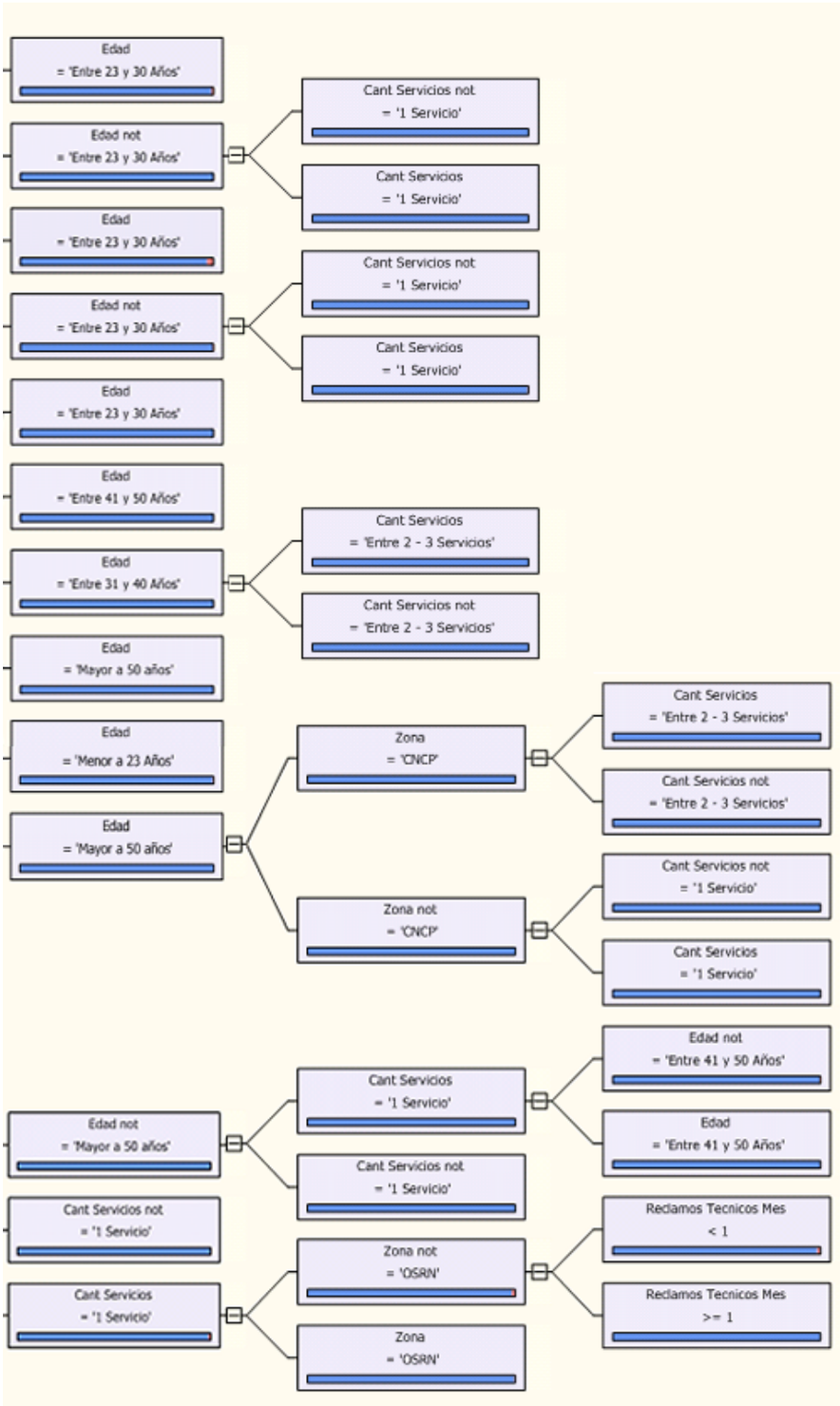


Figura 22: Segunda parte del Árbol de decisión del modelo de TPOS.

5.4.3.2 Modelado con el algoritmo de Redes Neuronales

El algoritmo de red neuronal Microsoft crea modelos de minería de datos de clasificación y regresión mediante la generación de una red de neuronas de tipo Perceptrón Multicapa. De forma similar al algoritmo de árboles de decisión de Microsoft, el algoritmo de red neuronal de Microsoft calcula las probabilidades de cada posible estado del atributo de entrada cuando se da cada estado del atributo de predicción. Posteriormente, puede utilizar estas probabilidades para predecir un resultado del atributo predicho basado en los atributos de entrada.

El proceso de construcción de la estructura y el modelo de minería de datos es idéntico al proceso explicado anteriormente con el algoritmo de arboles de decisión. Se comienza con la conexión a la base de datos que almacena las tablas de escenarios, una vista de la base de datos, seleccionando la tabla de escenario que corresponde al área de negocio analizado, la selección del algoritmo, la selección de las columnas de entrada, clave y predicción, la especificación del contenido y tipo de datos de estas columnas, la asignación de un nombre a la estructura de minería de datos y finalmente, antes de entrenar el modelo, la configuración de los parámetros del algoritmo de Red Neuronal de Microsoft como se aprecia en la figura 23.

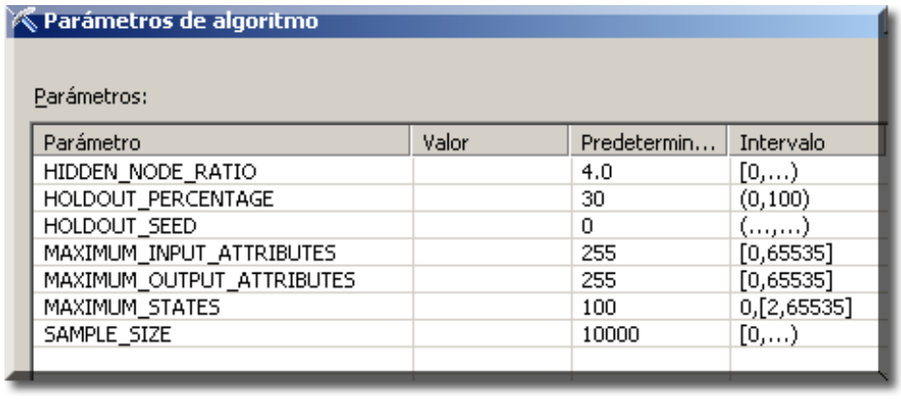


Figura 23: Determinación de los parámetros del algoritmo red neuronal para el modelo de TPOS.

Los valores que se especifican en los parámetros disponibles para el algoritmo, tienen mucha influencia en los pasos implicados en el entrenamiento del modelo de minería de datos. En primer lugar, el algoritmo evalúa y extrae los datos de entrenamiento del origen de datos. Un porcentaje de los datos de entrenamiento, denominado datos de exclusión, se reserva para medir la precisión de la estructura del modelo resultante.

Durante el proceso de entrenamiento, el modelo se evalúa frente a los datos de exclusión después de cada iteración en los datos de entrenamiento. Cuando la precisión del modelo deja de aumentar, el proceso de entrenamiento se detiene. Los valores de los parámetros *SAMPLE_SIZE* y *HOLDOUT_PERCENTAGE* se utilizan para determinar el número de escenarios de muestra de los datos de entrenamiento y el número de escenarios que se apartan para los datos de exclusión. El valor del parámetro *HOLDOUT_SEED* se utiliza para determinar aleatoriamente los escenarios individuales que se apartan para los datos de exclusión.

A continuación, el algoritmo determina el número y la complejidad de las redes que admite el modelo de minería de datos. Si el modelo contiene uno o más atributos que sólo se utilizan para la predicción, el algoritmo crea una única red que representa todos estos atributos. Si el modelo contiene uno o más atributos que se utilizan para la entrada y la predicción, el proveedor de algoritmos construye una red para cada uno de estos atributos. Si el número de atributos de entrada o de predicción es mayor que el valor del parámetro *MAXIMUM_INPUT_ATTRIBUTES* o el parámetro *MAXIMUM_OUTPUT_ATTRIBUTES* respectivamente, se utiliza un algoritmo de selección de características para reducir la complejidad de las redes que se incluyen en el modelo de minería de datos. La selección de características reduce el número de atributos de entrada o de predicción a los más relevantes estadísticamente para el modelo.

En el caso de los atributos de entrada y de predicción que tienen valores discretos, cada neurona de entrada o de salida representa respectivamente un único estado. En el caso de los atributos de entrada y de predicción que tienen atributos continuos, cada neurona de entrada o de salida representa respectivamente el intervalo y la distribución de valores del atributo. El número máximo de estados admitidos en cada escenario depende del valor del parámetro de algoritmo *MAXIMUM_STATES.* Si el número de estados para un atributo específico excede el valor del parámetro de algoritmo *MAXIMUM_STATES*, se eligen los estados más utilizados o relevantes para dicho atributo, hasta alcanzar el máximo; el resto de los estados se agrupa como valores que faltan para el análisis.

A continuación, el algoritmo utiliza el valor del parámetro *HIDDEN_NODE_RATIO* al determinar el número inicial de neuronas que se crearán para la capa oculta. El proveedor de algoritmos evalúa iterativamente el peso de todas las entradas de la red

simultáneamente, tomando el conjunto de datos de entrenamiento reservado anteriormente y comparando el valor real conocido de cada escenario de los datos de exclusión con la predicción de la red, en un proceso conocido como aprendizaje por lotes. Una vez que el algoritmo ha evaluado el conjunto completo de los datos de entrenamiento, revisa el valor predicho y real de cada neurona. El algoritmo calcula el grado de error, si lo hay, y ajusta los pesos asociados con las entradas de esa neurona, trabajando hacia atrás desde las neuronas de salida a las de entrada en un proceso conocido como propagación hacia atrás. A continuación, el algoritmo repite el proceso en todo el conjunto de datos de entrenamiento. Dado que el algoritmo puede admitir múltiples pesos y neuronas de salida, el algoritmo de gradiente conjugado se utiliza para guiar el proceso de entrenamiento en la asignación y evaluación de los pesos de las entradas.

Una vez que el entrenamiento del modelo se ha realizado, el visor de redes neuronales que posee SQL Server Analysis Services muestra el modelo de minería de datos que se genera con el algoritmo de Red neuronal. Puede utilizarse para seleccionar estados concretos de atributos de entrada y para investigar el modo en que los otros atributos de entrada del modelo afectan al estado del atributo de salida, también denominado atributo de predicción.

En la figura 24, se observa el modelo de redes neuronales creado para televisión digital inalámbrica, aquí se puede observar los estados de las variables de entrada más influyente en la variable predictiva. Se designa en la ventana Output el atributo del modelo de red neuronal que se ocupará como salida junto con dos estados que se desean comparar, en este caso, estado cliente y estado fugado. Luego en el panel de variables se disponen las columnas Atributo, Valor del atributo, Favorece [cliente] y Favorece [fugado], de manera que las columnas se ordenan por la importancia que tiene las columnas de entrada sobre el estado Fugado. La barra a la derecha del atributo muestra el estado del atributo de entrada que el estado del atributo de salida favorece. El tamaño de la barra muestra la intensidad con que el estado de salida favorece al estado de entrada.

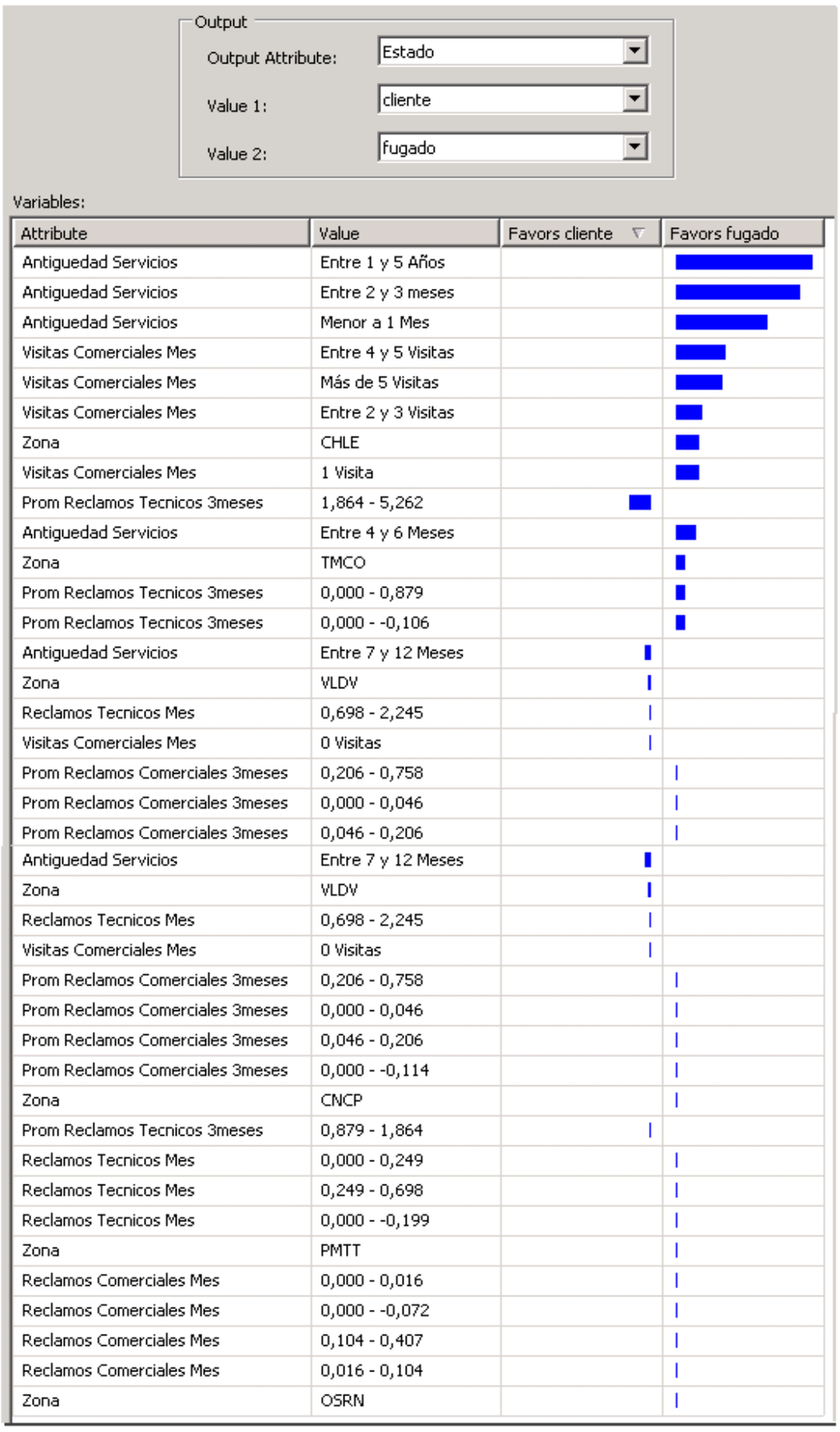


Figura 24: Visor del modelo de redes neuronales para Televisión Digital Inalámbrica.

5.5 Evaluación

En esta etapa del proyecto, ya se construyeron los modelos de minería de datos que parecen tener una alta calidad desde la perspectiva del análisis de datos.

Antes de proceder a la implementación final del modelo, es importante evaluar los modelos sobre aplicaciones reales, para ellos se realizan consultas predictivas al modelo, tomando un conjunto de clientes de la empresa que a la vez están presentes en las tablas de escenarios ocupadas para el entrenamiento del modelo.

El objetivo de este proceso fue conocer la efectividad predictiva de los modelos de minería de datos, descubriendo cuantos clientes que efectivamente se fugaron de la empresa durante el año dos mil nueve, son reconocidos el mes anterior con una probabilidad alta de fuga.

5.5.1 Consultas Predictivas

La evaluación del modelo se llevo a cabo creando un proyecto de Integration Services y programando una Tarea de Consulta de minería de datos, como se aparecía en la figura 25. Esta tarea es una herramienta que ofrece Microsoft Integration Services para realizar consultas a los modelos de minería de datos.

Figura 25: Creación de una consulta al modelo de minería de datos.

Para programar esta tarea, como primera etapa, se debe crear una conexión al proyecto donde se encuentra la estructura de minería de datos, como muestra la figura 26, posteriormente se selecciona el modelo a consultar.

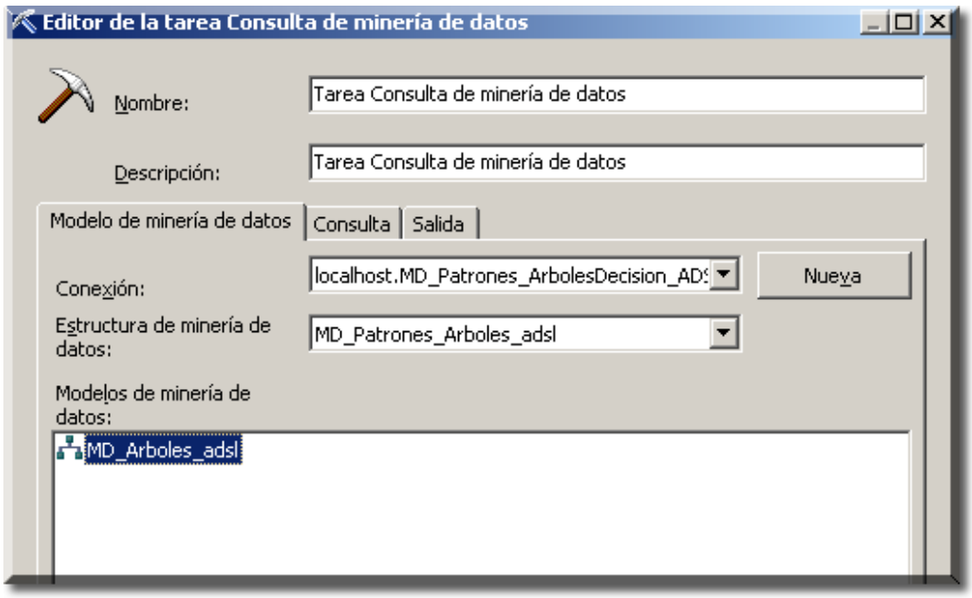


Figura 26: Selección de la estructura y modelo que serán consultados.

Una vez realizados los pasos anteriores, en la pestaña de consulta, se ingresa un conjunto de clientes con la misma estructura de información que poseen las tablas de escenarios, y se selecciona el modelo a consultar. Como se muestra en la figura 27, automáticamente la herramienta relaciona las variables que el modelo necesita, del conjunto de datos de entrada, para realizar la predicción.

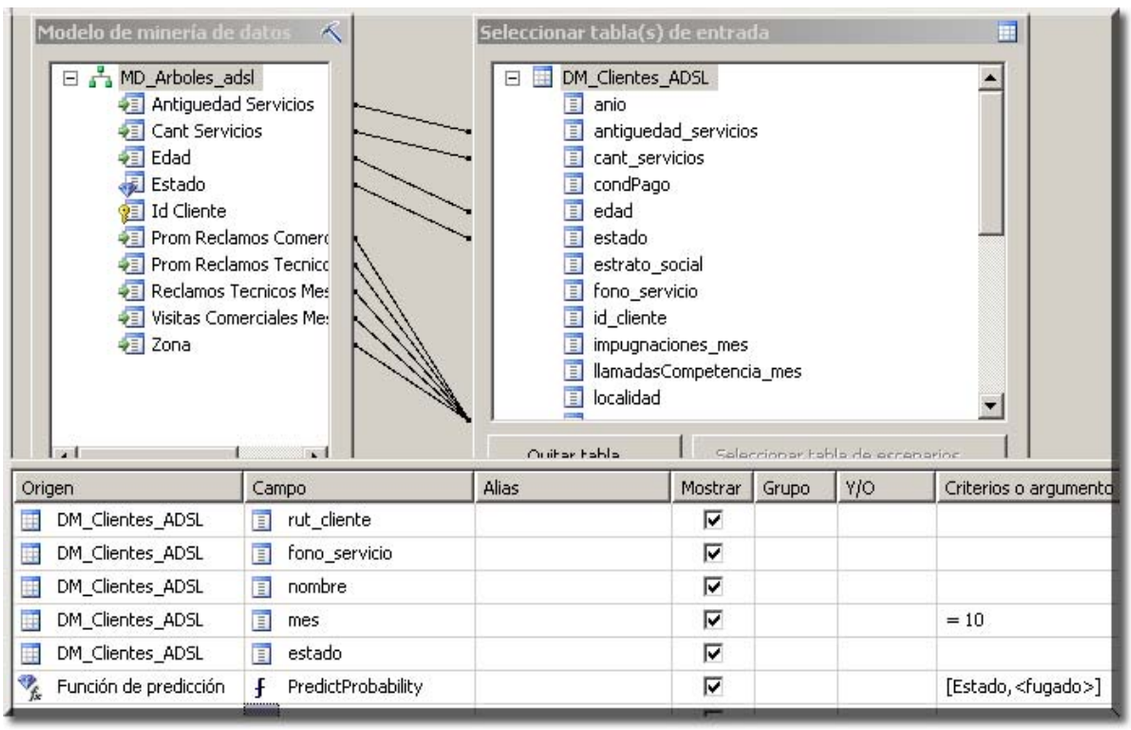


Figura 27: Selección del modelo de minería de datos, la tabla de entrada, las variables desplegadas en la consulta y la función de predicción.

Finalmente, en la figura 28, se observa la salida de la consulta predictiva, aquí se selecciona una base de datos y la tabla donde se almacenan los resultados que podrán ser entregados al departamento de fidelización de clientes por medio de un reporte.

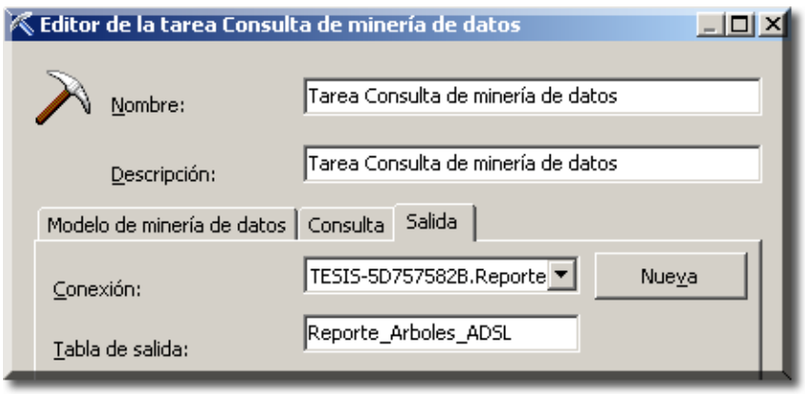


Figura 28: Selección de la salida para la consulta al modelo de minería de datos.

Todas las etapas recientemente explicadas conforman el proceso que se utilizó para obtener resultados que evalúan la calidad predictiva de un modelo de minería datos. Cada modelo evaluado, pertenece a una de las cuatro áreas de negocio que se analizaron, además como es sabido, cada área de negocio fue modela con la técnica de arboles de decisión y redes neuronales, por ende, existen ocho modelos distintos de minería de datos que se evaluaron. Para obtener mejores resultados y poder discernir con claridad que algoritmo manifiesta mejores resultados, sobre una misma área de negocio, se evaluó cada modelo para dos meses distintos escogidos aleatoriamente.

5.5.2 Resultados

El resultado de la evaluación a cada uno de los modelos se expreso por medio de gráficos circulares 3D, en donde se contraponen los clientes que efectivamente se fugaron de la empresa y los clientes reconocidos por el modelo de minería de datos con alta probabilidad de fuga. El porcentaje de clientes que el modelo reconoce como una posible fugas del total de clientes fugados determinó la calidad del modelo.

A continuación se muestran los gráficos para las áreas de negocio analizadas y los resultados que se obtienen con las dos técnicas de modelado. En la figura 29 se muestran los resultados de la evaluación para el área de negocio de internet banda ancha.

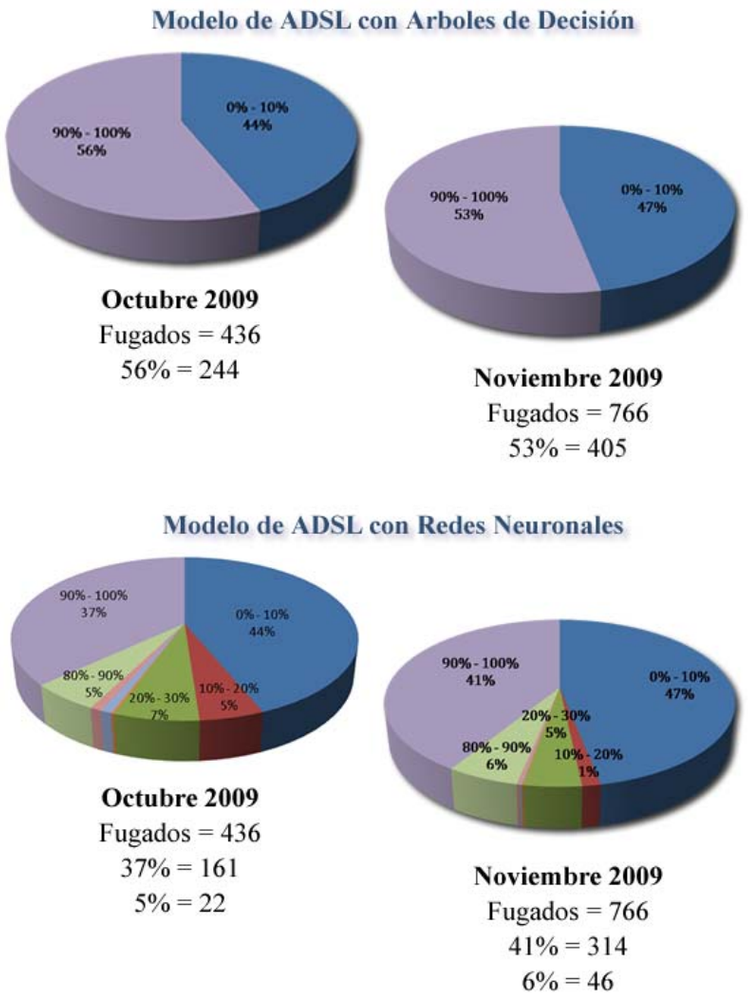


Figura 29: Evaluación del modelo creado para Internet Banda Ancha.

En la figura se aprecia que los clientes que se fugaron de la empresa durante el mes de octubre del 2009 son 436, de esos clientes el modelo de minería de datos entrenado con árboles de decisión detecta al 54% de esos clientes con una probabilidad de fuga sobre el 90%, correspondiendo a 244 clientes, a su vez el modelo entrenado con redes neuronales detecta un 37% de clientes con probabilidad de fuga sobre 90%, correspondiendo a 166 clientes y a otro 5% de ellos con una probabilidad de fuga entre 80% y 90%, lo que equivale a 22 clientes. Para el mes de noviembre del mismo año los clientes fugados de la empresa fueron 766, de aquí el modelo entrenado con árboles de decisión detectó al 53% de esos clientes con una probabilidad de fuga sobre el 90%, correspondiendo a 405 clientes, a su vez el modelo entrenado con redes neuronales detecta un 41% de clientes con probabilidad de fuga sobre 90%, correspondiendo a 314 clientes y a otro 6% de ellos con una probabilidad de fuga entre 80% y 90%, lo que equivale a 46 clientes.

Se puede apreciar, de acuerdo a los resultados de la evaluación, que el modelo de internet banda ancha entrenado con árboles de decisión manifiesta mejores resultados que el modelo entrenado con redes neuronales.

A continuación, en la figura 30, se aprecian los resultados de las evaluaciones realizadas a los modelos de minería de datos para telefonía post pago.

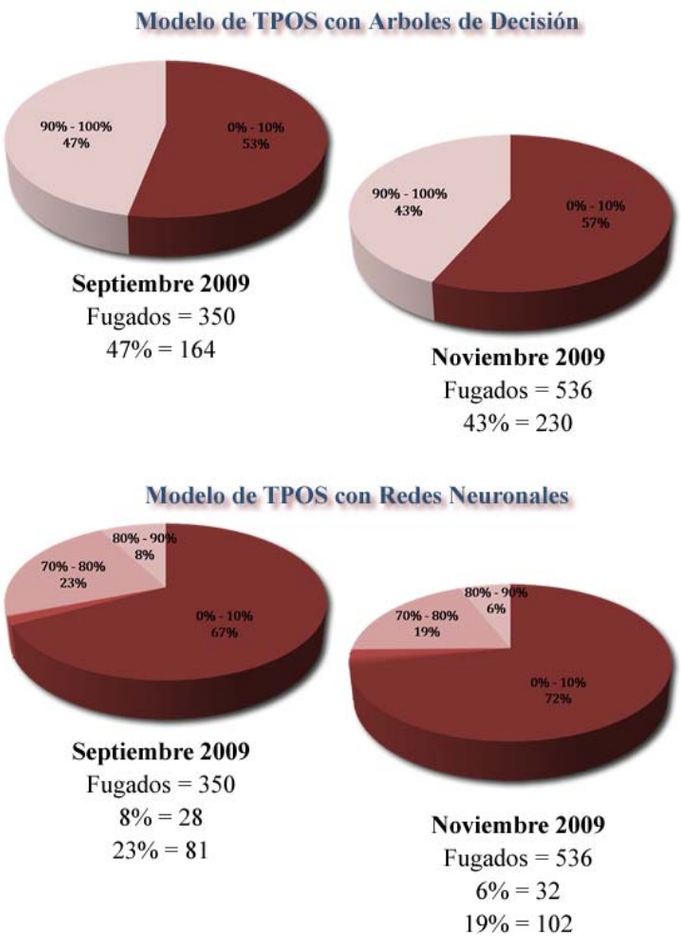


Figura 30: Evaluación del modelo creado para Telefonía Post Pago.

En la evaluación a los modelos de telefonía post pago, nuevamente el modelo entrenado con árboles de decisión manifiesta mejores resultados que el modelo entrenado con redes neuronales. Para el mes de septiembre del 2009, de 350 clientes fugados, el modelo creado con árboles de decisión reconoce a un 47% de ellos con una probabilidad de fuga sobre un 90%, equivalente a 164 clientes, a su vez el modelo creado con redes neuronales detectó una probabilidad de fuga entre el 80% y 90% al 8% de los 350 clientes efectivamente fugados y al 23% de ellos entre un 70% y 80% de probabilidad de fuga, lo que equivale a 81 clientes. Para el mes de noviembre de 530 clientes fugados, el

modelo entrenado con árboles de decisión detectó al 43% de ellos con una probabilidad de fuga sobre el 90%, lo que corresponde a 230 clientes, en cuanto al modelo entrenado con redes neuronales, para ese mismo mes, detectó al 6% de los clientes fugados con una probabilidad de fuga entre el 80% y 90%, equivalente a 32 clientes y al 19% de los clientes fugados con una probabilidad de fuga entre 70% y 80%, lo que equivale a 102 clientes.

En la figura 31, se exhiben los resultados de los modelos creados para el área de negocio de televisión digital inalámbrica.

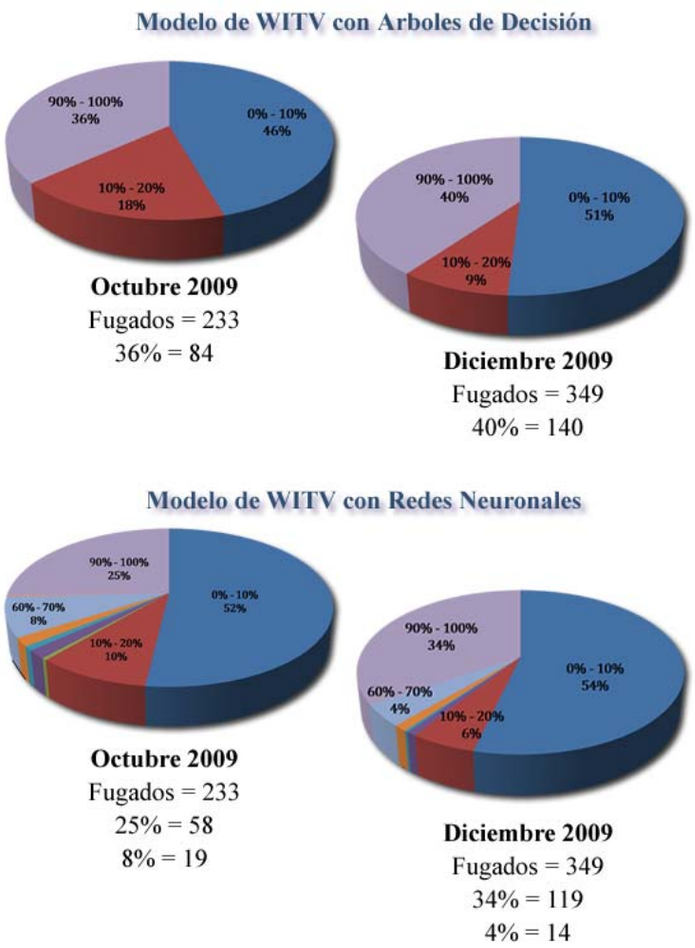


Figura 31: Evaluación del modelo creado para Televisión Digital Inalámbrica.

Al igual que en las evaluaciones anteriores, el modelo entrenado con árboles de decisión para televisión digital inalámbrica, manifiesta mejores resultados que el modelo entrenado con redes neuronales, no es una diferencia excesiva como ocurre en telefonía post pago, pero si existe un mejor rendimiento. Los meses en que se evaluaron estos modelos son octubre y diciembre, en el primer mes, el modelo creado con árboles de

decisión detectó una probabilidad de fuga sobre el 90% para el 36% de los clientes que efectivamente se fugaron de la empresa, lo que equivale a 84 clientes. En tanto el modelo entrenado con redes neuronales detectó a un 25% de los clientes fugados con una probabilidad de fuga mayor a 90%, equivalente a 58 clientes. Situación similar ocurre el mes de diciembre, en donde, el modelo entrenado con árboles de decisión detectó a un 40% de los clientes fugados con una probabilidad de fuga mayor a 90%, equivalente a 140 clientes y el modelo entrenado con redes neuronales solamente detectó a un 34% de los clientes fugados con una probabilidad de fuga mayor a 90%, correspondiendo a 119 clientes.

Finalmente, se observa en la figura 32 los resultados de la evaluación realizada a los modelos creados para el área de negocio de telefonía local inalámbrica.

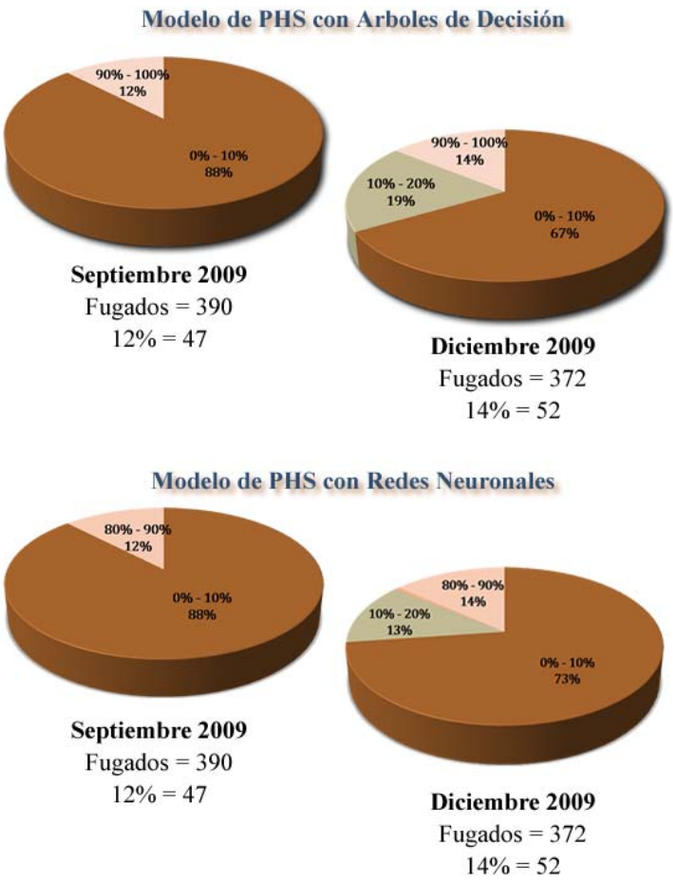


Figura 32: Evaluación del modelo creado para Telefonía Local Inalámbrica.

En la evaluación a los modelos de telefonía local inalámbrica, a diferencia de los casos anteriores, se obtiene un rendimiento muy bajo, las razones pueden ser variadas, desde que no se consideraron las variables que realmente influyan en la fuga de este tipo de servicio, que al ser un área de negocio que solo tiene telefónica del sur, a diferencia de

sus competidores, presenta comportamientos muy distintos a las otras áreas, o que simplemente las fugas de esta área de negocio no posee patrones de comportamientos claros, que puedan ser descritos o descubiertos. En cuanto a los resultados, para el mes de septiembre del 2009, de 390 clientes fugados, el modelo creado con árboles de decisión reconoce a un 12% de ellos con una probabilidad de fuga sobre un 90%, equivalente a 47 clientes, a su vez el modelo creado con redes neuronales detectó una probabilidad de fuga entre el 80% y 90% al 12% de los 390 clientes efectivamente fugados lo que equivale a 47 clientes. Para el mes de diciembre de 372 clientes fugados, el modelo entrenado con árboles de decisión detectó al 14% de ellos con una probabilidad de fuga sobre el 90%, lo que corresponde a 52 clientes, en cuanto al modelo entrenado con redes neuronales, para ese mismo mes, detectó al 14% de los clientes fugados con una probabilidad de fuga entre el 80% y 90%, equivalente a 52 clientes.

CONCLUSIONES

En este estudio se propuso una nueva herramienta a Telefónica del Sur para combatir el churn o tasa de abandono de clientes, mediante la aplicación de modelos predictivos creados con técnicas de minería de datos, que detectan anticipadamente la fuga potencial de clientes de la empresa.

Las áreas de negocios que se sometieron al análisis fueron Internet Banda Ancha, Telefonía Post Pago, Televisión Digital Inalámbrica y Telefonía Local Inalámbrica. En base a esto y a la información entregada por Telefónica del Sur, respecto a las variables relevantes que abarcan todos los ámbitos de información que se necesitan conocer del cliente, se investigó el desempeño de los algoritmos de minería de datos para utilizar aquellos que posean un rendimiento más adecuado para objetivo del proyecto.

De esta forma, se llegó a la conclusión que los algoritmos que debían ser utilizados para lograr el éxito del proyecto eran, los algoritmo de clasificación Bayes Naives, Redes Neuronales y Arboles de Decisión.

El algoritmo de Bayes Naives se utilizó para la exploración inicial de los datos, en donde se descubren relaciones entre las columnas de entrada y la columna de predicción, permitiendo descubrir en cada área de negocio, en base a los clientes fugados durante el año 2009, las variables más influyentes a la hora de construir un modelo predictivo con minería de datos.

En base a los resultados de la exploración inicial, se construyeron modelos de minería de datos, a través de la implementación de los algoritmos Redes Neuronales y Arboles de Decisión, en donde cada uno generó un modelo predictivo que indica, para cada área de negocio, el porcentaje de fuga que posee cada cliente.

La evaluación de cada modelo de minería de datos implementado, se efectuó comparando los clientes que efectivamente se fugaron de la empresa y los clientes reconocidos por el modelo de minería de datos con alta probabilidad de fuga. El porcentaje de clientes que efectivamente se fugaron de la empresa según los datos históricos de la empresa y que el modelo reconoció con probabilidad alta de fuga durante el mes anterior, determinó el indicador que mide la calidad del modelo.

En base a lo anterior, el algoritmo de Arboles de Decisión manifestó un rendimiento superior al exhibido por el algoritmo de Redes Neuronales, descubriendo en los resultados de los modelos predictivos Internet Banda Ancha, Telefonía Post Pago y Televisión Digital Inalámbrica un importante conjunto de clientes con alta probabilidad real de fuga. En cuanto al modelo de Telefonía Local Inalámbrica, los resultados no fueron los esperados, manifestando en la predicción un porcentaje pequeño de clientes con probabilidad real de fuga, muy inferior a los modelos anteriores. En esta área de negocio el modelo creado con Arboles de Decisión y el modelo creado con Redes Neuronales presentan resultados muy similares.

En cuanto a la evaluación individual de cada área de negocio, se obtiene que para el modelo de Internet Banda Ancha, creado con el algoritmo de arboles de decisión, se predice en promedio al 55% de los clientes efectivamente fugados, con una probabilidad de fuga por encima del 90%, a diferencia del algoritmo de redes neuronales que solo predice en promedio al 39%. Para Telefonía Post Pago el modelo creado con el algoritmo de arboles de decisión predice en promedio, el 45% de los clientes efectivamente fugados con una probabilidad de fuga sobre del 90%, a diferencia del algoritmo de redes neuronales que solo descube en promedio 28% de los clientes que efectivamente se fugaron de la empresa, con una probabilidad de fuga entre el 70% y 90%. En lo que respecta al área de negocio de Televisión Digital Inalámbrica el modelo creado con el algoritmo de arboles de decisión detecta en promedio, el 38% de los clientes efectivamente fugados, con una probabilidad de fuga por encima del 90%, a diferencia del algoritmo de redes neuronales que solo predice en promedio al 30%.

Considerando los resultados obtenidos y el grado de efectividad de cada modelo, se recomienda utilizar los modelos generados con el algoritmo de Arboles de Decisión, debido a que presenta un mejor desempeño que Redes Neuronales para descubrir aquellos clientes que son posibles candidatos de abandonar la empresa.

La información entregada por cada modelo será fundamental para aplicar anticipadamente acciones de retención, permitiendo a la empresa reducir los costos de retención, disminuir la tasa de abandono, focalizar los esfuerzos de marketing y mejorar la comunicación con sus clientes.

Finalmente se puede concluir que cada modelo de minería de datos tiene la capacidad de potenciar aún más su rendimiento. Incorporando a la tabla de escenarios más información relevante de los clientes, como por ejemplo, características personales, información del grupo familiar, comportamientos que posee como cliente y el grado de satisfacción con los servicios contratados. Debido a que los datos utilizados para entrenar el modelo, son la clave de éxito de un modelo predictivo creado con Minería de Datos.

REFERENCIAS

[JAR05] Jaramillo Ovalle Pamela, 2005. “Diseño de un Modelo Parametrizado, usando Data Mining, que permita a Telefónica del Sur Segmentar por Valor a los Clientes de Prepago”. Tesis de Grado. Universidad Austral de Chile.

[DOM09] María de la Luz Domper. (2009). “Problemas de competencia en las telecomunicaciones”. Foro Latinoamericano de Competencia. Disponible http://www.lyd.cl/lyd/controls/neochannels/neo_ch3754/deploy/luli%20tdlc.pdf. Consultado el 26 de Diciembre de 2010.

[SUB] Subsecretaria de Telecomunicaciones. (n. d.). Informe, Análisis del Mercado de Productos de Telecomunicaciones en Chile. Disponible en www.subtel.cl/prontus_subtel/site/artic/20090422/asocfile/20090422104633/informe_iii_analisis_mercado_prod_teleco.pdf. Consultado el 27 de Diciembre de 2010.

[DAT00] Dataprix., (2000). Prologo CRISP-DM. Disponible en http://www.dataprix.com/prologo-crisp-dm. Consultado el 29 de Diciembre de 2010.

[CHA99] Pete Chapman., (1999). CRISP-DM 1.0, Step-by-step data mining guide. Disponible en www.crisp-dm.org/CRISPWP-0800.pdf. Consultado el 29 de Diciembre de 2010.

[MDN08] Microsoft Developer Network., (2008). Algoritmos de Minería de Datos. SQL Server 2005. Disponible en http://msdn.microsoft.com/es-es/library/ms175595(SQL.90).aspx. Consultado el 4 de enero de 2010.

[Tan05] Tang Z & MacLennan J (2005). Data Mining with SQL Server 2005. Wiley Publishing, Inc.

[MDN08] Microsoft Developer Network., (2008). Usar las herramientas de minería de datos. SQL Server 2005. Disponible en http://msdn.microsoft.com/es-es/library/ms175312(v=SQL.90).aspx. Consultado el 5 de enero de 2011.

[San05] Sánchez S, Ayuso M & Caridad J. (2005). Software de Minería de Datos: Análisis de Características. Disponible en http://www.iadis.net/dl/final_uploads/200508C006.pdf. Consultado el 11 de Enero de 2011.

[Han05] Hancock J & Toren R (2006). Practical Business Intelligence with SQL Server 2005. Addison Wesley Professional.

[Fer06] Enrique J. Fernández, 2006. “Asistente para la Gestión de Documentos de Proyectos de Explotación de Datos”. Tesis de Magíster en Ingeniería de Software. Instituto Tecnológico de Buenos Aires. Disponible en http://www.itba.edu.ar/archivos/secciones/fernandez-tesisdemagister.pdf. Consultado el 13 de enero de 2011.

[CRI] CRISP-DM. (n. d.). CRoss Industry Standard Process for Data Mining, Metodología CRISP-DM. Disponible en http://www.crisp-dm.org/Process/index.htm. Consultado el 14 de enero de 2011.

[Mal03] Malagón Constantino, (2003). Clasificadores Bayesianos. El algoritmo de Naive Bayes. Disponible en http://www.nebrija.es/~cmalagon/inco/Apuntes/bayesian_learning.pdf. Consultado el 15 de febrero de 2011.

ANEXOS

Anexo1: Procedimientos almacenados

Tabla 8: Descripción de los procedimientos que realizan las tareas de preparación de datos.

NOMBRE DEL PROCEDIMIENTO	DESCRIPCIÓN
ProcMD_ADSL_datos_clientes ProcMD_PHS_datos_clientes ProcMD _TPOS_datos_clientes ProcMD _WITV_datos_clientes	La función de estos procedimientos es insertar a tablas temporales de minería de datos, todos los clientes vigentes de la empresa por medio de datos identificatorios de ellos, además elabora e inserta la variable antigüedad del servicio, zona y localidad de residencia. Es generado un procedimiento de este tipo para cada uno de los servicios analizado.
ProcMD_VisitaOficinasComerciales	Este procedimiento busca a los clientes de las tablas temporales que hayan visitado las oficinas comerciales e inserta la cantidad de veces que ocurrió una visita en el mes anterior al analizado. Se crea un procedimiento de este tipo para cada uno de los servicios de análisis.
ProcMD_Modificaciones_Servicios	Este procedimiento busca a los clientes de las tablas temporales de minería de datos que han realizado modificación de algún tipo a sus servicios e inserta la cantidad de modificaciones que realizó en el mes anterior al analizado. Se crea un procedimiento de este tipo para cada uno de las áreas de negocio analizadas.
ProcMD_Modificaciones_Servicios_3 meses	Este procedimiento busca a los clientes de las tablas temporales de minería de datos que han realizado modificación a sus servicios en los últimos tres meses anteriores al analizado, e inserta la cantidad de modificaciones que realizó. Se crea un procedimiento de este tipo para cada uno de los servicios de análisis.
ProcMD_reclamos_comerciales	Este procedimiento busca a los clientes de las tablas temporales de minería de datos que han realizado reclamos comerciales e inserta la cantidad de reclamos que realizó en el mes anterior al analizado. Se crea un procedimiento de este tipo para cada uno de los servicios de análisis.
ProcMD_reclamos_comerciales_3mes es	Este procedimiento busca a los clientes de las tablas temporales de minería de datos que han realizado reclamos comerciales en los últimos tres meses anteriores al

	analizado, e inserta la cantidad de reclamos que realizó. Se crea un procedimiento de este tipo para cada uno de los servicios de análisis.
ProcMD_reclamos_tecnicos	Este procedimiento busca a los clientes de las tablas temporales de minería de datos que han realizado reclamos técnicos e inserta la cantidad de reclamos que realizó en el mes anterior al analizado. Se crea un procedimiento de este tipo para cada uno de los servicios de análisis.
ProcMD_reclamos_tecnicos_3meses	Este procedimiento busca a los clientes de las tablas temporales de minería de datos que han realizado reclamos técnicos en los últimos tres meses anteriores al analizado, e inserta la cantidad de reclamos que realizó. Se crea un procedimiento de este tipo para cada uno de los servicios de análisis.
ProcMD_impugnaciones	Este procedimiento busca a los clientes de las tablas temporales de minería de datos que han tenido alguna impugnación en sus servicios e inserta la cantidad de impugnaciones que tuvo en el mes anterior al analizado. Se crea un procedimiento de este tipo para cada uno de los servicios de análisis.
ProcMD_impugnaciones_3meses	Este procedimiento busca a los clientes de las tablas temporales de minería de datos que han tenido alguna impugnación en sus servicios en los últimos tres meses anteriores al analizado, e inserta la cantidad de impugnaciones que tuvo. Se crea un procedimiento de este tipo para cada uno de los servicios de análisis.
ProcMD_llamadas_competencia	Este procedimiento busca a los clientes de las tablas temporales de minería de datos que han realizado llamadas a los números de las empresas de la competencia e inserta la cantidad de llamadas que realizó en el mes anterior al analizado. Se crea un procedimiento de este tipo para cada uno de los servicios de análisis.
ProcMD_llamadas_competencia_3meses	Este procedimiento busca a los clientes de las tablas temporales de minería de datos que han realizado llamadas a los números de las empresas de la competencia en los últimos tres meses anteriores al analizado, e inserta la cantidad de llamadas que realizó. Se crea un procedimiento de este tipo para cada uno de los servicios de análisis.
ProcMD_contratacionesPPV	Este procedimiento busca a los clientes de la tabla temporal de WITV que han contratado servicios de pay per view e inserta la cantidad de contrataciones que realizó en el mes

	anterior al analizado.
ProcMD_contratacionesPPV_3meses	Este procedimiento busca a los clientes de la tabla temporal de WITV que han contratado servicios de pay per view en los últimos tres meses anteriores al analizado, e inserta la cantidad de contrataciones que realizó.
ProcMD_trafico_llamadas	Este procedimiento selecciona a los clientes de las tablas temporales de TPOS Y PHS y busca la cantidad de llamadas telefónicas que han realizado en los últimos tres meses anteriores al analizado, e inserta la cantidad de llamadas que realizó, independiente de la duración o el destino.
ProcMD_EliminarRepetidos_Insertar_CantServicios	Este procedimiento realiza dos tareas, primero elimina los datos repetidos que puedan existir en las tablas temporales y posteriormente elabora e inserta la variable discreta que informa sobre la cantidad de productos que posee un cliente de un mismo servicio. Se crea un procedimiento de este tipo para cada uno de los servicios de análisis.
ProcMD_Clientes_Fugados	La función de este procedimiento es insertar a la tabla temporal de clientes fugados, todos los clientes que han dado de baja algunos de sus servicios durante el mes analizado. Junto con agregar datos identificatorios de ellos, inserta las variables antigüedad del servicio, zona y localidad de residencia.
ProcMD_Fugados_VisitaOficinasComerciales	Este procedimiento busca a los clientes de la tabla temporal de clientes fugados, que visitaron las oficinas comerciales e inserta la cantidad de veces que realizó la visita durante en el mes anterior al de la fuga. Se crea un procedimiento de este tipo para cada uno de los servicios analizados.
ProcMD_Fugados_Modificaciones_Servicios	Este procedimiento busca a los clientes de la tabla temporal de clientes fugados, que realizaron modificaciones en alguno de sus servicios e inserta la cantidad de modificaciones que realizó durante el mes anterior al de la fuga. Se crea un procedimiento de este tipo para cada uno de los servicios analizados.
ProcMD_Fugados_Modificaciones_Servicios_3meses	Este procedimiento busca a los clientes de la tabla temporal de clientes fugados, que realizaron modificaciones en alguno de sus servicios e inserta la cantidad de modificaciones que realizó durante los últimos tres meses anteriores al de la fuga. Se crea un procedimiento de este tipo para cada uno de los servicios analizados.
ProcMD_Fugados_reclamos_comerci	Este procedimiento busca a los clientes de la tabla temporal

ales	de clientes fugados, que realizaron reclamos comerciales e inserta la cantidad de reclamos que realizó durante el mes anterior al de la fuga. Se crea un procedimiento de este tipo para cada uno de los servicios analizados.
ProcMD_Fugados_reclamos_comerciales_3meses	Este procedimiento busca a los clientes de la tabla temporal de clientes fugados, que realizaron reclamos comerciales e inserta la cantidad de reclamos que realizó durante los últimos tres meses anteriores al de la fuga. Se crea un procedimiento de este tipo para cada uno de los servicios analizados.
ProcMD_Fugados_reclamos_tecnicos	Este procedimiento busca a los clientes de la tabla temporal de clientes fugados, que realizaron reclamos técnicos e inserta la cantidad de reclamos que realizó durante el mes anterior al de la fuga. Se crea un procedimiento de este tipo para cada uno de los servicios analizados.
ProcMD_Fugados_reclamos_tecnicos_3meses	Este procedimiento busca a los clientes de la tabla temporal de clientes fugados, que realizaron reclamos técnicos e inserta la cantidad de reclamos que realizó durante los últimos tres meses anteriores al de la fuga. Se crea un procedimiento de este tipo para cada uno de los servicios analizados.
ProcMD_Fugados_impugnaciones	Este procedimiento busca a los clientes de la tabla temporal de clientes fugados, que tuvieron alguna impugnación en sus servicios e inserta la cantidad de impugnaciones existentes durante el mes anterior al de la fuga. Se crea un procedimiento de este tipo para cada uno de los servicios analizados.
ProcMD_Fugados_impugnaciones_3meses	Este procedimiento busca a los clientes de la tabla temporal de clientes fugados, que tuvieron alguna impugnación en sus servicios e inserta la cantidad de impugnaciones existentes durante los últimos tres meses anteriores al de la fuga. Se crea un procedimiento de este tipo para cada uno de los servicios analizados.
ProcMD_Fugados_llamadas_competencia	Este procedimiento busca a los clientes de la tabla temporal de clientes fugados, que realizaron llamadas telefónicas a los números de la competencia e inserta la cantidad de llamadas que realizó durante el mes anterior al de la fuga. Se crea un procedimiento de este tipo para cada uno de los servicios analizados.
ProcMD_Fugados_llamadas_competencia_3meses	Este procedimiento busca a los clientes de la tabla temporal de clientes fugados, que realizaron llamadas telefónicas a los

	números de la competencia e inserta la cantidad de llamadas que realizó durante los últimos tres meses anteriores al de la fuga. Se crea un procedimiento de este tipo para cada uno de los servicios analizados.
ProcMD_Fugados_contratacionesPPV	Este procedimiento busca a los clientes de la tabla temporal de clientes fugados, que han contratado servicios de pay per view e inserta la cantidad de contrataciones que realizó durante el mes anterior al de la fuga. Este procedimiento se crea para el análisis del servicio de WITV solamente.
ProcMD_Fugados_contratacionesPPV_3meses	Este procedimiento busca a los clientes de la tabla temporal de clientes fugados, que han contratado servicios de pay per view e inserta la cantidad de contrataciones que realizó durante los últimos tres meses anteriores al de la fuga. Este procedimiento se crea para el análisis del servicio de WITV solamente.
ProcMD_Fugados_Trafico_llamadas_3meses	Este procedimiento busca a los clientes de la tabla temporal de clientes fugados que poseían servicios de TPOS y PHS, y cuenta la cantidad de llamadas telefónicas que realizaron durante los últimos tres meses anteriores al de la fuga e inserta esa cantidad de llamadas que realizó, independiente de la duración o el destino.
ProcMD_insertar_datos_fugados	Este procedimiento realiza dos funciones, primero elimina los datos repetidos que puedan existir de la tabla temporal de los clientes fugados y posteriormente inserta los datos de estos clientes en las tablas temporales de minería de datos. Aquí realiza la separación de los clientes fugados según el servicio por el cual se realizo la baja del servicio.
ProcMD_Fugados_cantidadServicios	Este procedimiento elabora e inserta la variable con la información de la cantidad de productos que poseían los clientes fugados de un mismo servicio. Esta tarea se realiza en cada una de las tablas temporales de minería de datos que representan a los cuatro servicios analizados.
ProcMD_Edad_cliente	Este procedimiento elabora la variable discreta de la edad de los clientes, tanto vigentes como fugados. Esta variable se obtiene mediante la estimación de la edad por medio del número de Rut del cliente, debido a que la empresa no tiene información sobre la fecha de nacimiento de sus clientes. Se crea un procedimiento de este tipo para cada uno de los servicios analizados.
ProcMD_Eliminar_Empresas	Este procedimiento elimina de las tablas temporales de

	minería de datos a todos los clientes que corresponden a una empresa o instituciones.
ProcMD_Insertar_Datos_DM	Este procedimiento inserta toda la información de las tablas temporales en las tablas definitivas de minería de datos, agregándole un identificador único a cada registro de las tablas. Esto es fundamental para el desarrollo del modelo de minería de datos. Se realiza esta tarea para cada una de las tablas que representan a los servicios analizados.