# Modelo de predicción de fuga de clientes en telefonía móvil prepagada

Presentado por
**Camilo E. Baena Blanco**



**Fundación Universitaria Los Libertadores**
Facultad de Ingeniería y Ciencias Básicas
Especialización en Estadística Aplicada
Bogotá D.C, Colombia
2018

# Modelo de predicción de fuga de clientes en telefonía móvil prepagada

Presentado por

**Camilo E. Baena Blanco**

en cumplimiento parcial de los requerimientos para optar al título de

## Especialista en Estadística Aplicada

*Dirigida por*

**Juan C. Santana**

Profesor

**Fundación Universitaria Los Libertadores**

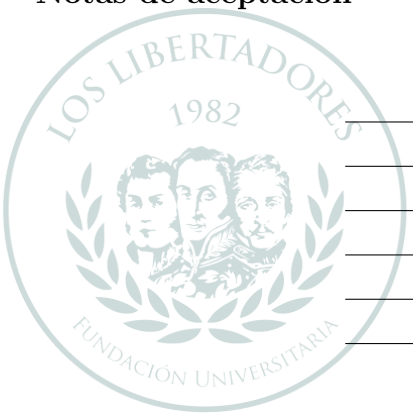
Facultad de Ingeniería y Ciencias Básicas

Especialización en Estadística Aplicada

Bogotá D.C, Colombia

2018

**Notas de aceptación**






Firma del presidente del jurado

Firma del jurado

Firma del jurado

Bogotá DC, Noviembre de 2018.

LOS LIBERTADORES

FUNDACIÓN UNIVERSITARIA

Las directivas de la Fundación Universitaria Los Libertadores, los jurados calificadores y el cuerpo docente no son responsables por los criterios e ideas expuestas en el presente documento. Estos corresponden únicamente a los autores y a los resultados de su trabajo.

# Dedicatoria

Dedicado a Dios, mi familia y para cada uno se esos amigos que han cruzado en mi camino, simplemente por que compartir es vivir...

# Agradecimientos



Agradezco a la Fundación Universitaria Los Libertadores por la oportunidad que me ha brindado de realizar la presente especialización en estadística, a mis compañeros que estuvieron apoyándome en el transcurrir de las jornadas académicas, y brindo un agradecimiento especial al profesor Juan C. Santana, quién brindó la directriz y la asesoría necesaria para desarrollar el presente trabajo.

# Índice general

# Índice de figuras

# Índice de cuadros



**SVM** Máquinas de soporte vectorial

# Modelo de predicción de fuga de clientes en telefonía móvil prepagada

## Resumen

En el presente estudio se analizan los datos de una muestra de clientes, en una de las cinco compañías más importantes de telefonía móvil en Colombia, con los cuales se usan técnicas estadísticas para construir un modelo de clasificación que permita identificar los clientes que presenten comportamientos determinados por el uso del servicio, similares a los que presentaron los que ya se tienen como fugados, a través del aprendizaje supervisado, se aprovechan los datos históricos de clientes etiquetados como fugados y también como activos (refiriéndose a los usuarios que permanecen con el servicio), para construir un modelo que permita hacer una clasificación de los usuarios con el menor margen de error posible.

**Palabras claves:** telefonía móvil prepagada, aprendizaje supervisado, regresión logística, redes neuronales, máquinas de soporte vectorial

# Capítulo 1

# Introducción

En el presente trabajo se busca determinar un modelo estadístico que permita clasificar los clientes de una compañía de telefonía móvil, aplicando técnicas como regresión logística, redes neuronales y máquinas de soporte vectorial, además hacer una caracterización de los clientes que se fugan y de los que permanecen encontrando las variables mas significativas que representan estos comportamientos.

La fuga de clientes se identifica en las compañías de telefonía móvil cuando un cliente presenta un tiempo de inactividad mayor a noventa días, esta inactividad significa que el usuario no usó la línea móvil para realizar o recibir llamadas o mensajes de texto y tampoco para consumir datos para navegar en internet o redes sociales. A partir de este hecho, se clasifican principalmente dos tipos de usuarios, los que desisten del uso del servicio y los que permanecen con el uso regular de los servicios que pueda ofrecer la compañía, todo sobre la condición de no cumplir los 90 días de inactividad.

A partir de los datos que se logran almacenar de cliente, donde se registran todas las operaciones que el usuario realiza con su línea, a este tipo de bases de datos de tráfico se le denominan como CDR (call detail record). Estas bases de datos almacenan masivamente la información de forma recurrente y transaccional, esto quiere decir que un cliente al terminar una conversación a través de su teléfono móvil, o de utilizar su red social favorita, se genera un registro que guarda el tiempo en segundos de la llamada o la cantidad megabytes que se gastaron.

Al disponer de la fuente de datos, se deben establecer variables referentes al consumo por cada servicio que se ofrezca, este quizás puede llegar a ser un paso de los más importantes y complejos, debido a que se debe asegurar la calidad y la completitud de la información que se va a extraer de la base de datos, para esto se toma provecho de los motores de bases de datos, con los cuales a través del lenguaje SQL (Structured Query Language), se pueden construir consultas que transformen los datos en conjuntos de información que agrupen variables, características o acumulados interesantes por cada cliente.

De esta forma de construyen variables correspondientes al total de tiempo de llama-

das, total de megabytes consumidos, cantidad de mensajes de texto enviados, productos comprados, etc, estas variables con acumuladas de forma semanal, con la intención de conocer las últimas ocho semanas de cada usuario. De forma similar que en otros trabajos realizados se usan este tipo de variables para conocer y de alguna forma representar el comportamiento individual de la base de clientes que disponga este tipo de compañías.

A partir de de tener un conjunto de datos organizado y refinado, se toma una muestra aproximada de 34.000 clientes los cuales se etiquetan si permanecieron activos o se fugaron, ya con estos datos etiquetados. Para los cuales se hace una caracterización de los comportamientos relevantes de los clientes que se fugan y de los que permanecen activos a través del análisis de componentes principales y correspondencias múltiples, y finalmente se implementan tres tipos de modelos de clasificación donde se encuentra que la técnica SVM (máquinas de soporte vectorial) es la que presenta una precisión del 87.86 %

# Capítulo 2

# Planteamiento del Problema

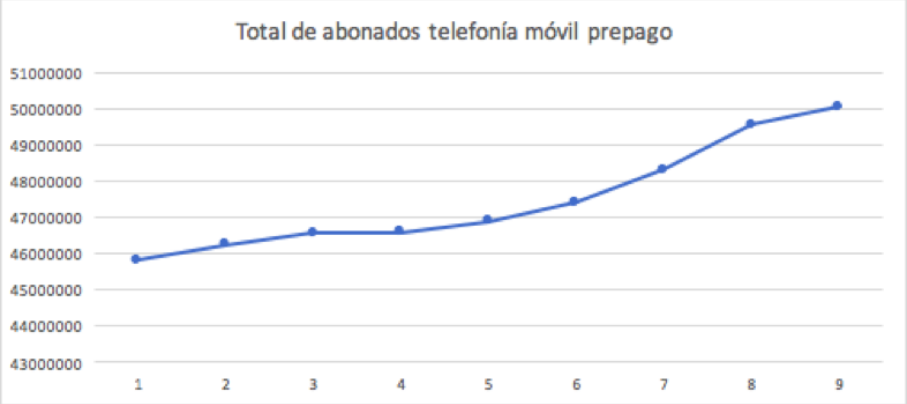
Con el avance imparable de la tecnología, los mecanismos de comunicación han generado una necesidad importante en los seres humanos de estar interconectados todos el tiempo, en los últimos quince años los dispositivos móviles específicamente los teléfonos inteligentes se han propagado a nivel mundial dadas las múltiples tareas que se logran hacer a través de ellos, con los cuales se tienen muchas más funciones que simplemente realizar o recibir llamadas o mensajes de texto, en la actualidad se tiene la posibilidad de conectarse a redes sociales, gestionar las cuentas de correo electrónico, hacer transacciones bancarias o hasta realizar tus compras del hogar, etc. En fin, se tienen infinidad de funcionalidades y propósitos que se le pueden dar a estos teléfonos inteligentes.

En las compañías de telefonía móvil y demás empresas que ofrecen servicios resulta mas costoso conseguir nuevos clientes que mantener con los que ya se cuenta, dado esto evitar la perdida o fuga de estos clientes se convierte en una tarea crucial en los objetivos del negocio. Con el propósito de establecer estrategias para el tratamiento adecuado y a tiempo de sus usuarios estas compañías hacen uso de datos que pueden recopilar para conocer un poco más sus clientes aplicando técnicas estadísticas que faciliten el entendimiento y la interpretación de comportamientos presentes en el uso del servicio de telefonía.

En Colombia se tienen al rededor de cincuenta millones de líneas móviles, lo cuál representa una penetración que crece rápidamente año a año, tomando partida de este hecho, existen compañías de telefonía móvil que ofrecen este servicio para que los usuarios puedan adquirir productos que permitan comunicarse no solo a través de llamadas o mensajes de texto sino también del uso de datos. Estas empresas se distribuyen la participación en el mercado lo cuál crea una competencia entre ellas para adquirir nuevos clientes y sobre todo, mantener y fidelizar los que cada compañía tiene. Esta cifra de abonados de telefonía viene en un acelerado crecimiento en los dos últimos años como lo podemos ver en la figura 1, de acuerdo a los reportes generados trimestralmente desde el ministerio de

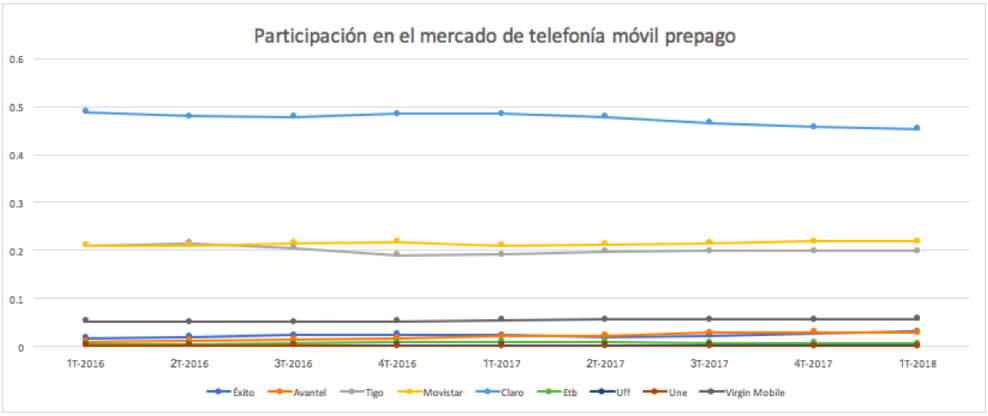
tecnologías de la información y la comunicación MinTic[1].

Figura 2.1: Abonados telefonía móvil




La participación en el mercado de telefonía móvil prepago en Colombia se constituye por nueve compañías que prestan este servicio concentrado principalmente en oferta de voz y datos, estas compañías son: Claro, Movistar, Tigo, Virgin Mobile, Avantel, Etb, Uff, Une y Éxito. Así como se puede observar en la figura 2, es evidente que la mayor participación en el mercado la tiene Claro con un 48 %, le sigue Tigo y Movistar con un 20 % cada uno, Virgin Mobile tiene 5 % y el otro 5 % lo comparten el resto de empresas. Se observa un dominio importante del sector por parte de las primeras cinco compañías, por lo tanto, es importante que los participantes de este sector dispongan de estrategias que vayan más de allá de suministrar un servicio de calidad.

Figura 2.2: Participación operadores de telefonía móvil

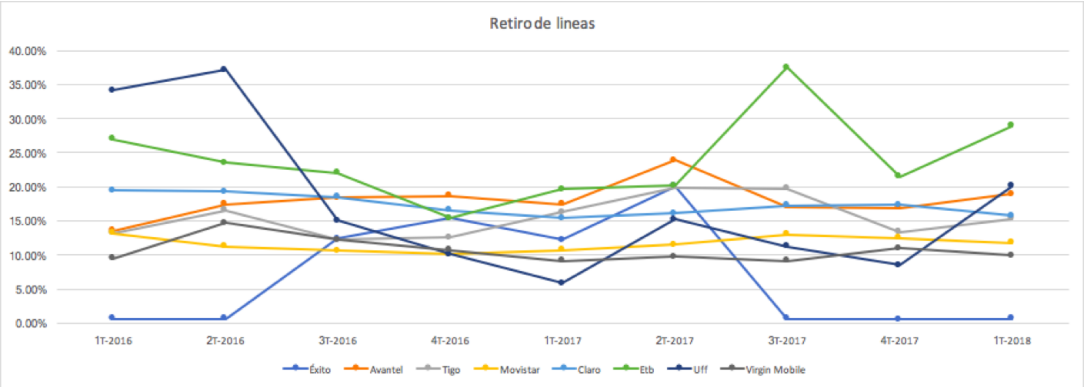



La parte del reporte de mayor interés para efectos de este trabajo tiene que ver con los retiros que reportan cada una de estas compañías trimestralmente al MinTic. En la figura 3 se evidencia que proporcionalmente la pérdida de clientes que presenta el sector está

[1]Consulta reportes trimestrales MinTic en:https://colombiatic.mintic.gov.co/679/w3-multipropertyvalues-36376-36410.html

entre el 10 % y 20 % en tres meses, por lo tanto las campañas de retención y fidelización están a la orden día para mitigar en lo posible este fenómeno, la diferencia entre estas compañías es básicamente el músculo financiero que apalanca cada operación, por lo tanto el presupuesto que cada empresa dispone para implementar costosas campañas de mercadeo estará sujeto a la capacidad que financieramente posean.

Figura 2.3: Retiros telefonía móvil




Ante el escenario evidente de tener clientes que se retiran o desisten del servicio, es necesario conocer el comportamiento de los clientes básicamente con dos objetivos, el primero es caracterizar y comprender las tendencias o patrones en sus consumos y uso; y en segundo lugar es lograr establecer segmentos de clientes los cuáles sean de interés para aplicar el tipo de campaña que se requiera, esto con la intención de minimizar costos pero sin dejar de tratar los subscriptores que para este análisis serían los que pueden presentar indicios de fuga. En este punto es donde distintas técnicas estadísticas toman lugar, las cuales, a través del análisis de los datos se construyen modelos que apoyan las decisiones estratégicas del negocio.

Este fenómeno de fuga de clientes en la telefonía móvil prepagada ha sido estudiado activamente aproximadamente desde el año 2000, dado al desarrollo tecnológico y oferta de nuevos servicios en comunicaciones a nivel mundial. Existen gran variedad de técnicas estadísticas que permiten hacer análisis sobre los datos, en este caso se necesita aprovechar de los algoritmos que suministran métodos de clasificación, dado que para efectos de este estudio se va a establecer una forma adecuada de hacer la clasificación de los clientes que se retiran y los clientes que permanecen, de esta forma construir un modelo que permita la identificación de clientes propensión a presentar fuga con un margen de error aceptable. Para esto se exploran algoritmos como regresión logística, redes neuronales y máquinas de soporte vectorial.

## 2.1 Objetivos

### 2.1.1 Objetivo General

Desarrollar un modelo estadístico que aplique técnicas de aprendizaje de máquina que permita identificar los clientes propensos a presentar fuga, a través de la recopilación de datos históricos.

### 2.1.2 Objetivos Específicos

- Caracterizar por medio de los datos suministrados los comportamientos relevantes de los clientes que se fugan y los que permanecen.
- Determinar las variables más significativas que representan la fuga de clientes.
- Comparar al menos tres tipos de modelos que ofrezcan un mecanismo clasificación.
- Seleccionar el mejor modelo (que presente la tasa más baja de error para clasificación)

## 2.2 Justificación

La fuga de clientes ha sido ampliamente estudiada a nivel mundial para todo tipo de negocios, especialmente aquellos que ofrecen servicios o productos que tienen alguna relación de subscripción, sin embargo, en Colombia la literatura que se ha encontrado hasta el momento sobre el estudio de la telefonía móvil es poca, concentrándose en trabajos que estudian los determinantes que fidelizan los clientes de telefonía móvil como por ejemplo el trabajo realizado por (Villamarin J, 2017)[1], donde hace un estudio de las características de los clientes de la compañía Claro.

Se debe tener en cuenta que la telefonía móvil se presta como un servicio que tiene especialmente dos modalidades, las cuales son prepago y postpago, considerando el trabajo realizado por (Villamarin J, 2017)[1], en Colombia y otros trabajos en Suramérica encontrados como los de (Alvarado J, 2011)[2] y (Contreras, 2017)[3] en Chile, se han realizado para telefonía móvil de modalidad pos pago, que aunque aplican técnicas similares a las usadas en este trabajo, el enfoque de trabajar en el estudio de la telefonía prepagada hace que estos trabajos no logren ser un referente importante para este estudio.

Por otro lado se referencian el trabajo de (Aimee,2017)[4], en el cuál incorporan variables correspondientes a la red de contactos de los usuarios de una compañía de telefonía móvil prepagada en Bélgica, haciendo una comparación de varios algoritmos como lo son regresión logística y redes neuronales obteniendo un AUC de 0.88. En Turquía, también se encontró un trabajo (Zehra, 2017)[5] similar de predicción de fuga de clientes de un operador móvil, en el cuál comparan dos grupos de variables, un grupo con variables de consumo y uso, y otro grupo utilizando segmentación RFM, donde a través del uso de regresión logística obtienen una precisión de clasificación del 85 % para el primer grupo y 96 % utilizando la segmentación RFM.

Dado el escenario, en el cuál se ha visto poca producción literaria acerca de la fuga de clientes prepagados en las compañías del sector de los operadores móviles en Colombia, se encuentra una oportunidad pertinente para entender el contexto y aplicar técnicas estadísticas que apoyen el análisis de los clientes y generación de información valiosa que permitan la clasificación de esos clientes que presentan comportamientos similares a los que ya han dejado de usar el servicio.

# Capítulo 3

# Marco Teórico / conceptual

Con el objeto de clarificar y establecer un punto de partida en común para el desarrollo del trabajo, se presentan los siguientes conceptos clave para desarrollar y entender la siguientes secciones.

## 3.1 Telefonía móvil prepagada

Este servicio es ofrecido por compañías de telecomunicaciones, especializadas en telefonía móvil, las cuales tienen como núcleo del negocio ofrecer servicios principalmente de llamadas a otras líneas móviles o fijas, enviar y recibir mensajes de texto y megabytes para navegación en internet y redes sociales. El crecimiento de esta industria va de la mano con el desarrollo tecnológico, tanto de dispositivos móviles más novedosos y el desarrollo de plataformas, por lo tanto muy seguramente encontraremos en la cotidianidad de una persona la necesidad de tener a la mano si teléfono móvil para cumplir con sus tareas del día a día, y estar disponible para ser contactado en cualquier momento ya sea a través de una llamada, un mensaje de texto o de una plataforma como una red social.

## 3.2 Fuga de clientes - churn

Como lo hemos venido discutiendo podemos entender el churn como el fenómeno en el cual los usuarios o subscriptores del servicio de telefonía móvil y demás productos (entiéndase como planes, paquetes de recursos en sms, voz y datos), deciden terminar su relación con la compañía y acceder a otro tipo de ofertas que se presenten lanzadas por otras compañías que prestan esta misma clase de servicio. Para clasificar un cliente como churneado se tiene en cuenta que la línea tenga noventa días de inactividad consecutivos, esta inactividad se considera cuando el usuario no presenta uso del servicio, es decir, llamadas, mensajes de texto o consumo de datos.

## 3.3 Base de datos CDR

En telecomunicaciones, las bases de datos que almacenan los datos referentes a todo lo que tenga ver con la interacción del cliente con todos los servicios que estén a su disposición como usuario de telefonía móvil, es decir, llamadas que recibe y realiza, mensajes de texto enviados y recibidos, consumo de datos, recargas en dinero y compra de paquetes de recursos (o comúnmente llamados planes). Este tipo de compañías establece un esquema particular para establecer sus bases datos, pero en últimas resulta siendo un tema transparente en el sentido que solo nos preocuparemos por tener acceso a estos sistemas de información y que nos permita realizar las consultas que se necesiten; este esquema generalmente cuenta con dos sistemas de gestión de base de datos, el primero será donde se encuentran los CDR's que generalmente es de solo lectura y otro donde traeremos copias parciales o segmentos de estos datos de CDR para realizar las transformaciones que se requieran y generar las variables que serán ingresados al algoritmo de aprendizaje, al conjunto de variables organizadas y tabuladas se le llama dataset.

## 3.4 Algoritmos de aprendizaje de máquina

Conocidos popularmente como algoritmos de machine learning (ML), los cuales parten de un principio y es el de predecir hacia el futuro de acuerdo al conocimiento proporcionado por datos históricos, existen dos grandes de clases de aprendizaje, supervisado y no supervisado, para este trabajo nos fijaremos en los algoritmos supervisados, donde este tipo de aprendizaje supervisado se alimenta de datos históricos etiquetados con la clase a predecir, para efectos de esta investigación debemos suministrarle una serie de datos históricos de clientes que se fueron a churn y también datos de los que se conservaron en la compañía, esto con el objetivo de que la máquina encuentre patrones en cada grupo y con este conocimiento adquirido pueda predecir la etiqueta definida en datos que no se conozca a que clase pertenecen, es decir, predecir si de acuerdo a lo aprendido si un cliente se comporta como uno que ya se fue de la compañía o si se está comportando como un cliente de los que permanece.

### 3.4.1 Regresión logística

De forma similar a los métodos de regresión lineales usualmente usados, se tiene este tipo de técnica para que a través la suma de las variables ponderadas con ciertos coeficientes que se calculan se obtenga un resultado, el cuál es la probabilidad de que el resultado sea uno. Esto se logra gracias a la función logit implementada en modelos lineales generalizados para estructuras donde la respuesta es de tipo binomial o multinomial.

### 3.4.2 Redes neuronales

Conforma a la evolución de los sistemas informáticos, este tipo de modelos se han venido usando y explorando constantemente, donde los (ANS, Artificial Neural Systems) se logran dada la inspiración en los sistemas neuronales biológicos, que tienen dendritas (entradas), soma (procesador, funciones de activación) y el axón (salida), de esta forma los nodos de una red nuronal computacional se conectan e interactúan para recibir los datos con los que se va a entrenar y ajustando los pesos en los enlaces que determinan la configuración emergente para mejorar el resultado de la clasificación.

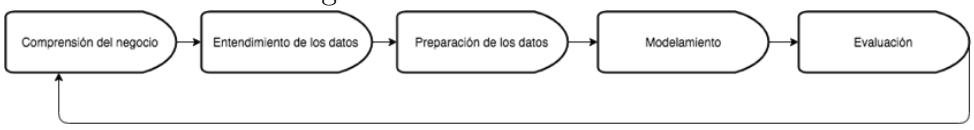
### 3.4.3 Máquinas de soporte vectorial SVM

Esta es una técnica que parte de una noción diferente a las abordada anteriormente en las dos técnicas señaladas, las cuales presenten representar los datos, específicamente, los datos asociados a cada valor que toma la variable respuesta. En las máquinas de soporte vectorial se busca establecer una línea, plano o hiperplano que separe al máximo las categorías ofrecidas por la variable respuesta.

## 3.5 Metodología CRISP-DM

Estas siglas en inglés abrevian Cross Industry Standard Process for Data Mining, metodología que surgió a finales de los años 90 gracias al gran interés de hallar conocimiento en los datos, proporciona una descripción normalizada del ciclo de vida de un proyecto de análisis de datos, este modelo CRISP-DM ayuda a establecer las fases de la metodología, sus tareas respectivas y las relaciones o dependencias entre tareas. De forma general esta metodología cuenta con 5 fases las cuales son: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelamiento, evaluación y despliegue; una de las características más relevantes en CRISP-DM es que es un proceso iterativo, en el sentido de que después de realizar la evaluación del modelo predictivo pueden emerger ajustes que conlleva a que se tengan que trabajar desde la fase del entendimiento del negocio, en el siguiente gráfico podremos ver de forma más fácil la metodología que se va a tener en cuenta.

Figura 3.1: Ciclo CRISP-DM




En primera medida se debe tener acceso a los datos, ya se había mencionado que a través de los CDR, se obtienen los datos de tráfico de los usuarios, el cuál es gestionado a través de un motor de base de datos, que permite realizar consultas sobre la información para extraer específicamente lo que se necesite, es relevante aclarar que la información

utilizada para este análisis es reservada y confidencial, en cumplimiento a políticas de habeas data exigidas desde la compañía prestadora del servicio.

Una de las razones para tomar de guía la metodología CRISP-DM, es por el hecho de que las variables que se van a usar se deben construir, extraer y transformar a partir de los datos que se encuentran en la base de datos, para esto se toma el trabajo de programar las consultas sql pertinentes para calcular adecuadamente la información necesaria correspondiente a la definición de variables que se realizó.

# Capítulo 4

# Marco Metodológico

En el ámbito de la metodología CRISP-DM, mencionado en el capítulo anterior, se establecen las diferentes etapas para el desarrollo de los objetivos propuestos en el presente trabajo.

### 4.0.1 Comprensión del negocio

Ya se tiene claro que el objetivo principal del modelo es la predicción de churn para los clientes de la compañía, por lo tanto, el criterio de éxito del modelo es que se logre determinar a partir de los datos existentes los patrones en común para reconocer un cliente que pueda estar en riesgo de desistir del servicio de telefonía móvil suministrado por la empresa, esto a favor de apoyar las estrategias de retención y fidelización de los usuarios. La principal fuente de datos con la que se cuenta son los CDR's, la cual debe estar disponible y accesible para su consulta de forma recurrente. Estos son los diferentes CDR's que podremos encontrar en una compañía de telecomunicaciones: CDR de tráfico y CDR de recargas.

### 4.0.2 Comprensión de los datos

Se revisa inicialmente la colección completa de datos, se seleccionan los campos que llegan a hacer útiles, se realiza una exploración de los campos seleccionados con el fin de determinar con qué tipo de datos estamos trabajando, es decir, hay diferentes tipos de datos como lo pueden ser valores numéricos (enteros o de punto flotante), datos booleanos, cadenas de caracteres, fechas o identificadores compuestos; habiendo definido esto aseguraremos que vamos a trabajos con datos íntegros y con un formato definido.

### 4.0.3 Preparación de los datos

Se realiza una caracterización de los datos existentes, para determinar la mejor manera de programar las consultas a la base de datos que haga las transformaciones que se requieran, y de esta manera se generen las variables con calidad e integridad, debemos verificar que, aunque estos procesos manipulan los datos no puede pasar que aparezcan

datos corruptos. Para esta selección de variables hemos mencionado que establecimos como referencia a (Alae, 2017)[6] y (Adnan,2017)[**adnan**], los cuales establecen un conjunto de variables fundamentales con las cuales se construirá el dataset, agrupadas semanalmente con el propósito de obtener las últimas 8 semanas, estas variables son las siguientes:

- Duración de llamadas salientes
- Duración de llamadas entrantes
- Productos vigentes
- Cantidad de recargas realizadas
- Valor acumulado en dinero de recargas
- Cantidad de productos comprados
- Cantidad de datos consumidos en megabytes
- Porcentaje de uso de la línea
- Saldo disponible

De acuerdo esto se conforma un conjunto de datos de 34362 registros que corresponden a 17000 usuarios que presentaron churn y 17362 que permanecieron activos, caracterizados a través de 104 variables que describen el comportamiento de estas líneas en las últimas ocho semanas de los clientes activos y también las últimas ocho semanas de los usuarios churneados antes de que iniciaran sus noventa días de inactividad.

### 4.0.4 Modelamiento

Para los algoritmos de aprendizaje de maquina supervisado se debe realizar una partición de los registros del dataset de la siguiente manera: los datos de clientes de los cuales se conoce que ya se fueron de la compañía, por lo tanto ya deben estar etiquetados como churn, estos datos son los que se pasan por la etapa de aprendizaje de máquina; un segundo grupo de datos los cuales corresponden al de los clientes activos en la empresa, los cuales son evaluados por el algoritmo y realiza la predicción de la etiqueta de churn [9]. En el presente trabajo se tiene como variable respuesta el churn y las demás variables mencionadas en la etapa anterior como las predictoras o independientes, esto con el objetivo de procesar de forma similar la información a través de los algoritmos, los cuales son: regresión logística, redes neuronales y máquinas de soporte vectorial.

### 4.0.5 Evaluación de resultados

Esta fase es fundamental para medir la calidad de las predicciones de nuestro algoritmo, dado que cuando este ha cumplido su parte de aprendizaje se vuelve a probar a si mismo con una pequeña parte de los registros con los que aprendió y que ya estaban previamente etiquetados como churn, con esta sencilla prueba podemos determinar qué tan acertadamente se están evaluando los clientes actuales. Para esto se definen los siguientes grupos, verdaderos positivos (TP) correspondiente a los casos que fueron correctamente evaluados por el clasificador, falsos positivos (FP) se refiere cuando el clasificador predice

como churn cuando en realidad no lo era, falsos negativos (FN) cuando un usuario que es churn no fue marcado con esta etiqueta y los verdaderos negativos, los cuales usuario que no tienen riesgo de abandono no se haya marcado como churn, estos indicadores conforman la matriz de confusión, la cual se usara para comparar el resultado de la clasificación de cada uno de las técnicas usadas.

### 4.0.6 Despliegue

En esta fase lo que establece es la puesta en marcha del modelo predictivo de forma recurrente e iterativa, que haga parte de los procesos en producción y reciba un monitoreo constante tanto a su parte de infraestructura como su parte lógica y de configuraciones. Se determina que el modelo hará una evaluación mensual de sus clientes los cuales serán analizados en una profundidad de tres meses (90 días). Se automatizan las tareas que diariamente consolidarán el día inmediatamente vencido para la construcción del dataset de clientes activos que serán evaluados, el resultado de la predicción se publicará en el sistema de información a manera de base de datos donde estará relacionado el identificador del usuario con su respectiva calificación de churn.

# Capítulo 5

# Análisis y Resultados

## 5.1 Análisis de componentes principales

Se realiza este análisis con el objetivo de conocer el comportamiento tanto de los clientes como de las variables en una misma representación, en la figura 5.1 se presenta el biplot del PCA para las variables cuantitativas respecto a los individuos que permanecen activos (color azul) y los que se fueron a churn (color rojo). Se notan dos puntos de concentración fuertes, uno para cada grupo, siendo la de los clientes fugados la que se encuentra en la parte negativa del primer componente principal, lo cuál indica la correlación inversa que tienen los clientes sobre la proyección en las variables por ejemplo el valor de recarga acumulado y la cantidad de recargas acumulada, que de alguna forma son las mas representativas sobre las componentes. Lo cuál significa que los individuos con cierto valor promedio de recargas realizadas periódicamente representan los clientes que van a permanecer activos, mientras que, los que tienen un comportamiento inverso, representan usuarios potencialmente a presentar fuga.

## 5.2 Análisis de correspondencias múltiples

Se continúa haciendo esta caracterización de usuarios ahora con el análisis MCA para incorporar variables categóricas, por ejemplo, si el cliente tuvo algún producto vigente en cierta semana. El análisis intenta hacer una representación a través de elipses, en la figura 5.2,se muestran los grupos de clientes que mantienen productos vigentes y los que no, observamos que en su gran mayoría de los clientes fugados, tienden a estar en el grupo de las variables que indican que no se tuvieron productos vigentes, los clientes activos tienen un espectro más amplio, aunque la mayor parte se concentra en las variables que representan que si se tienen productos vigentes, puede que un usuario activo no tenga productos vigentes todas las semanas, por tal razón vemos la elipse azul de clientes activos más amplia que la de clientes fugados.

Finalmente se evidencia que la proyección de las variables en dos componentes princi-

Figura 5.1: Gráfica PCA Biplot - Variables cuantitativas

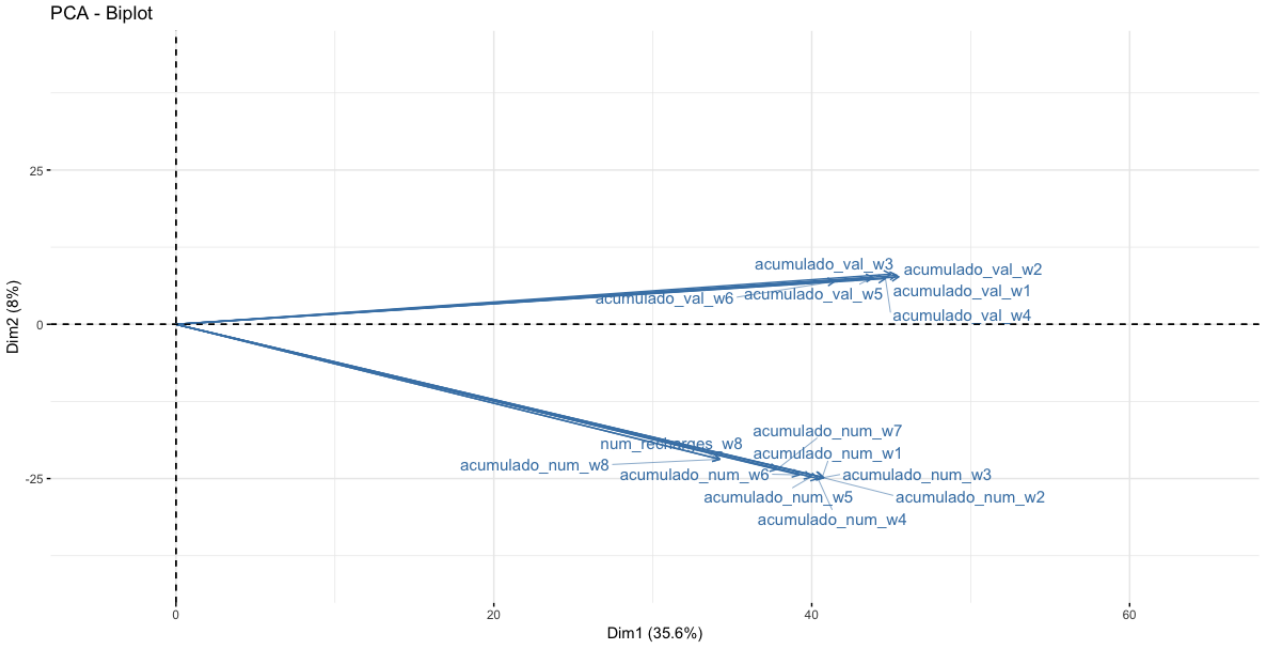
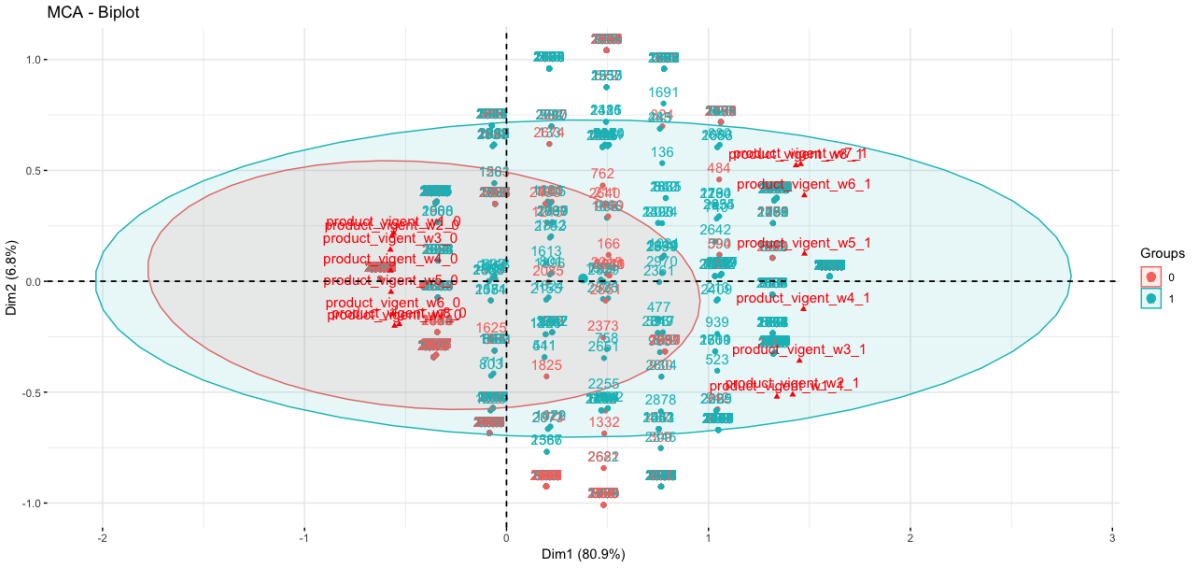



Figura 5.2: Gráfica MCA Biplot - Variables cualitativas




pales que explican casi el 88 % de la variabilidad de los datos, lo cuál confirma lo que se ha venido entendiendo con las variables de producto vigente, que determinan fuertemente el hecho que un usuario permanezca activo, es decir clientes que presenten irregularidad en las recargas y adquisición de productos, llegan a comportarse de forma similar a un cliente que va a abandonar el servicio.

## 5.3 Análisis de significancia de las variables

Un tema importante al momento de caracterizar un fenómeno o comportamiento a través de los datos es conocer cuáles son las variables significativas en las que se debe apoyar el análisis y construcción de modelos. Gracias a los coeficientes generados a través de la regresión logística se optimiza un modelo con las variables más significativas a través del estadístico de T-Student, con los cuales se calculan los odds ratio, que sencillamente representan la posibilidad de ocurrencia de un evento en este caso el hecho que un usuario permanezca activo frente a la aparición o magnitud cada variable.

| | | | |
|---|---|---|---|
| product_vigent_w8 | 2.2048007 | originating_w8 | 1.0036344 |
| product_vigent_w3 | 1.351812 | originating_w4 | 1.0022137 |
| product_vigent_w4 | 1.2904175 | terminating_w4 | 1.001022 |
| product_vigent_w5 | 1.2273286 | data_mb_w1 | 1.0007584 |
| product_vigent_w7 | 1.2237291 | data_mb_w8 | 1.0004986 |
| purchases_w1 | 1.1267449 | data_mb_w3 | 1.000411 |
| acumulado_num_w1 | 1.0743427 | data_mb_w2 | 1.0002852 |
| forwarding_w1 | 1.0478729 | data_mb_w5 | 1.0002472 |
| sms_w5 | 1.0187536 | total_money_w1 | 1.0000981 |
| forwarding_w8 | 1.0169762 | total_money_w8 | 1.0000699 |
| use_p_w1 | 1.0135304 | val_recharges_w6 | 1.0000453 |
| sms_w3 | 1.0120658 | val_recharges_w5 | 1.0000393 |
| use_p_w7 | 1.0058987 | total_money_w2 | 1.0000393 |
| terminating_w2 | 1.0042897 | total_money_w4 | 1.0000285 |
| terminating_w8 | 1.0039614 | total_money_w3 | 1.0000195 |
| originating_w2 | 1.0037475 | val_recharges_w7 | 1.00001 |
| originating_w6 | 1.0037366 | | |

Cuadro 5.1: Tabla de odds ratio

Al revisar el cuadro 5.1 observamos las variables de producto vigente en primer lugar, especialmente la que representa el tenencia del producto en la semana octava, lo cual indica que si un cliente tiene producto vigente ahora y tuvo también producto en la semana octava, tiene dos veces mayor de probabilidad de seguir activo que un cliente que no presente adquisición de productos. Adicionalmente se encuentran variables interesantes como lo son el número de compras, la cantidad de tiempo en llamadas, la cantidad de datos consumidos y el porcentaje de uso de la linea semanalmente.

Luego de la ejecución de los algoritmos se obtienen los siguientes resultados, los cuales se presentan a través de una matriz de confusión, la cuál muestra el resultado de la clasificación en cada técnica, donde las filas representan la etiqueta verdadera y las columnas la etiqueta que ha resultado de la predicción, de esta forma se puede obtener la precisión predictiva del algoritmo calculando la proporción de predicciones acertadas sobre el total

de observaciones:

## 5.4 Regresión Logística

En estadística, la regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores. El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados (GLM por sus siglas en inglés) que usa como función de enlace la función logit. Las probabilidades que describen el posible resultado de un único ensayo se modelan, como una función de variables explicativas, utilizando una función logística.

| | | Predice 0 | Predice 1 | Total | Precisión clase |
|---|---|---|---|---|---|
| Actual | 0 | 14362 | 1987 | 16349 | 87.85 % |
| Actual | 1 | 2340 | 13671 | 16011 | 85.39 % |
| | | | | Precisión | 86.63 % |

Cuadro 5.2: Matriz de confusión regresión logística - in sample.

| | | Predice 0 | Predice 1 | Total | Precisión clase |
|---|---|---|---|---|---|
| Actual | 0 | 879 | 134 | 1013 | 86.77 % |
| Actual | 1 | 156 | 833 | 989 | 84.23 % |
| | | | | Precisión | 85.51 % |

Cuadro 5.3: Matriz de confusión regresión logística - out sample.

## 5.5 Redes neuronales

La redes neuronales son un modelo computacional que representan de forma análoga el comportamiento conocido al de una red neuronal biológica, la cuál sometida varias acciones logra producir diversos valores de salida. Las neuronas están conectadas a través de enlaces, en estos enlaces el valor de salida de la neuronal predecesora se multiplica por un valor de ponderación, estos pesos pueden amplificar o reducir la activación de las demás neuronas vecinas. De la misma forma, la salida de cada neurona posee una función de activación la cuál modifica el valor resultado o establece un umbral límite que se debe cumplir para que la magnitud se propague a los demás nodos.

Este tipo de modelos se forman así mismos y aprenden, muy diferente a tener reglas elaboradas o preestablecidas que se pueden programar, las redes neuronales van un poco más allá donde estás reglas de asociación son difíciles de establecer por la programación convencional. Para incrementar el aprendizaje de la red se hace a través de la función

de perdida, la cuál califica que tan lejos se esta del objetivo, y se tiene en cuenta para reajustar los pesos de los enlaces.

| | | Predice 0 | Predice 1 | Total | Precisión clase |
|---|---|---|---|---|---|
| Actual | 0 | 11830 | 3319 | 15149 | 78.09 % |
| Actual | 1 | 698 | 14153 | 14851 | 95.30 % |
| | | | | Precisión | 86.61 % |

Cuadro 5.4: Matriz de confusión redes neuronales - in sample.

| | | Predice 0 | Predice 1 | Total | Precisión clase |
|---|---|---|---|---|---|
| Actual | 0 | 772 | 222 | 994 | 77.67 % |
| Actual | 1 | 58 | 950 | 1008 | 94.25 % |
| | | | | Precisión | 86.01 % |

Cuadro 5.5: Matriz de confusión redes neuronales - out sample.

## 5.6 Máquinas de soporte vectorial SVM

La SVM busca un hiperplano que separe de forma óptima a los puntos de una clase de la de otra, los cuales se han representado en un espacio en cuál cada punto será un individuo observado simultáneamente con todas las variables que lo representan.

En la idea de separación óptima es donde reside la característica más importante de las máquinas de soporte vectorial: este tipo de algoritmos buscan el hiperplano que tenga la máxima distancia con los puntos que estén más cerca de él mismo. Por eso también a veces se les conoce a las SVM como clasificadores de margen máximo. De esta forma, los puntos del vector que son etiquetados con una categoría estarán a un lado del plano y los casos que se encuentren en la otra categoría estarán al otro lado.

| | | Predice 0 | Predice 1 | Total | Precisión clase |
|---|---|---|---|---|---|
| Actual | 0 | 14642 | 1304 | 15946 | 91.82 % |
| Actual | 1 | 1707 | 14707 | 16414 | 89.60 % |
| | | | | Precisión | 90.70 % |

Cuadro 5.6: Matriz de confusión SVM - in sample.

Aplicando este clasificador para la variable de fuga se busca optimizar un hiperplano que logre separar al máximo los puntos de ambas categorías, de esta forma un dato desconocido se ubicará en el respectivo grupo que corresponda, clasificándolo como activo o churneado, tal cuál se ve en el cuadro 5.6.

| | | Predice 0 | Predice 1 | Total | Precisión clase |
|---|---|---|---|---|---|
| Actual | 0 | 862 | 92 | 954 | 90.36 % |
| Actual | 1 | 151 | 897 | 1048 | 85.59 % |
| | | | | Precisión | 87.86 % |

Cuadro 5.7: Matriz de confusión SVM - out sample.

# Capítulo 6

# Conclusiones y Recomendaciones

Del anterior capítulo en el cuál se presentaron los análisis correspondientes al desarrollo del trabajo, se establecen las siguientes conclusiones, dando lugar a resolver lo que se tenía planteado en las conclusiones:

- De acuerdo al análisis PCA y MCA, los clientes propensos de fuga se empiezan a diferenciar de los clientes activos cuando el valor acumulado de recargas es bajo y por ende también su cantidad de recargas realizadas, además, al tener saldo puede adquirir y mantener productos vigentes lo cuál promueve el uso del servicio y evitar que el cliente tenga periodos de inactividad.

- En el cálculo de los odds ratio ofrecidos por los coeficientes generados en la regresión logística, se evidencia que las variables producto vigente entre las semana 3 hasta la 8, indican que el cliente que adquiera un producto en estas semanas tiene menos probabilidad de fugarse que uno que adquiera productos.

- Al probar los tres métodos de clasificación, obtenemos que SVM ofrece una tasa de error mas baja, con una precisión fuera de la muestra del 12.14 %, escogiendo éste como el clasificador adecuado entre los que se evaluaron.

- La etapa para hacer el despliegue del modelo y las acciones que se deban tomar para hacer el tratamiento de los clientes que puedan llegar a presentar indicios de presentar fuga, se dejan a criterio y responsabilidad de la compañía operadora de telefonía móvil, responsable de los datos suministrados.

- El trabajo continúa en el sentido de seguir incorporando otro tipo de variables que no consideraron para el presente trabajo e invitar a la comunidad profesional y académica a seguir explorando otros tipos de técnicas que nos permitan resolver problemas de clasificación.

# Bibliografía


[1] VILLAMARIN Vega J. *ANÁLISIS DE LOS DETERMINANTES QUE FIDELIZAN A LOS CLIENTES DE TELEFONÍA MOVIL EN LA CIUDAD DE BOGOTÁ: CASO CLARO*. Citado en Noviembre 17 de 2018. URL: https://repository.usta.edu.co/bitstream/handle/11634/10442/Villamar%C3%ADnjavier2018.pdf?sequence=1&isAllowed=y.

[2] ALVARADO Bustos J. *DISEÑO E IMPLEMENTACIÓN DE UN MODELO PREDICTIVO PARA DETECTAR PATRONES DE FUGA EN LOS SERVICIOS DE TELEFÓNICA DEL SUR*. Citado en Noviembre 17 de 2018. URL: http://cybertesis.uach.cl/tesis/uach/2011/bmfcia472d/doc/bmfcia472d.pdf.

[3] CONTRERAS Morales E. y FERREIRA Correa F. *DISEÑO DE UN MODELO PREDICTIVO DE FUGA DE CLIENTES UTILIZANDO ÁRBOLES DE DECISIÓN*. Citado en Noviembre 17 de 2018. URL: http://revistas.ubiobio.cl/index.php/RI/article/view/3055/3075.

[4] BART Baesens y GERDA Claeskens AIMEE Backiel1. *Predicting Time-To-Churn of Prepaid Mobile Telephone Customers Using Social Network Analysis*. Citado en Noviembre 17 de 2018. URL: https://feb.kuleuven.be/public/u0043181/papers/BackielBaesensClaeskens_PredictingChurn_JORS_2016.pdf.

[5] ZEHRA Can y ENRIç Albey. *Churn Prediction for Mobile Prepaid Subscribers*. Citado en Noviembre 17 de 2018. URL: http://www.scitepress.org/Papers/2017/64253/64253.pdf.

[6] ALAE Chouiekh y EL HASSANE Ibn El Haj. *Machine learning techniques applied to prepaid subscribers: Case study on the telecom industry of Morocco*. Citado en Noviembre 17 de 2018. URL: https://ieeexplore.ieee.org/document/8054923.