Aplicación de Minería de Datos para Predecir Fuga de Clientes en la Industria de las Telecomunicaciones

Francisco Barrientos *
Sebastián A. Ríos *

Resumen

Hoy día, la minería de datos está cobrando relevancia creciente en empresas u organizaciones para resolver problemas complejos de negocio. Por ejemplo, el procesamiento de volúmenes masivos de datos donde se esconde información valiosa respecto del comportamiento de compra de productos o servicios, hasta generar nuevos productos usando dichos comportamientos de los clientes. Esto es particularmente cierto en el mercado de negocios de las telecomunicaciones donde el número de clientes normalmente llega a varios millones y un analista humano es incapaz de realizar su labor sin metodologías y algoritmos que permitan automatizar o semi-automatizar el descubrimiento de conocimiento. El problema es aún más complejo, si tomamos en consideración que estas empresas cuentan con muchos sistemas internos desde los cuales debe ser obtenida la información necesaria y suficiente para poder realizar cualquier modelo predictivo. Este paper tiene por objetivo mostrar una metodología para poder realizar predicción de fuga de clientes ó Churn en un ambiente multiplataforma en la industria de las telecomunicaciones. Además, se usaron diversos algoritmos como Redes Neuronales, Support Vector Machines y Árboles de Decisión y se evaluó la calidad como el porcentaje de aciertos en la variable predicha. Esta fue aplicada a una empresa de telecomunicaciones real; donde se utilizaron los resultados para poder generar estrategias de retención de clientes y así realizar la evaluación de la calidad de los resultados obtenidos.

Palabras Clave: Proceso KDD, Minería de Datos, Modelos Predictivos, Predicción de Churn

*Centro de Investigación en Inteligencia de Negocios (www.ceine.cl). Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile, Santiago, Chile.

1. Introducción

El estudio del churn o fuga de clientes es un área en la cual año a año se invierten grandes recursos: equipos de especialistas, consultoras externas, software especializado, etc. Siempre, con la intención de poder descubrir de manera anticipada, si es que un cliente va a decidir cambiarse de una compañía a su competencia. Por ejemplo, pasar del Retail A al B, o de la Farmacia X a la Y. En particular, en el área de las telecomunicaciones, se ha hecho cada vez más necesario estudiar la fuga de clientes, dada la alta competitividad que se está desarrollando a nivel mundial y las nuevas legislaciones que están surgiendo en Chile (como, la portabilidad numérica). En la industria de las telecomunicaciones durante los años 2008 a 2010, la fuga de clientes llegó a ser del 30 % anual [28, 37] (estudios previos a la portabilidad numérica). A partir de lo anterior, se desprende que la fuga de clientes es un problema de la industria y donde se hace necesaria la aplicación de herramientas avanzadas que permitan predecir y describir de algún modo, qué clientes tienen mayor potencial de riesgo de cambiarse de compañía.

Las líneas de negocio presentes que encontramos comúnmente en empresas de telecomunicaciones son: tráficos de larga distancia, telefonía local, internet, cargos de accesos, servicios privados, facturas y cuentas corrientes atención de clientes (referente a las solicitudes de atención y reclamos), clientes y contratos, modelos de operaciones, participación de mercado y suscriptores. A nivel mundial, el tamaño de esta industria es de 4,03 trillones de dólares, cifra que se pronostica que crezca a un 6 % hasta el año 2013 [4]. Respecto al tamaño local en cuanto a inversiones realizadas en este mercado, la cifra rodea 778.153 millones de dólares [13].

Dentro de las principales razones para que un cliente deje de comprar productos de una compañía se destacan la disconformidad y la falta de políticas de retención efectivas expresadas en un mejor trato hacia ellos [27]. Para posicionar lo anterior en un contexto local se detectan cerca 11.809 reclamos en el organismo regulador de los cuales 5.242 corresponderían a disconformidades, ya sea con la suscripción, continuidad y calidad del servicio o bien con cobros irregulares [14]. Otra de las características del sector es la gran cantidad de información que generan y almacenan sus empresas [1], por ende, nuevas tendencias e ideas han aparecido para hacer uso de la minería de datos en una gran cantidad de áreas, en especial, en marketing, detección de fraude y control de calidad [1], donde la fidelización juega el rol más importante. Las empresas del sector están

especialmente sensibilizadas con la pérdida de clientes que escogen una compañía de la competencia, varios autores señalan que es más costoso conseguir un nuevo cliente que mantener uno antiguo [1, 28, 31, 43]. Por ello la finalidad de este paper es introducir el procedimiento KDD dentro de la empresa de telecomunicaciones para que pueda ser aplicado en todas sus áreas, en particular, con el objeto de detectar la fuga de clientes. Es así como se define el problema a tratar el cual consiste en predecir la fuga de clientes en esta empresa para un producto particular (NGN) cuyo segmento objetivo son las pequeñas y medianas empresas (PYMES). Una de las principales dificultades es que la información necesaria para predecir la fuga de clientes se encuentra distribuida en un sistema multiplataforma, el cual consiste en múltiples plataformas con información de clientes, transaccional, reclamos, entre otras. Estas plataformas no disponen de transferencias automáticas de información lo que dificulta aun más este trabajo. Dentro de este problema el término churn hace su aparición, el cual se es usado en el sector de telecomunicaciones para describir el cese de servicios de la suscripción de un cliente. Se habla de churner o fugado para denominar a aquel cliente que ha dejado la compañía [5].

2. Revisión de literatura

2.1. Conceptos básicos y tipos de fuga de clientes

La fuga de clientes, dentro de las telecomunicaciones, se produce cuando un cliente cancela el servicio prestado por la compañía [31]. En dicha cancelación, el cliente puede decidir renunciar a la empresa (voluntaria), o bien, la empresa puede expulsarlo (involuntaria). En particular, la connotación de churn hace referencia la fuga de los clientes, por lo que, para efectos de este estudio, se cuenta el churn en base a la decisión del cliente en abandonar la empresa por medio de la cancelación de un servicio. También, se puede entender como churn a aquel término *"usado para describir colectivamente el cese de servicios de la suscripción de un cliente...donde el cliente es alguien que se ha unido a la compañía por al menos un período de tiempo...un churner o fugado es un cliente que ha dejado la compañía"* [5]. Los principales tipos de fuga acorde a [5, 40] son:

- Absoluta: suscriptores que se han desligado sobre la base de datos total en un período.
- De línea o servicio: Este tipo de churn el número de servicios disconti-

nuados sobre la base de datos total

- Primaria: Referente al número de fallas
- Secundaria: Descenso en el volumen de tráfico
- Fuga de paquete: Esta fuga se caracteriza por el hecho de que cambian de planes y/o productos dentro de la compañía [5].
- Fuga de la compañía: Sin lugar a dudas el más costoso, en este caso el cliente se fuga hacia la competencia, por ende, no solamente se pierde el ingreso no percibido, sino que también el prestigio de la compañía expresado en la participación de mercado de la competencia.

2.2. El proceso KDD

La minería de datos se establece como una de las etapas de un proceso más genérico denominado Knowledge Discovery in Databases (KDD), el cual, es el proceso de análisis de bases de datos que busca encontrar relaciones inesperadas que son de interés o valor para el poseedor de dicha base de datos [20]. En términos simples es encontrar relaciones no triviales en los datos. Este proceso iterativo consiste en cinco etapas, en donde la minería de datos es definida como una fase más de este procedimiento. Al ser un proceso iterativo es posible volver a una etapa previa en caso de que no se tengan resultados satisfactorios al final de una. A continuación se describe dichas etapas o fases bajo la perspectiva de [11, 16]:

- Integración o Selección: Aquí se escogen las variables y las fuentes a considerar en el proceso completo, por lo que se refiere a la creación del conjunto de datos como la base de datos de estudio en el proceso. Dentro de este paper, esta etapa adquiere gran importancia, pues está sujeta a la interacción de múltiples plataformas como parte de selección de fuentes de información. La sección 2.2.1 presenta una descripción detallada de esta etapa.
- Preprocesamiento: El análisis y limpieza de los datos son las líneas principales a seguir en esta sección, donde se produce el tratamiento de valores ausentes (missing), los valores fuera de rango (outliers). Para ello, se emplean distintas técnicas de imputación de datos que van desde un tratamiento valor a valor (simple imputation) hasta un reemplazo contemplando múltiples variables y sus valores (multiple imputation). La sección 2.2.2 presenta una descripción detallada de esta etapa.

- Transformación: Acá se generan nuevas variables a partir del estudio de la naturaleza de las variables originales; desde la perspectiva de la escala, nominal o continua, o bien de la distribución de los valores presentes. La sección 2.2.3 presenta una descripción detallada de esta etapa.

- Minería de datos: Este paso en el proceso de KDD, consiste en la aplicación de análisis de datos para descubrir un algoritmo ad-hoc que produzca una particular enumeración de patrones a partir de los datos y que los produzca considerando restricciones de capacidad computacional [16]. Por ende, se selecciona el modelo y algoritmo a utilizar, bajo los supuestos que mantienen los objetivos primarios del estudio. La sección 2.2.5 presenta una descripción detallada de esta etapa.

- Interpretación y Evaluación: Esta última fase involucra las medidas de evaluación y la trasposición de resultados técnicos a niveles comerciales, de tal manera, que la aplicación del procedimiento converja a acciones correctivas en el negocio, que solucionen el fenómeno estudiado. Respecto a la evaluación, ésta se puede aplicar desde dos aristas: técnica y comercial. La primera se subdivide acorde al tipo de validación y sus métricas que se aplican al modelo, mientras que la evaluación comercial no se encuentra estandarizada y generalmente se puede utilizan encuestas o grupos de blindaje para medir la efectividad práctica del procedimiento. Las principales técnicas de evaluación técnica son el "holdout" y la validación cruzada. La sección 2.2.6 presenta una descripción detallada de esta etapa.

2.2.1. Data Warehousing

En la primera etapa del KDD (Integración), se requieren fuentes de información consolidadas, por ello, es que generalmente se aplica este procedimiento posterior a la implementación de un data warehouse (DWH) en la compañía. Este concepto se define como la colección de tecnologías de soporte decisivo que permite al trabajador tomar buenas y rápidas decisiones [9]. Esta colección debe ser orientada al sujeto, integrada, variante en el tiempo y estable, por ende, generalmente, se mantiene apartada de las bases de datos operacionales, pues se busca la consolidación de los datos [9, 41]. Por lo tanto, es lógico pensar que un data warehouse contiene datos consolidados a partir de múltiples bases de datos operacionales, durante extensos períodos de tiempo, por lo que es común que su tamaño alcance varios gigabytes o terabites [9]. Es importante destacar que cada estructura en un data warehouse posee una dimensión temporal [41].

Por lo anterior, la implementación de un data warehouse en la empresa es un proceso largo y complejo [9]. En ocasiones, las compañías optan por usar data marts en vez de construir un data warehouse, estos son subconjuntos de datos orientados a un departamento o área determinada, por lo mismo, no requieren de un consenso general a nivel de empresa, sin embargo, si no se incorpora estratégicamente la utilización del data warehouse se pueden producir problemas complejos de integración en el largo plazo [9]. Las principales diferencias se muestran a continuación en la Tabla 1:

DATA WAREHOUSE	DATA MART
Implementación a nivel de corporación o empresa	Implementación a nivel departamental
Aplicación a la unión de todos los data marts	Aplicación a un proceso de negocios singular
Consultas en recurso de presentación	Tecnología óptima para el acceso de datos y análisis
Estructura orientada a una vista empresarial de los datos	Estructura orientada a una vista departamental de los datos
Organización en base a un modelo entidad relación	

Tabla 1: Diferencias entre un Data warehouse y un Data mart

El diagrama general de la confección de un data warehouse puede observarse tal y como procede [9] en la Figura 1:

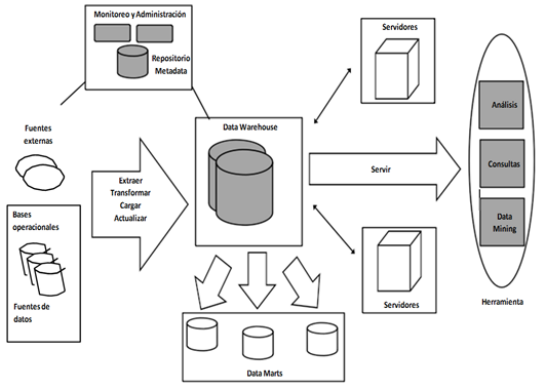

Figura 1: Diagrama estructural de un Data Warehouse

Ahora bien, la importancia de esta tecnología en las telecomunicaciones se debe a que generan un cambio en la forma en que la gente ve el desarrollo en sistemas de los [38], es decir, permite acercar el modelo de negocios a los empleados. Eventualmente a partir de esta colección de datos se requiere extraer información útil a los usuarios finales para lo cual se puede aplicar data mining [38].

2.2.2. Imputación de datos

Posterior a la integración de datos facilitada por un data warehouse, se analizan los conjuntos respectivos, cuyos primeros fenómenos observables son la aparición de valores anómalos, ausentes, o con variabilidad en el tiempo. Los valores ausentes son aquellos que no se encuentran en la base de datos analizada, su subdivisión se efectúa en base a las posibles razones de dicho fenómeno y es la siguiente [42]:

- Missing Completely at random (MCAR): Son aquellos datos perdidos que son completamente al azar, es decir, no poseen ningún tipo de relación con los datos presentes en otras variables. En este escenario, se asume que las distribuciones de probabilidad de los datos faltantes y de todos los datos son idénticas.

- Missing at Random (MAR): Estos datos o valores a diferencia de los anteriores, presentan relaciones con el resto de las variables, por lo que la distribución de probabilidad de los valores perdidos puede tener una distribución distinta a la variable en general. Los valores de este tipo pueden ser predichos usando datos completos[12, 30].

- Not Missing at random (NMAR): Son valores perdidos que tienen significado, es decir, el hecho de que esté ausente no es un error, sino información relevante para la variable

Otro tipo de fenómeno son los valores fuera de rango u outliers que se definen como aquellos valores en el conjunto de datos cuyo comportamiento es anómalo con respecto a lo observado en la mayoría de los registros [2].

Finalmente, es usual la existencia de las variables temporales, las cuales son variables cuyos valores pueden ser diferentes en distinto momentos del tiempo [8].

El trato que se le da a los datos afectos a dichos fenómenos es distinto dependiendo del tipo descrito previamente. Para el tratamiento de los valores ausentes existen distintas alternativas, dentro de las cuales, las más relevantes son [12]:

- Descarte de los registros con datos faltantes: Esta alternativa suele utilizarse cuando la información que aporta la variable es baja, o bien, la cantidad de valores perdidos es baja y la variable tiene poca varianza.

- Reemplazo de los datos faltantes con otro valor: Esta alternativa suele usarse para identificar al valor ausente.

- Imputación de los datos faltantes: Esta alternativa es factible cuando la cantidad de atributos con datos faltantes es relativamente pequeña en relación al número de registros que presentan dicha condición. Este método, sí influye en la información de la variable.

Otra perspectiva o alternativa de solución, frente a los datos perdidos, la presentan Roderick Little et al. [30], que establece las siguientes técnicas para tratar los valores ausentes:

1. Procedimientos basados en instancias completas: Esto se refiere a que cuando algunas variables no están guardadas para determinadas instancias, se sugiere simplemente eliminar éstas y analizar solamente las instancias con atributos completos. Los análisis usuales de este tipo son:

   - Listwise deletion: Este análisis sugiere que al momento de realizar la predicción, se debe trabajar con las observaciones que disponen de la información completa para todas las variables [17].
   - Parwise deletion: Este análisis considera la información completa, pero usando distintos tamaños de muestra. A diferencia del análisis listwise, solamente se eliminan aquellas observaciones que no poseen ningún dato, y los cálculos se realizan con diferentes tamaños de muestra lo que limita comparación de resultados [17].

2. Procedimientos basados en la imputación: En estos procedimientos *"los valores perdidos son llenados y la base de datos completada es analizada por métodos estandarizados. Los métodos comúnmente usados incluyen Hot Deck, Imputación por promedio e imputación por regresión"* [30], además, se agregan otro métodos como el cold deck. El hot deck imputa para cada ejemplo que contenga un valor perdido, se encuentra el ejemplo más similar y los valores perdidos son imputados de dicho ejemplo [12]. El método de la media consiste en una imputación de los valores anómalos por la media de la variable. La regresión sugiere *"imputar la información de una variable Y a partir de un grupo de covariables $X_1, X_2, \ldots, X_n$"* [29]. El cold deck consiste en seleccionar los valores o relaciones de uso obtenidos de fuentes distintas al conjunto de datos actual [44].

3. Procedimiento de asignación de pesos: Este método es usado cuando se desea adquirir una probabilidad de selección, pues "las inferencias aleatorias de la muestra de una encuesta sin respuesta comúnmente están basadas en pesos de diseño que suelen ser inversamente proporcionales a la probabilidad de selección" [30]. Usualmente se usa para estimar la población promedio a partir de una muestra. Reponderación: Este tipo de

imputación adquiere importancia cuando se tienen muy pocas respuestas o valores de alguna categoría de interés. Las ponderaciones se interpretan como el número de unidades de la población [17].

4. Procedimientos basados en modelos: Este procedimiento se aplica cuando se intuye una relación, lineal o no lineal, entre un subconjunto de variables. No obstante, el procedimiento completo se sintetiza en *"definir un modelo para los valores parcialmente perdidos y basando las inferencias en la similitud del modelo, con parámetros estimados bajos procedimientos tales como el de máxima verosimilitud"* [30].

2.2.3. Transformaciones especiales: ACP, Segmentación, RFM

- Transformaciones para variables temporales: Estas variables son referenciadas como secuencias temporales, las cuales *"se forman con los datos recopilados en una base sobre la evolución en el tiempo de un conjunto de características"* [39]. Las transformaciones que se utilizan en estas variables son: Índices, por funciones que preserven la ortonormalidad o promedio ponderados.

- Análisis factorial: Este análisis tiene el propósito de *"simplificar las numerosas y complejas relaciones que se puedan encontrar en un conjunto de variables cuantitativas observadas"* [29]. Esto quiere decir que no se encarga de reducir las variables, sino que busca encontrar el significado de los nuevos factores generados producto del análisis de componentes principales. Por ende, su definición converge a *"un procedimiento matemático mediante el cual se pretende reducir la dimensión de un conjunto de p variables obteniendo un nuevo conjunto de variables más reducido, pero capaz de explicar la variabilidad común encontrada en un grupo de individuos sobre los cuales se han observado las p variables originales"* [29].

- Modelo Recency, Frequency, Mount (RFM): este modelo data antes del año 2000 en el marketing cualitativo, como forma de medir el comportamiento del consumidor. Esta medición se hace desde tres perspectivas [23]. La primera es Recency, que indica hace cuánto el cliente respondió [23], que en otras palabras, significa el tiempo transcurrido desde la última vez que el cliente registró un accionar con la compañía. La segunda perspectiva es la Frecuencia, que provee una métrica de cuán seguido el cliente ha respondido a recibir mails, que pretende indicar, el tiempo transcurrido entre interacciones del cliente con la compañía. Finalmente, la tercera perspectiva es el valor monetario, que mide el monto en dinero o

el número de productos que el cliente ha gastado o consumido en respuesta a los mails enviados [23], lo que expresado de forma distinta, indica el valor monetario o cantidad que el cliente gasta, emplea o consume en cada accionar con la compañía. De esta manera, se puede reformular el modelo, para el caso de los reclamos en las telecomunicaciones las siguientes variables:

- Recency (R): La última vez o mes que el cliente emprendió un reclamo hacia la compañía
- Frecuencia (F): El número de meses en las que el cliente reclamó.
- Monto (M): El número de reclamos promedio involucrado en cada ocasión.

Generalmente M va asociado a un valor monetario, no obstante, acorde con J.-J. Jonker en [23], también se puede ocupar como una variable de monto de acciones.

2.2.4. Problema de clases desbalanceadas

De vez en cuando, dependiendo del tipo de mercado, aparece este problema expresado en la base de datos respecto a las clases, es decir, a las categorías o valores de la variable objetivo. Esta rareza de clases o desbalanceo de clases se da cuando existe escasez de una de las clases, esta escasez puede ser de dos tipos [18]:

- Rareza de clases: se define como la ocasión en que un valor de la variable objetivo se encuentra fuera del común denominador o en los extremos de la distribución de la variable objetivo.
- Rareza de casos: el segundo tipo corresponde a un conjunto de datos significativo pero a su vez pequeño [18], en otras palabras, son aquellas instancias que escapan al común, en cuanto a su comportamiento. Ambos tipos de escasez son consideradas una desbalanceo interno de las bases de datos. Este tipo de problema conlleva a dificultar la labor de la implementación del KDD, debido a que existen consecuencias asociadas a la ignorancia de dicha problemática, entre las cuales destacan [6, 18]:
  - Métricas de evaluación inapropiadas: En ocasiones, la métrica que ayuda a construir el modelo se basa en obtener una certeza adecuada, sin embargo, en el caso de que existe una rareza de clases del 1 %, estos modelos se construirán para obtener el 99 % de certeza, dejando de lado la clase rara que puede ser aquella de interés para el

proyecto. En otras palabras, esto indica que las clases raras tienen menor impacto en el accuracy (o certeza) que las clases comunes [18].

- Escasez de datos (Rareza absoluta y relativa): Esta consecuencia se da en las bases de datos en donde la cantidad de instancias que pertenecen a la clase rara es mucho menor con respecto al resto de las clases.
- Fragmentación del conjunto de datos: es un problema adjudicable al momento de aplicar un algoritmo en la etapa de minería de datos, porque las regularidades pueden ser solamente encontradas en particiones individuales que contienen menos datos [18], esto quiere decir que los patrones finales terminan bajo la influencia de los patrones internos de cada partición.
- Tendencia inducida: En la minería de datos para comprender el patrón general subyacente en el problema, se tiende a inducir tendencias, de hecho, muchos algoritmos de aprendizaje usan una tendencia general de manera de encontrar la generalización y evitar el sobreajuste, por lo tanto, la tendencia puede impactar la habilidad de aprender de casos o clases raras [6].
- Ruido: El ruido, cuando es consecuencia de rareza, tiene un mayor impacto sobre los casos raros que sobre los casos comunes, puesto que los casos raros tienen menos instancias para empezar, por lo tanto, requerirán menos ejemplos ruidosos para impactar el subconcepto aprendido [18].

2.2.5. Técnicas de Minería de Datos

En la etapa de minería de datos, se encuentra una variedad de modelos y perspectivas a aplicar, es en la sección modelos donde entran los algoritmos. De los cuales, se proceden a describir los más usados en el mundo de la investigación:

- K-Nearest Neighboor (KNN): El algoritmo del K vecino más cercano o KNN es uno de los algoritmos más simples. Este algoritmo no requiere de ningún parámetro fuera del número de vecinos a considerar. En pocas palabras, el algoritmo puede resumirse en que "reúne los K vecinos más cercanos y los hace votar, la clase con más vecinos gana, ..., mientras más vecinos consideramos, menor la tasa de error" [22]. Dicha cercanía, generalmente se mide en base a alguna distancia, por lo que se pueden obtener distinto resultados dependiendo de la distancia escogida, pues

diferentes métricas definirán diferentes regiones [26]. Su esquema general se propone a continuación:

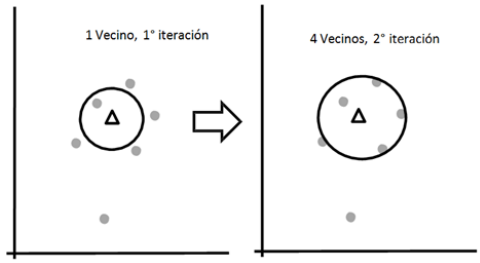

Figura 2: Modelo estructural del KNN

- Naive Bayes: En general los algoritmos de clasificación que utilizan el aprendizaje bayesiano resultan complejos en el sentido de la cantidad de parámetros. Sin embargo, el método de naive bayes convierte dicha complejidad en una simpleza factible, *"debido a que hace un supuesto de independencia condicional que reduce el número de parámetros a estimar, cuando se modela* $P(x \mid y)$*"* [36]. De forma cuantitativa, si la variable a predecir tiene dos valores pasa de estimar $2(2n-1)$ parámetros a $2n$. La utilidad de los algoritmos de aprendizaje bayesiano es que *"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"* [32]. Una explicación de la matemática subyacente de este algoritmo se encuentra en cite36.

- Árboles de Decisión: Los árboles de decisión son modelos que usualmente se representan en forma de grafos. Es *"un modelo predictivo que puede ser usado para representar tanto modelos regresivos como aquellos de clasificación, se refiere a un modelo jerárquico de decisiones y sus consecuencias"* [33]. Dentro de un esquema general, el árbol de decisión consiste en un grafo donde existe un nodo único o parental, el cual, contiene las instancias a contemplar en el modelo. Un ejemplo de este tipo de modelos es el LADTree, que es un tipo de árbol de decisión que itera sobre el ADTree que es un árbol que en vez de establecer criterios y dividir la muestra, asigna una puntuación a las categorías relevantes de determinadas variables.

- Support Vector Machines (SVM): A diferencia de los algoritmos anteriores, la máquina de soporte vectorial, o bien, Support Vector Machines, utilizan planos complejos para encontrar la mejor división de las instancias que permita clasificarlas de manera óptima. Cuya formulación es un

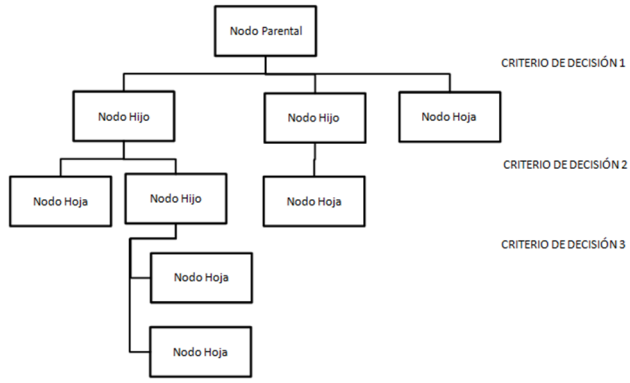

Figura 3: Modelo estructural de un Árbol de Decisión

problema de minimización cuadrática con un numero de variables igual al número de casos de entrenamiento [7, 35]: Existen además formulaciones para entrenar SVM usando programación lineal, estas formulaciones están basadas en la consideración de las normas L_1 y L_∞ en lugar de la norma L_2 [3]. Por lo tanto para grandes números de datos, se debe usar un equipo de gran capacidad. Además, las SVM utilizan la rama de optimización de la matemática, puesto que el problema que abordan involucra optimización de una función convexa [45], esto quiere decir que no contiene mínimos locales. Otra particularidad que comprende este algoritmo es que no requiere información acerca de la distribución del conjunto de datos. Es decir, se busca el balance entre certeza y cantidad de datos a aceptar. Es esta combinación la que induce el origen del problema de minimización del riesgo estructural. Cuya solución se traduce posteriormente en un problema de optimización sobre encontrar el hiperplano separador entre las instancias, mostrado en la siguiente figura:

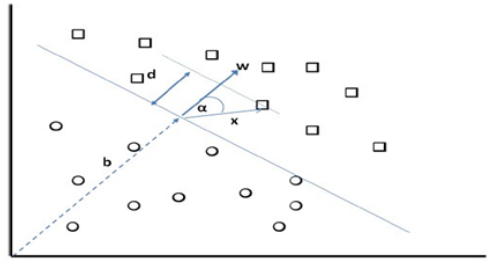

Figura 4: Modelo estructural de una SVM linealmente separable

Donde la formulación para un caso linealmente separable es:

$$\min_{w} \frac{1}{2}\vec{w}^t * \vec{w}$$

$$\text{Sujeto a } y_i(\vec{x} * \vec{w} + b) - 1 \geq 0$$

$$\forall i = 1, \ldots l$$

- Redes Neuronales: Este modelo de minería de datos es una de las estrategias más populares para aprendizaje supervisado y clasificación. Sin embargo, debido a la complejidad que posee, no se puede saber con exactitud el origen de sus resultados, lo que es una dificultad a la hora de explicar su funcionamiento. En un sentido directo, una red neuronal artificial (o denominada simplemente red neuronal, o ANN) *"consiste en procesar elementos (llamados neuronas) y las conexiones entre ellos con coeficientes(pesos) ligados a las conexiones, las cuales constituyen una estructura neuronal, y un entrenamiento y algoritmos recordatorios adjuntos a la estructura"* [24], lo que en palabras simples puede ser descrito como *"una piscina de unidades simples de procesamiento que se comunican enviando señales entre ellas sobre un gran número de conexiones ponderadas"* [25]. Un esquema general de este modelo se presenta a continuación:

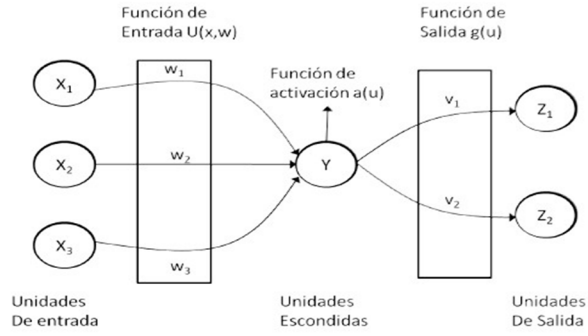

Figura 5: Modelo estructural de una red neuronal

- Regresión: La regresión, en la actualidad, consiste en *"el estudio de la dependencia de la variable dependiente, respecto a una o más variables (las variables explicativas), con el objetivo de estimar y/o predecir la media o valor promedio poblacional de la primera en términos de los valores conocidos o fijos (en muestras repetidas) de las últimas"* [19]. Por lo que este modelo sirve para predecir y clasificar, donde su uso típico

es el de predecir la demanda o el inventario futuro de una empresa. En el caso de la clasificación, la regresión que se utiliza comúnmente no resulta muy efectiva, puesto que la variable a predecir posee una connotación nominal, sin embargo, existe un tipo de regresión que se encarga de predecir variables nominales y se denomina regresión logística.

- Multiclasificadores: Los multiclasificadores, a diferencia de los modelos anteriores, busca encontrar formas o combinaciones de volver una predicción o clasificación efectiva en una predicción eficiente. Para ello se pretende explorar la mayor cantidad de caminos abordables, abarcando toda la información posible. Sin embargo, al buscar esta eficiencia, este tipo de modelos suele caer en algoritmos de gran complejidad, además, puede suceder que el modelo no sea válido a un nivel de fundamentos matemáticos.

2.2.6. Métricas de evaluación

Las medidas de evaluación técnica que generalmente se usan, se basan en una tabla de contingencia que describe las instancias predichas acertadas y erróneas. Esta tabla de contingencia se denomina matriz de confusión que *"contiene información acerca de las clasificaciones actuales y las predichas, realizadas por un sistema de clasificación"* [21]. El esquema de ésta para un caso de clasificación binaria es:

		Clase Actual	
Categorías		0	1
Clase Hipotética	0	TN	FN
	1	FP	TP
Columnas Totales		N=FP+TN	P=TP+FN

Tabla 2: Esquema de tabla de confusión caso binario

En base a esta tabla se definen las siguientes métricas de carácter técnico [6]:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{P}$$

$$Accuracy = \frac{TP + TN}{P + N}$$

$$F - Measure = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

$$Lift = \frac{Precision}{\frac{P}{P+N}}$$

Curvas Roc: Son curvas que muestran la habilidad del clasificador para posicionar las instancias verdaderas respecto a las falsas[44]. En una definición más acertada se puede decir que las Curvas ROC son las que miden la relación de la tasa de verdaderos positivos (predicciones acertadas) versus la tasa de falsos positivos (predicciones erradas). Siendo el positivo el referente a la clase de fuga cuando se trata de un problema de clasificación binario. Estas curvas no tienen una fórmula asociada. No obstante, sí tienen una métrica, la cual llamada *"Area Under the curve"(AUC)*, que se define como el rea bajo la Curva ROC, además, tiene la siguiente propiedad estadística: *"La AUC de un clasificador es equivalente a la probabilidad que el clasificador posicionará una instancia aleatoria positiva mejor que una instancia aleatoria negativa"* [15].

3. Metodología del proceso

El procedimiento KDD ejecutado se basa en una experiencia previa existente en la empresa, que permitió identificar de forma rápida y efectiva gran parte de las fuentes de información utilizadas. Un breve esquema presenta el KDD aplicado con las bases de datos que contempla:

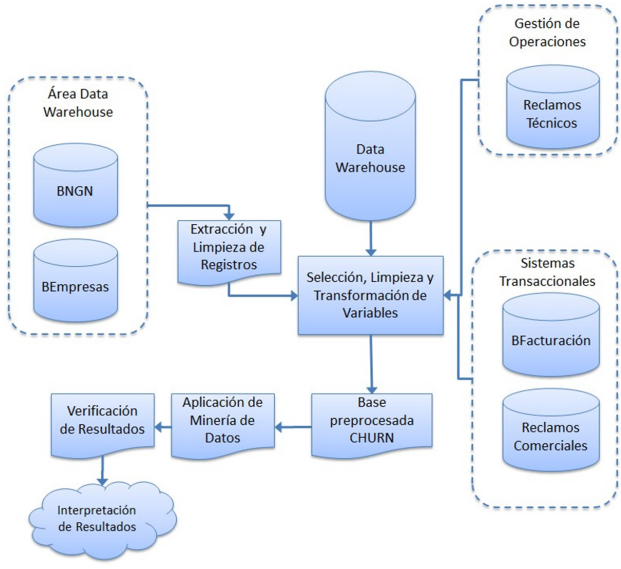

Figura 6: Esquema de procedimiento KDD Implementado

En donde las bases de datos explicitadas son descritas tal como sigue:

- Reclamos comerciales (RC): Esta base de datos contiene los registros con fecha de las solicitudes comerciales de los clientes ya sea de término de contrato, este tema no se tiene en su totalidad debido a que en esta base de datos sólo se registran las postventas en los call center, por lo tanto, es aquella que comprende las transacciones que el cliente efectúa vía telefónica con la empresa, es decir, la postventas, el cambio del servicio, término de contrato entre otros. Su periodicidad es mensual.

- BFacturación: Es la base de datos generada por el área del data warehouse destinada a temas de facturación y consumo. Su periodicidad es mensual, no obstante, su almacenamiento se mantiene cada tres meses (es decir, para obtener la base de datos correspondiente a Abril, se "borra" Enero del año en curso), es decir, si se desea averiguar la facturación del año 2009, se debe consultar directamente al data warehouse comercial.

- Reclamos Técnicos(RT): Base de datos destinada a hacer reportes acerca de los reclamos desde una perspectiva técnica. En otras palabras, en ella se encuentran todas las reparaciones a fallas en que el cliente ha avisado a la compañía. Por ende, su origen es el área de operaciones. Su periodicidad es mensual.

- BEmpresas: Esta base de datos contiene la totalidad de clientes y sus características descriptivas, es decir, su tamaño, clasificación, categoría, plan de retención, entre otros. Su origen es el área del data warehouse. Su periodicidad es mensual.

- BNGN: Esta base de datos contiene solamente los detalles de los clientes del producto NGN y sus planes correspondientes para cada cliente (con la vigencia respectiva referente a un contrato que puede contener múltiples planes). Su periodicidad es trimestral.

- Suscriptores: Proveniente directamente del DWH, es una base de datos que consta en sus registros de todos los teléfonos fijos existentes visibles en el mercado de las telecomunicaciones nacionales, por lo tanto, es una base de datos de 7,5 millones de registros. Debido a esto, se toma como una base de datos cuya variación es despreciable, es decir, estática y se obtiene por petición al área del data warehouse.

El conjunto de datos usado entregado desde distintas fuentes, poseía un total de 208 variables distribuidas en las múltiples bases de datos utilizadas. La cantidad de registros era aproximadamente 9000 clientes totales (vigentes y no

vigentes). Cabe destacar que las relaciones entre las bases de datos descritas en la figura 6 no necesariamente son de 1 a 1, puesto que un cliente puede registrar múltiples reclamos o solicitudes (RT y RC), así como también, tiene varios planes para un mismo contrato (BNGN sección paquetes), varios tickets de facturación (BFacturación) y teléfonos (Suscriptores). Las aristas que comprenden estas bases de datos son el comportamiento comercial y técnico del cliente, la información demográfica, las transacciones efectuadas y los equipos instalados. El horizonte de toma de datos contemplado para el estudio de fuga del servicio fue de 6 meses de historia.

El churn que se calcula en este paper es aquel referente al servicio, debido a que para que sea aplicable a nivel de cliente se debe implementar el KDD como procedimiento relevante en la compañía, además, ésta debe presentar una cultura más orientada a la retención en vez de a la fuerza de venta. Dicho churn en la compañía es de un 1 % aproximadamente, lo cual implica un problema de rarezas. La forma en que se trabajó dicho obstáculo comprendió desde el sobremuestreo que conllevó a caer en el overfitting de los modelos, la eliminación de los registros catalogados como fuera de rango y la segmentación de clientes y su predicción individual. Es esta última idea la que se aplicó como solución final. De las 9000 instancias se llegó a 5730 clientes, en una primera instancia, contemplados como vigentes. Otro punto relevante a destacar es que la ejecución del KDD contempló un total de 7 experimentos realizados en varios instantes de tiempo buscando medir la veracidad, la robustez y los resultados del procedimiento, dentro de estos experimentos se destacan aquel referente a un estudio histórico sobre la certeza del modelo en cuestión en ese momento (LADTree) y su evaluación técnica y comercial, así como también, el último experimento, el cual se describe primordialmente en este paper. En el experimento final declarado como cierre se ejecutó sobre 5692, una disminución que indica la migración de los clientes del servicio, lo cual se debió primordialmente a la aparición de un producto superior interno en la compañía cuyos clientes objetivos eran las empresas de mediano y gran tamaño que se encontraban insatisfechas con el producto NGN por la capacidad.

Respecto a las variables contempladas en el experimento final, se concretaron 43 variables, 17 catalogadas como nominales (incluyendo a la variable objetivo y el identificador) y 26 continuas.

Para el preprocesamiento se interpolaron variables de facturación posicionadas en la base de datos BEmpresas para obtener un indicador de los productos aparte que el cliente consumía aparte del NGN, se reemplazó por ceros para los valores ausentes de la BFacturación y para el caso de las bases de datos RT y RC. Se eliminaron variables en base a la alta varianza que poseían en el caso nominal, para el caso continuo, se eliminaron las variables con alta corre-

lación También se usó la moda para algunas variables nominales y tablas de contingencia para efectuar un reemplazo evitando la propagación de error en las relaciones inter-variables. En la base de datos de Suscriptores los valores perdidos se reemplazaron con un valor 0 solamente a modo de identificación para su posterior tratamiento.
Respecto a las transformaciones, las más relevantes, son la segmentación de planes NGN (cuya ubicación está en la sección paquetes de la base de datos BNGN) con un algoritmo de clusterización denominado Two-Step Clúster. De esta forma, se agruparon los planes y posteriormente la variable que se pudo relacionar directamente con el ID del cliente fue la cantidad de planes de un tipo determinado. Otra de las transformaciones fue la utilización del ACP para reducir la dimensionalidad de las variables provenientes de la BFacturación. Una tercera transformación se ejecutó para las variables de la RT bajo el uso del modelo RFM. Respecto a la base de datos de los Suscriptores la transformación que se utilizó detalla si el cliente tiene algún teléfono (ANI) en alguna de las compañías competidoras de la empresa que ejecutó este estudio.
En lo que se refiere a la etapa de modelamiento, se probaron varios enfoques, uno directo (base de datos + variable fuga), uno indirecto (base de datos + variable de segmentación + variable fuga), uno agregado (base de datos + variable de segmentación creada a partir de la variable fuga) y uno separado (se separa el conjunto de datos en dos en base a la variable fuga para segmentar cada subconjunto por separado y posteriormente predecir con esta variable incluida. Este último fue el usado en el experimento final y es bosquejado a continuación:

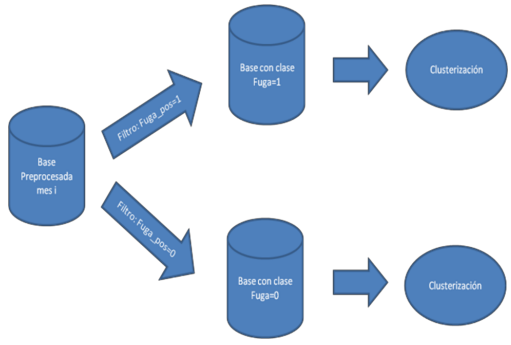

Figura 7: Modelamiento del experimento final

Para la etapa de evaluación, se contemplaron 2 perspectivas, la técnica y la comercial. La primera se orientó primordialmente en validar los resultados la mayor cantidad de veces que fuese posible, es decir, se procedió bajo la siguiente secuencia:

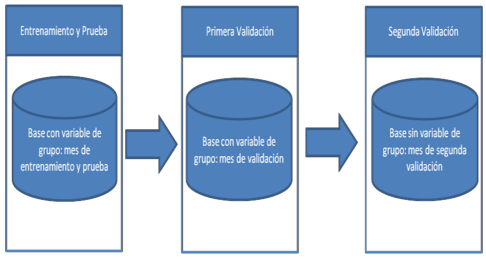

Figura 8: Esquema de evaluaciones

Donde la fase de entrenamiento-prueba es donde el modelo y aprende con datos preexistentes, la fase de primera validación es donde se predice sin conocimiento del valor de la variable objetivo, pero conociendo el grupo (del mes anterior) y al momento en que se posea dicho valor se contrasta para establecer la comparativa correspondiente. La fase de segunda validación, es en el caso en que aparte de no contar con la variable objetivo no se cuenta con la variable grupo del mes anterior. Las medidas técnicas utilizadas fueron todas las descritas en el capítulo anterior. Cabe mencionar que la predicción se midió marginalmente para ver la robustez de los modelos de minería de datos en el corto plazo (un mes).
Para la evaluación comercial se tomaron dos muestras, la primera se realizó posterior al estudio histórico efectuado en pos de obtener la veracidad en el tiempo del modelo escogido, la cual consistió en un muestreo aleatorio simple (MAS). Mientras que la segunda muestra se ejecutó en el experimento final, mediante un muestreo estratificado por grupo.

4. Aplicación experimental del proceso

En esta sección se presentan los resultados de la aplicación del proceso KKD para la predicción de fuga de clientes. Se comienza con la presentación de los resultados de aplicar transformaciones a las variables en estudio. Luego, se presentan una serie de experimentos que sirvieron como base para el experimento final. Por último, se analizan estos resultados en pos de proveer una clasificación que sea consistente en el tiempo.

4.1. Resultados de transformaciones relevantes

Previo a mostrar los resultados finales, se destacan los resultados de las transformaciones más relevantes. Para el caso de la segmentación de planes, se tomaron 5 variables: BA (Banda Ancha), velocidad, cantidad de teléfonos (ANIS), ADSL 2+(expansión del ADSL común) y Tecnología (Wiimax o de cobre). Usando el algoritmo de clusterizacion Two-Step clúster, se obtuvo lo siguiente:

Características por Grupo							
Grupo	Total	Anis [N]	Velocidad Glosa([KB])	BA	Tecnología	ADSL	Nombre
1	6422	(1 a 5)	Media (392 a 1714)	SI	COBRE	NO	Plan Estándar
2	1155	(1 a 5)	Alta (2000)	SI	MIXTO COBRE	SI	Plan ADSL
3	1518	(1 a 6)	Media Baja (587 a 685)	SI	WIMAX	NO	Plan Wimax
4	3833	(1 a 5)	Nula (0 a 0)	NO	COBRE	NO	Plan Sin Banda Ancha
5	830	(1 a 25)	Baja (0 a 179)	NO	MIXTO WIIMAX	BAJO	Plan Personalizado

Tabla 3: Resultado de segmentación de planes

Esta clusterización convertida a segmentación se realizó el mes de Julio, para la clasificación de nuevos planes se implementó un árbol de decisión simple. El número de clústeres se tomó a partir del resultado con las mismas variables de un clúster jerárquico.
La segunda transformación, es decir, el ACP, agrupo 6 variables de facturación y 6 variables de consumo en dos factores que describían la facturación y el consumo respectivamente. Los resultados que validaron el ACP para los meses acordes al último experimento, se muestran a continuación:

KMO y prueba de Bartlett				
Meses		dic-10	ene-11	feb-11
Medida de adecuación muestral de Kaiser - Meyer-Olkin.		0.9	0.9	0.92
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	178316.36	186998.92	202874.63
	gl	66	66	66
	Sig.	0	0	0

Tabla 4: Indicador KMO para el análisis ACP

Esta tabla representa que la transformación por ACP de estas 12 variables fue adecuada.

4.2. Experimentos previos

Los primeros experimentos permitieron establecer una estrategia coherente en las etapas de Integración y preprocesamiento en el KDD, además, fueron el primer acercamiento a una predicción validada. Los resultados que comparan los modelos se muestran a continuación, para el período de Marzo 2010:

	Modelos			
Criterios técnicos	J48	LADTree	Random Tree	Random Forest
Accuracy [%]	89.54 %	89.76 %	68.29 %	88.76 %
Medida F [%]	88.94 %	89.31 %	62.64 %	87.70 %
AUC	0.928	0.948	0.672	0.943
Lift [%]	232.41	231.19	164.47	238.07
TN	1545	1536	1489	1581
TP	852	867	339	795
FN	88	73	601	145
FP	192	201	248	156

Tabla 5: Comparativa entre múltiples modelos

Esta tabla se obtuvo al hacer un muestreo estratificado del total de clientes vigentes. En ella se puede apreciar que el modelo tentativo a escoger es el LADTree. Utilizando este modelo, se llegó al siguiente resultado validado para el período de Abril 2010:

Categorías predichas	Vigencia real	Fuga real	Precisión de la clase
Vigente	5587	12	99.79 %
Fuga	89	42	32.06 %
Recall de la clase	98.43 %	77.78 %	
Medida F clase 1 (Fuga)	45.41 %	F TOTAL	75.42 %
Medida F clase 0 (Vigente)	99.10 %		
Accuracy	98.24 %		
Correctamente Clasificadas	5629	98.24 %	
Incorrectamente Clasificadas	101	1.76 %	

Tabla 6: Tabla de confusión en la validación del mes de Abril de 2010

Este modelo se entrenó y probó con la base de datos del período Marzo 2010. Sin embargo, se deseaba establecer un modelo continuo que se sustentase

en el tiempo como ventana móvil. Para el período de Mayo 2010 se efectuó una segunda validación, cuyos resultados fueron:

Validación entrenando con Marzo			
	Categorías Real		
Categorías Predicción	Vigente	Fuga	Precisión de la clase
Vigente	5627	40	99.29 %
Fuga	127	1	0.78 %
Recall de la clase	97.79 %	2.44 %	
Accuracy	97.12 %		
Medida F Total	50.08 %		
Medida F clase SI (Fuga)	1.18 %		

Tabla 7: Tabla de confusión para el mes de Mayo de 2010

A partir de los cuales se desecha el procedimiento ejecutado anteriormente. Esto permitió deducir que si bien un modelo permite predecir en una ventana de tiempo determinada, no implica necesariamente que ese aprendizaje y forma de predecir se mantenga en el tiempo.

4.3. Estudio histórico

En pos de buscar una solución a la problemática anterior de que el resultado fuese sustentable en el tiempo, se realizó un estudio histórico, para el cual se rescató información del data warehouse en el caso de las bases de datos BNGN, mientras que para las otras bases de datos se solicitó información solo para los meses correspondientes. Debido a errores en el manejo operacional del data warehouse, se optó por añadir a los fugados posteriores a Julio como clientes vigentes de Junio, por ello, las bases de datos anteriores a Octubre presentan un sesgo y una mayor cantidad de clientes catalogados como vigentes. Esta situación se puede observar en el gráfico resultante de este estudio representado a continuación:

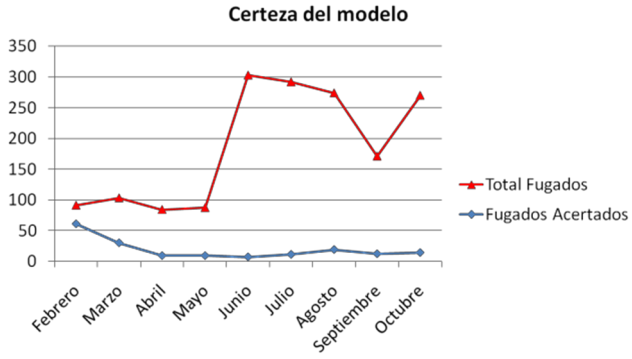

Figura 9: Gráfico histórico de certeza del modelo LADTree para un período de 11 meses

Cabe destacar que el modelo cuyos resultados fueron consistentes en el tiempo fue el LADTree, por lo tanto, se aplicó una predicción para los posibles fugados del mes de Diciembre de 2010, cuyos resultados fueron validados técnicamente con la variable fuga extraída a partir de los términos de contrato de la base de datos RC.

	Validación entrenando con Marzo		
Categorías Predicción	Categorías Real Vigente	Fuga	Precisión de la clase
Vigente	5627	40	99.29 %
Fuga	127	1	0.78 %
Recall de la clase	97.79 %	2.44 %	
Accuracy	97.12 %		
Medida F Total	50.08 %		
Medida F clase SI (Fuga)	1.18 %		

Tabla 8: Tabla de confusión post-estudio histórico

Aparte de esta validación técnica, se realizó un MAS para efectuar una encuesta telefónica para generar una validación comercial. Los resultados de este MAS se representan en la siguiente tabla:

Glosa	Cantidad de respuestas
Sin información	2
No	33
Sí	15
Universo total	50
Efectividad del modelo	30.00 %

Tabla 9: Resultados de la evaluación comercial post-estudio histórico

De esta manera se deduce que la variable fuga resulta no ser confiable para evaluar la certeza del modelo, pero sí es útil para participar en el aprendizaje del modelo, además, en un problema de desbalanceo las medidas de evaluación tradicionales pueden ser no apropiadas para medir la calidad del modelo en cuestión [18].

4.4. Experimento final

El experimento presentado a continuación corresponde a los resultados de la aplicación de un piloto. Por lo tanto, resume en cada una de sus etapas los aprendizajes adquiridos a partir de los experimentos anteriores. Las bases de datos preprocesadas contempladas fueron para los períodos de Diciembre 2010, Enero 2011, Febrero 2011 y Marzo 2011, siendo esta última aquella en la que se buscaba predecir el comportamiento de fuga de servicio. Particularmente en la etapa de modelamiento se agregó la clusterización de clientes en forma separada, tal y como se ejemplifica en la figura 7. Cabe señalar que los problemas de la marginalidad detectados en las fases anteriores fueron solucionados en base a un seguimiento de las bases de datos incrementales (BNGN) para establecer la vigencia de los clientes en cada mes sin duplicar cliente como en el caso histórico, ni desestimar la cantidad de clientes fugados, de esta manera, la distribución de fugados por meses:

Número de registros por sub-bases								
Mes	Base Diciembre 2010		Base Enero 2011		Base Febrero 2011		Base Marzo 2011	
Categoría Variable	Fuga	Vig	Fuga	Vig	Fuga	Vig	Fuga	Vig
Frecuencia	52	5457	57	5707	68	5653	0	5692

Tabla 10: Distribución de fuga por bases preprocesadas para último

Donde en el último mes no se detectan fugados por el hecho de que el experimento final requiere el cierre del mes de Abril 2011 para ver los fugados

en el mes de Marzo 2011, por lo que se asumen como vigentes. En la clusterización también se observó que los grupos establecidos en cada una de las sub-bases de datos (sub-base de datos F para fugados y NF para vigentes) contenían características similares, denominados los grupos "Grandes", "Reactivos" y "Pasivos" para cada sub-base de datos. Para la elección del modelo en la etapa de minería de datos, se optó por un modelo estilo regresivo u SVM, debido a que aquellos de aprendizaje por reglas (Árboles de decisión) y probabilísticos (Bayes) no habían entregado resultados suficientemente buenos y válidos. La comparativa del modelo final con el resto se muestra a continuación:

Modelos	TN	FP	FN	TP	Accuracy	Recall clase 1	Precision clase 1	Medida F clase 1
SVM 6	4949	15	0	67	99.70 %	100.00 %	81.71 %	89.93 %
SVM 10	4948	16	2	65	99.64 %	97.01 %	80.25 %	87.84 %
SVM 3	4872	92	0	67	96.05 %	100.00 %	42.14 %	59.29 %
Naive Bayes	4666	298	1	66	94.06 %	98.51 %	18.13 %	30.63 %
LADTree	4933	31	0	67	99.38 %	100.00 %	68.37 %	81.21 %

Tabla 11: Comparativa de modelos para último experimento

En esta tabla, SVM hace referencia a las Support Vector Machines, para cada ejecución del SVM se asignan distintos parámetros, SVM 3 es aquella configuración donde el kernel es rbf (Radial basis function),$C = 0$, y $\gamma = 0{,}5$, en SVM 6 el kernel es rbf, $C = 0$, y $\gamma = 0{,}0$ y en SVM 10 el kernel es rbf, $C = 10$, y $\gamma = 0{,}95$, donde C es el parámetro que controla la compensación entre errores de entrenamiento y generalización y γ es el error de clasificar erróneamente. De estos modelos se escoge el SVM 3 (por la cantidad de fugados posibles que predice para Marzo 2011 más que por los resultados en la base de datos preprocesada de Febrero 2011). De esta manera se obtiene tras la segunda validación los siguientes resultados:

	Categorías Reales		
Categorías Predichas	Vigente	Fuga	Precisión de la clase
Vigente	5492	38	0.993 %
Fuga	157	5	0.031 %
Recall	0.972	0.116	
Accuracy	96.57 %		
Medida F Total	52.76 %		
Medida F clase 1(Fuga)	4.88 %		

Tabla 12: Tabla de confusión en el último experimento

Adicional a las métricas usuales se agrega la Curva ROC correspondiente cuyo valor de área bajo la curva es 0.975 y su gráfico es:

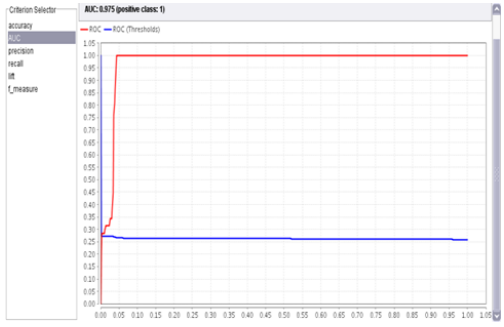

Figura 10: Gráfico de curvas ROC para modelo SVM seleccionado (gamma=0 y C=0)

Esta métrica presenta la calidad del modelo, no obstante, la tabla de confusión señala que no lo es, lo que conlleva a dudar de las métricas anteriores, esto se justifica para un problema de rarezas, en donde Weiss en [18] sugiere utilizar esta última métrica a modo de evaluación técnica, puesto que las otras no contemplan el efecto de la fuga de clientes como clase desbalanceada. Por consiguiente, se ejecuta una evaluación comercial vía telefónica, en donde los clientes objetivos fueron seleccionados a través de un muestreo estratificado por la categorización establecida por la clusterización del mes anterior (en este caso el clúster al que pertenecían en Febrero). Los resultados de esta encuesta se bosquejan posteriormente:

Descripción de ANIS	Cantidad	Porcentaje %
ANIS Vigentes	21	48.837
ANIS Sin tono	14	32.558
ANIS Contactado que declara mala atención	1	2.326
ANIS Con tono pero por sistema se retira	7	16.279
Total de Anis en la muestra	**43**	**100**
ANIS en Vigencia	21	48.837
ANIS en Retiro	22	51.163

Tabla 13: Resultados de evaluación comercial en el último experimento

Donde el concepto ANIS se refiere a número de teléfono. Esta encuesta señala que aquellos clientes que se intentó retener ya habían desconectado sus servicios, esto se debió a que el proceder de la encuesta se vio retrasada por los cambios organizacionales de la empresa en la que se desarrollaba el piloto. No obstante, declara que la certeza del modelo es cercana a un 51 %, distante del 5 % que señala la medida F como medida técnica. Es decir, se comprueba que el

error posiblemente se encuentra en la variable escogida como fuga (en este caso la variable original señala el término de contrato del producto). Con lo anterior, se puede deducir que la metodología aplicada muestra que en caso de tener el problema de rareza y, además, se tenga una alta presencia de valores fuera de rango con información útil, no son aplicables las técnicas de submuestreo, ni sobremuestreo como los demostrados en [6], y las técnicas de aplicación de pesos como las mostradas en [34], requieren de un estudio sobre la asignación para cada instancia (en este caso con alto porcentaje de valores fuera de rango), lo cual se traduce en un procedimiento de alta complejidad. También resulta infactible en un entorno de multiplataforma efectuar un procedimiento riguroso contemplando el costo que posee cada cliente fugado de un servicio particular como el propuesto en [10], puesto que los datos de facturación se encuentran dispersos en las distintas plataformas por las que se manejan los datos del cliente, a lo que se agrega la veracidad completa de dichos datos.

5. Recomendaciones

En este paper se cuenta con una experiencia previa que permitió el conocimiento de las fuentes de información, mas no así de las múltiples plataformas existentes como parte del proceso del producto de telefonía fija. En el caso de un producto cuyos datos se distribuyen entre distintas plataformas, la etapa de integración es la más relevante, por ende, se propone una serie de pasos que permiten una integración correcta:

- Investigar las variables relacionadas con el producto directamente con el área del producto.
- Conciliar dichos datos con otras variables pertenecientes a otras áreas (por ejemplo, facturación, reclamos, órdenes de trabajo, etc.)
- Averiguar si existen las variables fugas, renegociación y migración dentro del servicio, en caso de que no existan efectuar un seguimiento manual de ellas o proponer que se inserte como requerimiento informático dentro de las plataformas.
- Dado que la interpretación también es relevante en un proyecto con el KDD aplicado transversalmente se sugiere originar una descripción de las instancias que permita establecer causales de la fuga voluntaria, de tal manera de generar retenciones con valor agregado y no depender del modelo a utilizar. Una herramienta relevante para estos casos es la segmentación, debido a que bosqueja una exploración rápida del mercado

y permite una familiarización con el producto o servicio. Además, sirve como integrador entre bases de datos cuyas relaciones son de 1 a n (por ejemplo, un cliente muchos planes).

- Establecer la naturaleza de las bases de datos que cada plataforma genera, es decir, si la base de datos es marginal (dinámica) o incremental, en particular, se concluye que usar bases de datos operacionales instantáneas (no sujetas a cambio) son más útiles que usar bases de datos incrementales cuando se trata de datos transaccionales, así como también, la fuga correspondiente. Además, para llevar a cabo un proyecto de minería de datos se debe contar con dos tipos de fuentes de datos como mínimo y estas son un Data Warehouse (o algunas bases de datos incrementales) y las bases de datos marginales, donde las primeras no implican las segundas.

- Si se tiene una base de datos incremental y se desea establecer la marginalidad, este paper señala que es posible, mas el resultado presenta un comportamiento erróneo en pequeña escala agregando valores fuera de rango o con mayor porcentaje de valores perdidos lo que dificulta establecer un modelo no influenciable por la temporalidad de los datos. En otras palabras, si un Data Warehouse se implementa bajo una cultura que prescinde de otras bases de datos para darle prioridad, no tendrá la misma confiabilidad que presenta en la teoría.

En este trabajo, se descubre que la fuga de clientes del servicio, calculado en la empresa, tiene dos características relevantes: primero no es la fuga de clientes (pues un cliente en telefonía fija tiene muchos productos y servicios) y segundo, no se sabe qué clientes se van, es decir, el cálculo es global. No existe una única forma de bosquejar la variable que describe la fuga, no obstante, se entregan distintas perspectivas que pueden resultar útiles:

1. Verificar la base de datos relacionada con el call center de la empresa para observar si posee registro de fugas y analizar la cantidad de clientes que usan ese canal para terminar con el servicio de la compañía.

2. Verificar si existe una base (de datos) con los datos del contrato en toda empresa, en caso de que exista, validar la veracidad de los campos que contiene así como también, la temporalidad del mismo, si está explícita en una fecha de modificación (válida) se sugiere usar esta base de datos como principal.

3. En caso de que no se tenga registro en la empresa del cálculo de la variable fuga voluntaria individual, se aconseja calcularla en base a los registros

que desaparezcan o cambien su estado en una base de datos incremental consolidada en un período establecido para el estudio.

Las perspectivas anteriores aplican netamente a casos en los que las bases de datos trabajadas por el área encargada del producto sean de naturaleza incremental. Así como el cálculo de fuga voluntaria es clave, también lo es el establecimiento de las instancias vigentes, el cual, requiere de una estandarización o variables que declaren la certeza de su vigencia. Dentro de las consecuencias inmediatas de los errores de vigencia se encuentran las apariciones de valores perdidos y fuera de rango. Posterior a la aplicación del Preprocesamiento y la Transformación. El modelamiento plantea un desafío en cuanto a la cantidad de aprendizaje facilitado para el modelo, cada cual tiene su capacidad, pero en el mercado de las telecomunicaciones se suele dar una fuga de clientes mensual baja, lo que conlleva a la aparición del problema de rarezas o clases raras. El aplicar técnicas de muestreo simples (sub o sobre muestreo) así como también, aplicar sin muestreo conllevan a resultados no viables en el tiempo, por lo que en este paper se muestra que la segmentación válida de clientes y su seguimiento permiten mitigar este problema. La etapa de evaluación en el KDD se presenta como solamente de tipo técnica, no obstante, aparte de esta perspectiva se concluye que la vista comercial puede entregar validaciones con mayor confiabilidad que las técnicas en estos casos de sistemas con multiplataforma, puesto que tanto la variable objetivo como las instancias vigentes poseen pérdida de información al provenir de distintas bases de datos y plataformas.

6. Conclusiones

Este trabajo presenta una descripción metodológica de la aplicación de KDD para predecir la fuga de clientes en una empresa de telecomunicaciones. Por lo tanto, busca ser una ayuda a la hora de implementar un proyecto con características similares. Uno de los principales aprendizajes de este trabajo corresponde a la determinación de la integración como una de las etapas más importantes del proceso KDD para ser aplicado en empresas con información repartida en múltiples plataformas, debido a que esta etapa determina el resto de los resultados. Además, con la integración correcta aparece otra serie de problemas relacionados con variables puntuales, el trato de valores perdidos y las transformaciones respectivas. Se concluye que la clave, para llevar a cabo ambas etapas posteriores en el KDD sin error, es la variable a predecir en este caso la fuga del servicio, para lo cual se debe bosquejar la situación actual de cómo se calcula esta. Finalmente, otra de las conclusiones de este paper es que

tanto el preprocesamiento como la transformación se deben ejecutar de manera rigurosa en estas situaciones, porque el hecho de que los datos se encuentren disgregados en múltiples plataformas facilita la pérdida interna de valores para la mayoría de las variables por lo que un valor perdido no debe ser eliminado, sino que reemplazado, estimado o ignorado. Como resultado de nuestra aplicación se logró consolidar información de multiples plataformas y corroborar su calidad para ser utilizada en los modelos predictivos. Utilizando estos datos se realizaron varios experimentos entre Marzo del 2010 y Marzo 2011 para predecir la fuga de clientes, utilizando diversos algoritmos como SVM, LADTree, Naive Bayes, J48, Random Tree, entre otros. Con estos resultados fue posible construir un benchmark amplio, de tal modo de poder descubir que modelo se comporta mejor para predecir la fuga de clientes. Finalmente, se realiza un experimento tomando los RUTs de los clientes predichos como posibles fuga por el mejor modelo; y se realizó una encuesta para validar si efectivamente esto era cierto. Los resultados de la encuesta muestran que el modelo acierta en un 51 % de los RUTs predichos como fuga. Esto es un muy buen resultado pues aunque el 51 % de aciertos parece bajo para un modelo predictivo. El modelo permite que la empresa en lugar de generar acciones al azar de la base de 9000 clientes del producto NGN, simplemente se focalice en un número reducido de clientes empresas (114 en nuestro esperimento final) que el modelo predijo como posibles fugas y por ende, generar estrategias más personalizadas para aumentar su retención. Como trabajo futuro se propone la utilización de esta metodología para el cálculo de fuga de clientes con un periodo de observación mucho mayor con el fin de medir la capacidad predictiva de los modelos propuestos que solo fueron implementados a nivel de plan piloto.

Agradecimientos.

Los autores desean agradecer el continuo soporte del Instituto Sistemas Complejos de Ingeniería (ICM: P-05-004- F, CONICYT: FBO16) (www.isci.cl). Así mismo nos gustaría dar las gracias al Sr. Andrés Chacón, Sr. Eduardo Duran y a la Sra. Sandra Molina por su dispocisión para el buen desarrollo de este trabajo. Finalmente, a Don Ricardo Muñoz por su valiosa ayuda para editar la versión final de este trabajo.

Referencias

[1] J. Álvarez Menéndez. Minería de datos: Aplicaciones en el sector de las telecomunicaciones. Technical report, Universidad Carlos III, 2008.

[2] V. Barnett and T. Lewis. Outliers in Statistical Data. Wiley, 1994. 584 P.

[3] J. Pedroso, N. Murata. Support vector machines for linear programming: motivation and formulations. BSIS Technical Report, August 1999.

[4] Broadband stimulus kickstarts ICT industry growth del sitio: 2010 ict market review & forecast extraído el 29 de octubre del 2010 fuente: http://www.tiaonline.org/market intelligence/mrf/webinar/tia broadband webinar 20100319 final.pdf.

[5] B. Q. Huang, T. M. Kechadi, B. Buckley, G. Kiernan, E. Keogh, and T. Rashid. A new feature set with new window techniques for customer churn prediction in land- line telecommunications. Expert Syst. Appl., 37:3657– 3665, May 2010.

[6] J. Burez and D. Van den Poel. Handling class imbalance in customer churn prediction. Expert Syst. Appl., 36:4626–4636, April 2009.

[7] S. Maldonado, R. Weber. Modelos de Selección de Atributos para Support Vector Machines. Revista Ingeniería de Sistemas, Volumen XXVI, Septiembre 2012.

[8] S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering /segmentation algorithms. Technical report, IEEE, 2003.

[9] S. Chaudhuri and U. Dayal. An overview of data warehousing and olap technology. SIGMOD Rec., 26:65–74, March 1997.

[10] Y. Jie Dong, X. hua Wang, y J. Zhou, Cost BP Algorithm and its Application in Customer Churn Prediction, in 2009 Fifth International Joint Conference on INC, IMS and IDC, Seoul, South Korea, 2009, págs. 794-797.

[11] C. Bravo, S. Maldonado, R. Weber. Experiencias prácticas en la medición de microempresarios utilizando modelos de credit scoring. Revista Ingeniería de Sistemas, Volumen XXIV, Junio 2010.

[12] A. Farhangfar, L. Kurgan, and J. Dy. Impact of imputation of missing values on classification error for discrete data. Pattern Recogn., 41:3692–3705, December 2008.

[13] Estadísticas de inversión y empleo, sitio: Subtel (subsecretaría de telecomunicaciones). Extraído el 20 de octubre del 2010. fuente: http://www.subtel.cl/prontus subtel/site/artic/20100608/asocfile/20100608122246/1 series inversión y empleo dic09 191010 v1.xls.

[14] Estadísticas de reclamos recibidos por el dpto. Gestión de reclamos de la subtel. sitio: Subtel. Extraído el 20 de octubre del 2010. Fuente:http://www.subtel.gob.cl/prontus_oirs/site/artic/20100503/asocfile/20100503154918/estadisticas_reclamos_2010.pdf

[15] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth. From data mining to knowledge discovery in databases. AI Magazine, 17:37–54, 1996.

[16] T. Fawcett. Roc graphs: Notes and practical considerations for researchers. Technical report, HP Laboratories, 2004.

[17] M. Galván and F. Medina. Imputación de datos: teoría y práctica. Technical report, CEPAL Naciones Unidas, 2007.

[18] G. M. Weiss. Mining with rarity: a unifying framework. SIGKDD Explor. Newsl., 6:7–19, June 2004.

[19] D. N. Gujarati. Econometría. McGraw Hill, 2003. 921 P.

[20] D. Hand. Data mining: Statistics and more. The American Statistician, 52:112–118, 1998.

[21] G. Huerta. Balanceo de datos para la clasificación de imágenes de galaxia. Technical report, Universidad Politécnica de Puebla, 2010.

[22] Y. Jiangsheng. Method of k-nearest neighbors. Technical report, Institute of Computational Linguistics, Peking University, 2002.

[23] J.-J. Jonker, N. Piersma, and D. V. d. Poel. Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. Working papers of faculty of economics and business administration, ghent university, belgium, Ghent University, Faculty of Economics and Business Administration, 2003.

[24] N. K. Kasabov. Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering. MIT Press, Cambridge, MA, USA, 1st edition, 1996. 550 P.

[25] B. Kröse and P. van der Smagt. An introduction to Neural Networks. None, 1996. 135 P.

[26] L. I. Kuncheva. Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, 2004. 350 P.

[27] J. Larsen , J. Rossat, D. Ruta, and M.Wawrzynosek. Customer loyalty, a literature review and analysis. Technical report, UNIPEDE, 1998.E. H.

[28] M. A. P. M. Lejeune. Measuring the impact of data mining on churn management. Internet Research, 11(5):375–387, 2001.

[29] J. Lévy Mangin and J. Varela Mallou. Análisis multivariable para las ciencias sociales. Prentice Hall, 2003. 896 P.

[30] R. J. A. Little and D. B. Rubin. Statistical Analysis With Missing Data. Probability and Statistics. Wiley, New Jersey, second edition, 2002. 381 P.

[31] J. Lu. Predicting customer churn in the telecommunications industry an application of survival analysis modeling using sas. Technical report, Sprint Communications Company, 2001.

[32] C. M. Luque. Clasificadores bayesianos. El algoritmo naive bayes, 2003.

[33] L. Rokach and O. Maimon. Data Mining With Decision Trees:Theory And Applications. World Scientific Publishing, 2008. 244 P.

[34] Geoffrey J. McLachlan, V. Nikulin, «Classification of Imbalanced Marketing Data with Balanced Random Sets», JMLR: Workshop and Conference Proceedings 7: 89-100, KDD Cup 2009

[35] S. Maldonado. Utilización de support vector machines no lineal y selección de atributos para credit scoring. Master's thesis, Universidad de Chile, 2007. 118 P.

[36] T. M. Mitchell. Generative and discriminative classifiers: Naive bayes and logistic regression. In Machine Learning, 2010.

[37] S. Molina. Aplicación de técnicas de minería de datos para predicción del churn de clientes en una empresa de telecomunicaciones. Master's thesis,

Escuela de Ingeniería de la Pontificia Universidad Católica de Chile, 2009. 114 P.

[38] R. Mattison. Data Warehousing and Data Mining for Telecommunications. Artech House, Inc., Norwood, MA, USA, 1st edition, 1997. 282 P.

[39] J. A. Ortega Ramírez. Patrones de comportamiento temporal en modelos semicualitativos con restricciones. PhD thesis, Universidad de Sevilla, 2000.

[40] M. Richeldi and A. Perrucci. Analyzing churn of customers.

[41] P. Ponniah. Data Warehousing Fundamentals. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 2001. 518 P.

[42] D. B. Rubin. Inference and missing data. Biometrika, 63:581–590, 1976.

[43] J. Miranda, P. Rey, R. Weber. Predicción de Fuga de Clientes para una Institución Financiera mediante Support Vector Machines. Revista Ingeniería y Sistemas, Volumen XIX, Octubre 20005.

[44] J. Wang, editor. Data mining: opportunities and challenges. IGI Publishing, Hershey, PA, USA, 2003.

[45] L. Wang. Support Vector Machines: Theory and Applications (Studies in Fuzziness and Soft Computing). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. 431 P.