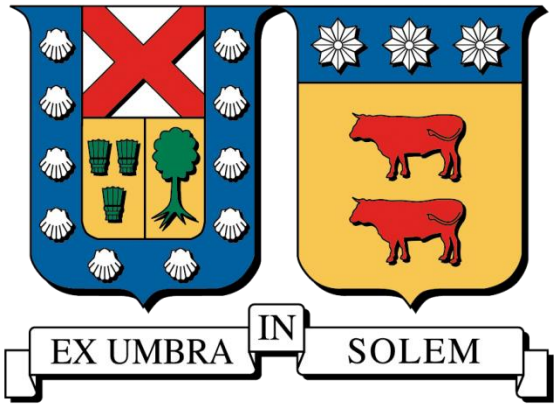
2018-12

PREDICCIÓN DE FUGA DE CLIENTES EN APLICACIONES MÓVILES

PUMARINO PALOMBO, CONSTANZA VALENTINA

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INGENIERÍA COMERCIAL
SANTIAGO-CHILE



PREDICCIÓN DE FUGA DE CLIENTES EN APLICACIONES MÓVILES

CONSTANZA VALENTINA PUMARINO PALOMBO

MEMORIA PARA OPTAR AL TÍTULO DE:
INGENIERO COMERCIAL

PROFESOR GUÍA: SR. LUIS ACOSTA
PROFESOR CORREFERENTE: SR. DAVID ALMENDRAS

DICIEMBRE 2018

Agradecimientos

Quiero agradecer a mi familia, especialmente a mi madre Isabel Palombo, por el apoyo incondicional en este proceso, por siempre creer en mí e incentivar un espíritu de superación y crecimiento personal en mi vida.

Agradezco también a mis compañeros, Matías Campos, Alonso Canales y Felipe Finkelstein, quienes no solo fueron parte del camino, sino que verdaderos amigos que enriquecen mi vida. Sin las eternas tardes de estudio, el intento de hacer ejercicio para una vida sana y equilibrada y los momentos de esparcimiento no estaría hoy terminando la carrera.

Es un orgullo poder llamarlos amigos y mantener la cercanía que desde un comienzo nos unió.

Índice

Índice de Ilustraciones

Índice de Gráficos

Índice de Tablas

1 Resumen Ejecutivo

La presente memoria busca determinar un modelo de predicción de fuga y de segmentación de clientes basados en información histórica de usuarios de una aplicación móvil mediante técnicas de minería de datos utilizando la metodología KDD.

Esta aplicación, perteneciente al rubro del entretenimiento, entrega un medio de comunicación entre empresas de variados rubros y los consumidores finales, permitiendo hacer campañas con promociones y beneficios comerciales las cuales poseen un club de fidelización que premia a los usuarios por medio de puntos acumulables canjeables por productos.

El experimento consistió en la recolección de 3 meses de datos de nuevos usuarios registrados, para analizar el estado de los usuarios al 4to mes, clasificándolo como fugado si este no interactuaba con la aplicación en 30 días, obteniendo un nivel de fuga de clientes de un 44,25%. Se complementó con información relacionada al comportamiento de uso de la aplicación, como cantidad de mensajes recibidos y enviados, cantidad de puntos acumulados y canjeados. También se utilizó información demográfica, características del dispositivo móvil y variables temporales de tiempos de uso. Por otra parte, los datos se sometieron a un proceso de clasificación de clusters, analizando los tipos de usuarios y las características de cada grupo, obteniendo 4 tipos de clientes, segmentados por el nivel de uso de la aplicación, identificando que los primeros 10 días de uso son críticos para la retención.

Para la predicción de fuga se compararon 7 modelos distintos, siendo Gradient Boosted Tree el con mejor rendimiento, alcanzando una exactitud de 91,1% pero con un tiempo de ejecución muy elevado en comparación al resto, seguido por Deep Learning que alcanzando un 89,2% 27 veces más rápido que el primer modelo.

Se concluye el estudio con un análisis de los resultados, identificando los atributos claves para la retención de clientes y proponiendo posibles cursos de acción para el manejo de clientes a futuro.

2 Introducción

En la actualidad, los teléfonos celulares son una necesidad, la versatilidad es tal, que una persona puede manejar todos los aspectos de su vida desde su mano. Según la publicación "*The Future of Internet III*" (01), en el año 2020 se espera que los móviles sean el principal medio de acceso a internet.

Esta tendencia, genera oportunidades de negocio, siendo las aplicaciones móviles una de esas, y con un mercado en expansión, hay mucho trabajo todavía por hacer.

Otra tendencia mundial es el análisis de datos, que permite inferir el comportamiento de los clientes basándose en sus acciones pasadas. Según lo expuesto en la encuesta *Big Data Survey 2015: 4 Core Insights for More Success With Data* (02), de las 186 empresas encuestadas, el 17% ya usa *Business Intelligence* en su organización, y para el 2018 se esperaba que sea el 40%.

En general las aplicaciones móviles poseen una baja retención de clientes, en la industria, se tiene una tasa de deserción del 71% en un periodo de 90 días, definiendo como usuario activo aquel que utiliza la aplicación por lo menos una vez cada 30 días a partir desde que un usuario instala una aplicación en su celular (03).

Otro indicador que arroja el estudio de Localytics, es que 21% de los usuarios solo usa 1 vez las aplicaciones móviles, dándole una mayor importancia a la interacción del usuario por sobre la cantidad de descargas de la app. El análisis de la información que se recopila día a día de los

usuarios y su comportamiento es la clave para perdurar en una industria muy competitiva y en constante movimiento.

Adicionalmente, Facebook, uno de los principales actores de marketing, recientemente liberó un servicio que permite optimizar la retención de clientes para las aplicaciones móviles a través del gestor de campañas (04).

3 Problema de Investigación

La fuga de clientes conlleva varios problemas para las empresas, un ejemplo es el costo asociado a la captación de nuevos clientes, que es 5 veces más costoso en comparación con el costo de una campaña de retención de uno existente, y el aumento de un 5% en la retención de clientes, puede aumentar entre un 25% y un 95% el retorno de la empresa (05).

Mediante la presente memoria, se estudiará la estadística descriptiva del perfil de fugador en una aplicación móvil mediante estadística bivariable y modelos no supervisados de clustering, así como encontrar un modelo predictivo de fuga que permitan identificar un perfil de usuarios para una aplicación, y recomendaciones de negocio a nivel estratégico y acciones que permitan aumentar la retención de usuarios.

El estudio se realizará con una aplicación móvil del rubro de entretención con una influencia importante sobre usuarios jóvenes. La aplicación envía contenido con panoramas, beneficios comerciales y descuentos. Adicionalmente la aplicación posee convenios con clubes de beneficios

en el área de retail y permite la acumulación de Puntos que pueden ser canjeados por diversos productos o servicios.

Creando un modelo de predicción de fuga, se pueden implementar campañas antes de perder el cliente, disminuyendo costos operacionales, como personal de captación de clientes y de marketing asociado.

4 Objetivos

4.1 Objetivo General

Determinar un modelo predictivo de fuga para aplicaciones móviles, usando técnicas de minería de datos para disminuir la tasa de deserción y proponer un plan de retención.

4.2 Objetivos Específicos

Estudiar estadística descriptiva mediante técnicas bivariadas y modelo no supervisados.

Identificar patrones que expliquen el perfil del fugador.

Construir el *dataset* de entrenamiento, validación y testeo.

Construir un modelo de fuga.

Validar el modelo.

Proponer posibles cursos de acción para reducir la fuga.

5 Marco Teórico

5.1 ¿Qué es una aplicación móvil?

Una aplicación móvil o app en inglés, es un programa informático diseñado para ser ejecutado en un teléfono inteligente (*smartphone*), tabletas y otros dispositivos móviles, que permiten al usuario efectuar una tarea concreta de cualquier tipo.

5.2 Clasificación de las aplicaciones móviles

Existen 3 tipos de apps (06), que se clasifican como:

Aplicaciones Nativas: Son aquellas que se desarrollan para un sistema operativo específico (iOS, Android, etc.). Algunas ventajas de estas aplicaciones, es que están totalmente integradas con el dispositivo para el cual se crearon, pudiendo hacer uso total de las distintas características como el GPS, 3D Touch, cámara, acelerómetro, entre otros. No requieren conexión web para ser ejecutadas y están más expuestas a público al ser distribuidas en los distintos *Stores* de cada sistema operativo.

Aplicaciones Web: Este tipo de aplicaciones, a diferencia de las nativas, requiere de una conexión web para ser ejecutadas, lo que permite ser usadas en cualquier dispositivo que tenga conexión a internet.

Aplicaciones Híbridas: Como su nombre lo dice, es una mezcla entre los 2 tipos anteriores, usa tecnología multiplataforma de las Aplicaciones Web pero permiten acceder a buena parte de los dispositivos y sensores del teléfono mediante comunicadores como “PhoneGap”. Un ejemplo de

aplicación híbrida es Facebook, se descarga del *Store*, cuenta con las características de una aplicación nativa, pero requiere constantes actualizaciones.

5.3 Evolución de las aplicaciones móviles

Las primeras aplicaciones móviles fueron creadas en los años ’90. Eran los calendarios, juegos de arcade, editores de *ringtones* e incluso e-mail ya finalizando la década. Cumplían funciones elementales y su diseño era bastante simple (07).

La llegada del Protocolo de Aplicaciones Inalámbricas (*WAP* por sus siglas en inglés, detalles del protocolo en el Anexo N°1), la tecnología *EDGE* para el uso de internet y la creación de los *smartphones*, fueron los principales impulsores del mercado de las *apps* y su evolución a lo actual, que encontramos aplicaciones para todo tipo de intereses, desde juegos, diseños, arte, medicina y noticias entre otros.

El lanzamiento del iPhone en el año 2007, abrió las puertas a los desarrolladores para crear y distribuir sus aplicaciones a través del App Store, lo que permitió la evolución de un mercado estancado, ya que antes sólo los fabricantes de sistemas operativos podían desarrollarlas (08).

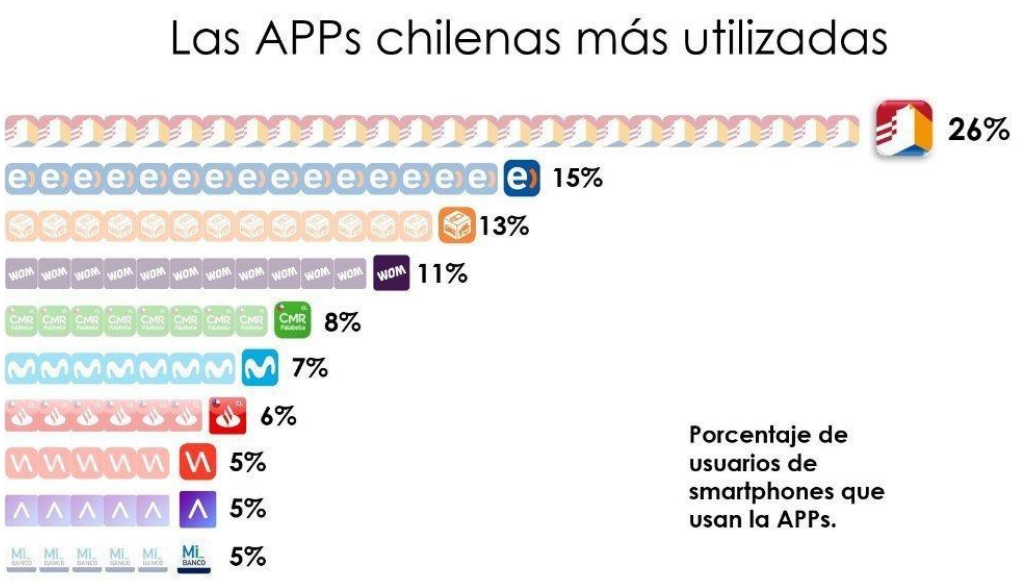
5.4 Mercado de las aplicaciones móviles en Chile

No es una sorpresa que en Chile el mercado de las aplicaciones móviles esté en expansión. La masificación de la tecnología, la llegada del 4G y la mayor accesibilidad para desarrollar, han

generado un espacio para desarrollo. Al 2015, se han creado más de 3 mil aplicaciones en Chile y a nivel mundial se crean 30 mil todos los meses (09).

En Chile, según un estudio realizado por Nielsen, indicó que 93% de los celulares en Chile son smartphones (10). Otro estudio de la misma compañía indica que en promedio, un usuario de smartphone, utiliza 27 aplicaciones al mes, un claro indicador de que hay que mejorar la oferta disponible y entregar un servicio que genere valor. (11).

Tabla 1: Apps chilenas más utilizadas



Fuente: Criteria Reasearch 2018

De las 10 aplicaciones más usadas en Chile, 9 están asociadas a servicios que su *core* no es la tecnología, principalmente bancos (Banco Estado, Falabella, Santander, Mach (BCI) y Banco de Chile) y telecomunicaciones (Entel, WOM y Movistar), lo que indica que el uso de aplicaciones

es principalmente un medio de comunicación con empresas establecidas por sobre aplicaciones desarrolladas exclusivamente para smartphones.

No es extraño que la aplicación más utilizada sea del Banco Estado, ya que todo ciudadano tiene una cuenta bancaria gratuita, permitiendo una exposición mayor y un mercado objetivo más extenso que las otras empresas.

5.5 Modelos de fuga

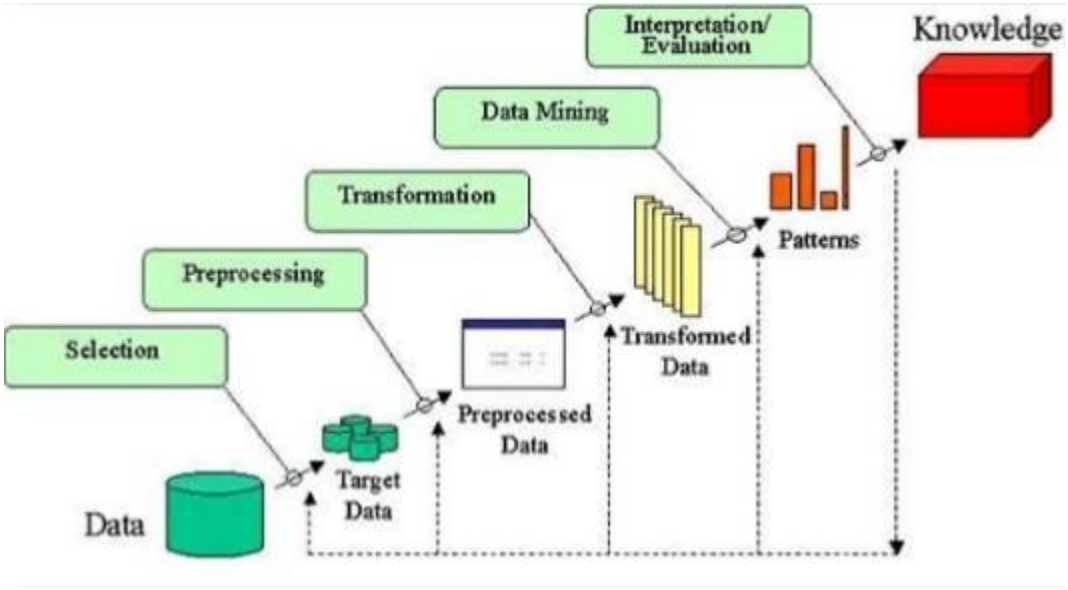
Los modelos de fuga o *churn* en inglés, buscan predecir el abandono de un cliente. Utilizando datos recolectados a través de la experiencia de los usuarios, se trata de modelar los patrones de comportamientos que puedan indicar si un nuevo usuario es un potencial "fugado" o "*churner*". Utilizando técnicas de clasificación y regresión se crea un modelo predictivo, identificando los atributos que agregan o no valor a la aplicación móvil y tomar medidas acordes a la experiencia del usuario para lograr una retención mayor en el tiempo.

Hay distintos tipos de fuga, pero el relevante para este estudio es el absoluto, que es el caso en que el usuario deja de usar la aplicación en un determinado periodo de tiempo.

5.6 Knowledge Discovery in Databases

El proceso denominado Knowledge Discovery in Databases (KDD), es el proceso iterativo de análisis de bases de datos y consta de 5 etapas como explica (12):

Ilustración 1:Proceso KDD



Fuente: www2.cs.uregina.ca

a. Integración o Selección: es la definición de las variables objetivo (lo que se quiere predecir, calcular o inferir) las variables independientes (las que se usarán para el cálculo o proceso) y el muestreo de los registros disponibles.
b. Preprocesamiento: es el análisis de las propiedades de los datos, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos.
c. Transformación: preparar la información para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema.
d. Minería de Datos: es la extracción de conocimiento, se obtiene un modelo que representa patrones de comportamiento observado en los valores de las variables del problema o relaciones de asociación entre dichas variables. Pueden usarse varias técnicas a la vez para generar distintos modelos, pero cada técnica requerirá un preprocesamiento diferente de los datos.

e. Interpretación y Evaluación: una vez obtenido el modelo se debe validar, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias (para todos los modelos creados en caso de tener múltiples). Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

5.7 Modelos Predictivos

Existen 2 tipos de modelos predictivos, los de aprendizaje supervisados o minería predictiva y los no supervisados o minería descriptiva.

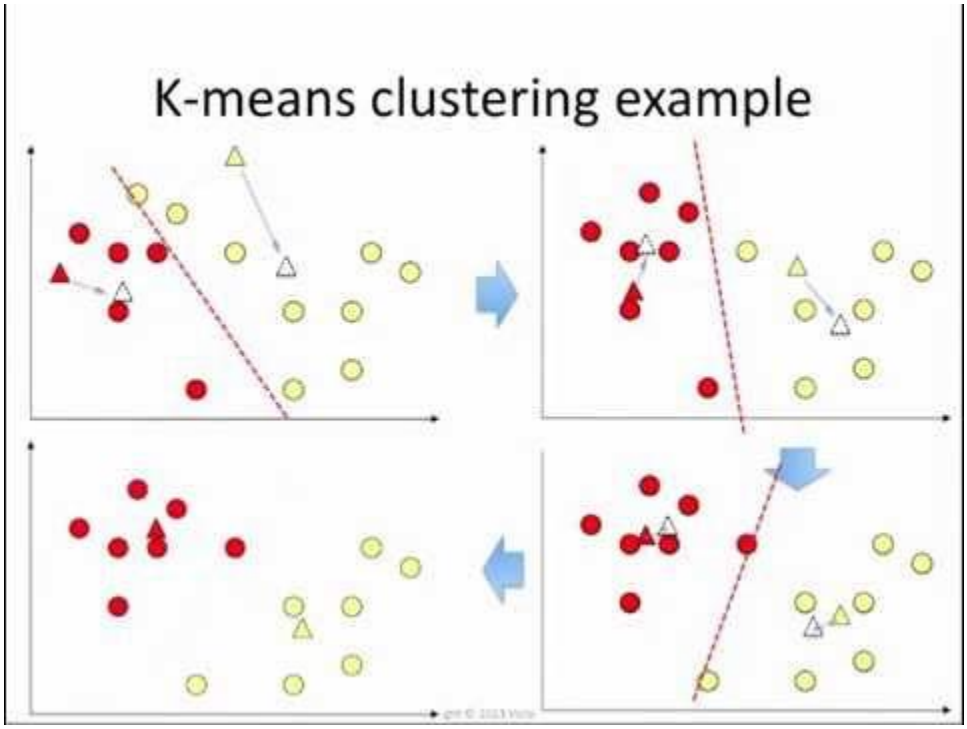
5.7.1 Modelos No Supervisados

Estos modelos buscan agrupar los datos según similitud, en base a las propiedades de estos, sin ninguna definición previa, segmentación o clasificación de estos.

5.7.1.1 Modelo K-Means

Dentro de estos modelos, encontramos el modelamiento de clusters, que a través del proceso de K-medias, que busca minimizar la varianza del sistema, definiendo centroides y agrupando los datos según cercanía a cada punto. El número y posición de centroides se encuentra con un proceso iterativo para asegurar la mínima varianza.

Ilustración 2: Modelo K-Medias



Fuente: http://www.inf.ed.ac.uk/teaching/courses/iaml/slides/kmeans-2x2.pdf

Una variación del modelo de K-medias es X-medias (para el caso de software Rapidminer), que calcula el número de centroides óptimos para la clasificación en vez de entregar un parámetro fijo para su segmentación.

5.7.2 Modelos Supervisados

Este tipo de modelos busca clasificar los datos basándose en datos procesados anteriormente. Estos modelos tienen 2 etapas, una de entrenamiento y otra de validación, o sea, del universo de datos disponibles, se utiliza una parte para crear y entrenar el modelo y la otra parte para validarlo (medir

el porcentaje de éxito en la clasificación) y asegurar un nivel de cumplimiento mínimo. Dentro de estos modelos, se detallan siete tipos:

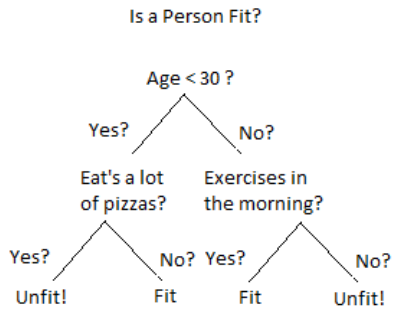
5.7.2.1 Árboles de Decisión

Se forman mediante un conjunto de nodos de decisión (ramas) y nodos respuesta (hojas).
Los nodos de decisión están asociados a los atributos y tienen 2 o más ramas que salen del, cada una de ellas representando los posibles valores que puede tomar el atributo asociado. De alguna forma, un nodo de decisión es como una pregunta que se le hace al ejemplo analizado, y dependiendo de la respuesta que dé, el flujo tomará una de las ramas salientes.

Un nodo respuesta está asociado a la clasificación que se quiere proporcionar, y nos devuelve la decisión del árbol con respecto al ejemplo de entrada.

Un ejemplo de árbol de decisión es la siguiente ilustración, dónde en base a las respuestas de una persona, la clasifican como En Forma o No En Forma.

Ilustración 3: Ejemplo de Árbol de Decisión

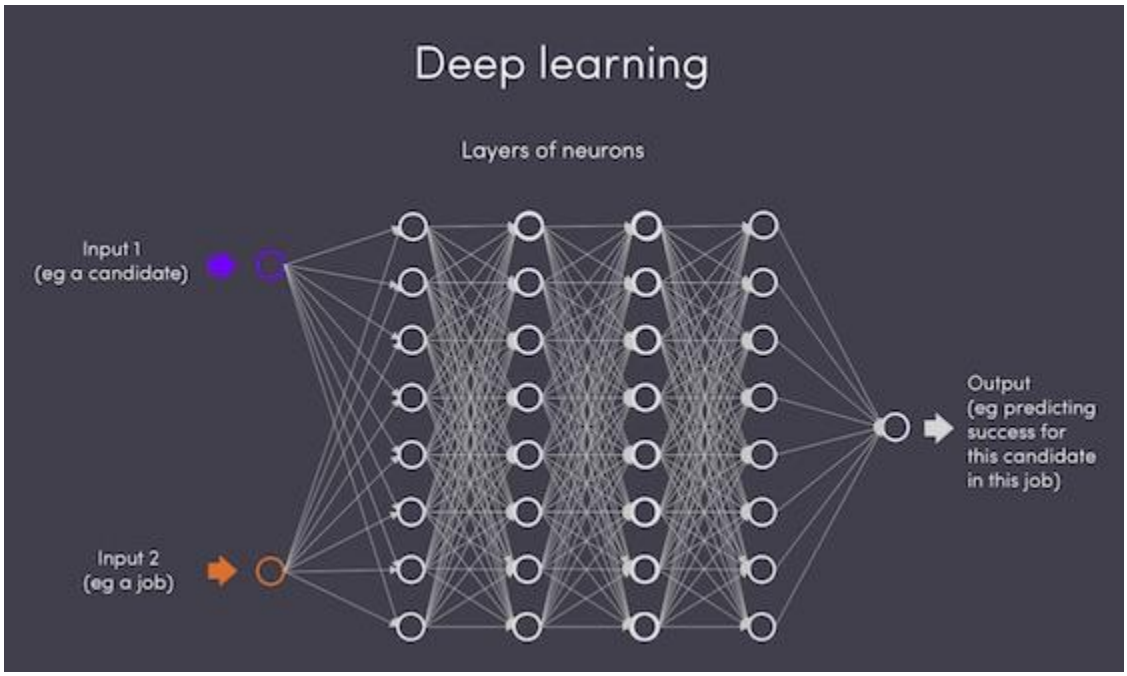


Fuente: https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html

5.7.2.2 Deep Learning

El Deep Learning es un tipo de algoritmos de aprendizaje automático estructurado o jerárquico, dicho de otra forma, toma modelos existentes para predecir el futuro con los datos disponibles. El proceso de predicción se realiza mediante el aprendizaje, no con reglas programadas previamente y casi siempre ligado al procesamiento de texto, voz, imagen y vídeo. No dan siempre la misma respuesta. A partir de la interacción con el usuario, "aprenden" si la respuesta que han dado es la adecuada o si deben ofrecer otra alternativa (13).

Ilustración 4: Ejemplo Deep Learning



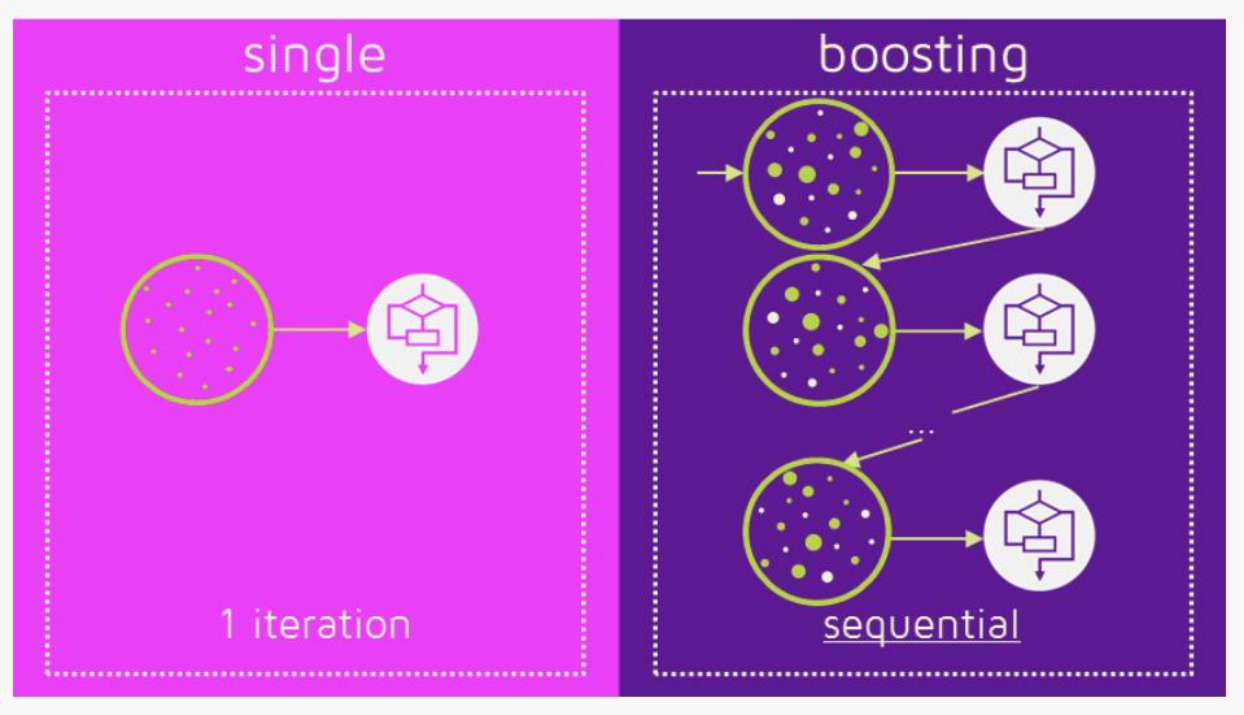
Fuente: https://www.untapt.com/industry/2017/04/29/how-ai-impacts-hr/

5.7.2.3 Gradient Boosted Tree

Árboles de Decisión Potenciados en español, es una técnica de aprendizaje automático utilizado para el análisis de la regresión y para problemas de clasificación estadística, el cual produce un

modelo predictivo en forma de un conjunto de modelos de predicción débiles, típicamente árboles de decisión. Construye el modelo de forma escalonada como lo hacen otros métodos de potenciamiento, y los generaliza permitiendo la optimización arbitraria de una función de pérdida diferenciable.

Ilustración 5: Ejemplo de Árbol de Decisión Potenciado



Fuente: https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d

5.7.2.4 Naive Bayes

Clasificador Bayesiano Ingenuo es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. En términos simples, asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica, dada la clase variable y considera que cada una de estas

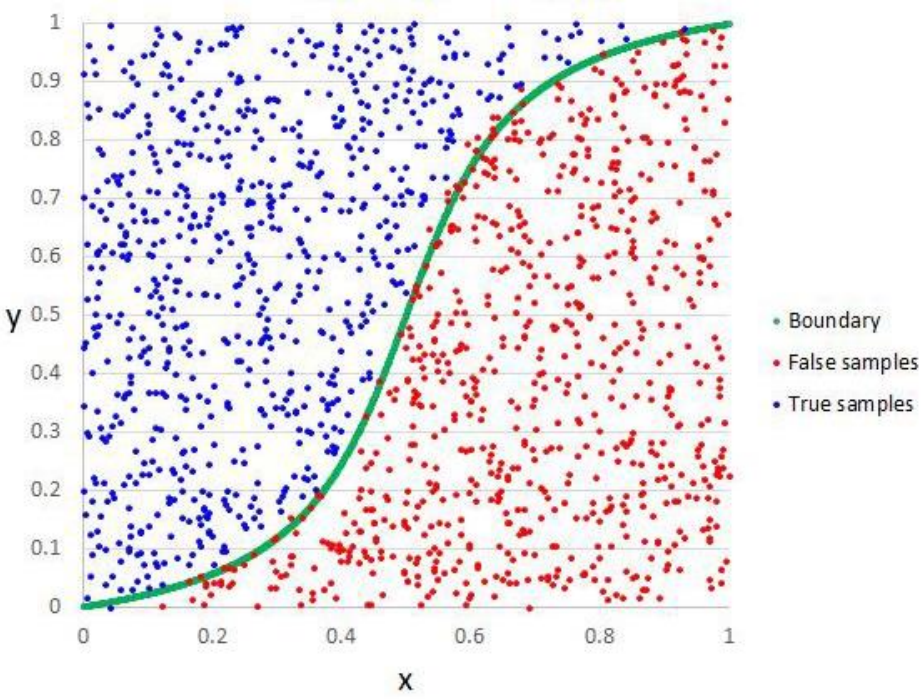
características contribuye de manera independiente a la clasificación, independientemente de la presencia o ausencia de las otras características (14).

5.7.2.5 Logistic Regression

Es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica o dependiente en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores.

La regresión logística es usada extensamente en las ciencias médicas y sociales. Otros nombres para regresión logística usados en varias áreas de aplicación incluyen modelo logístico, modelo logit, y clasificador de máxima entropía.

Ilustración 6: Ejemplo de Regresión Logística

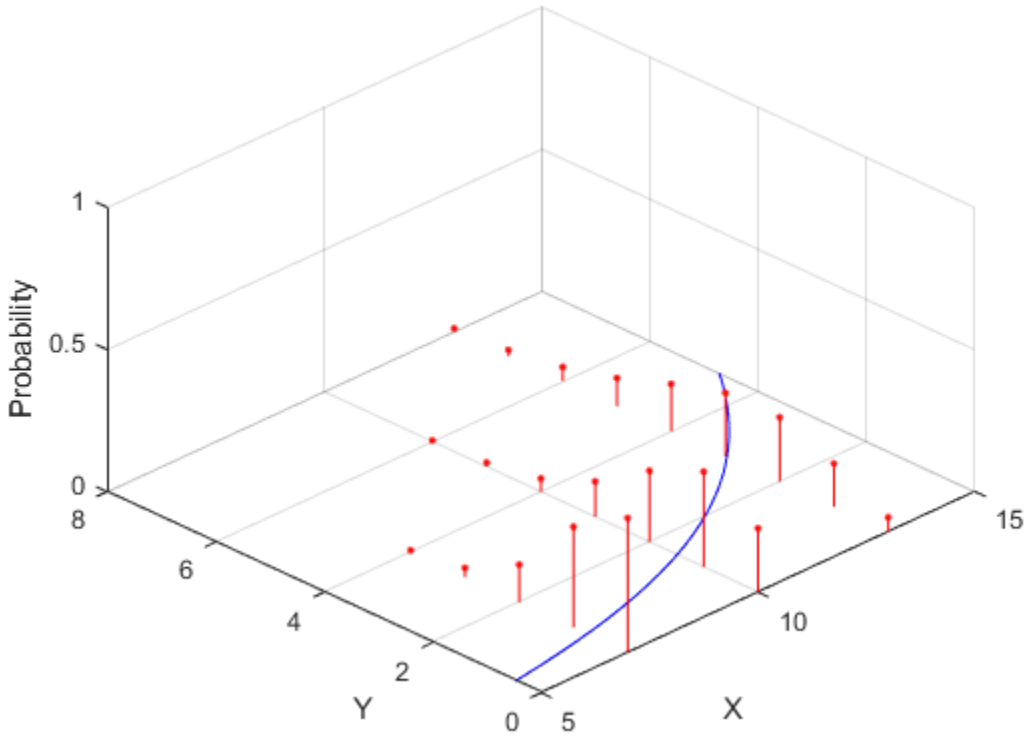


Fuente: https://medium.com/greyatom/logistic-regression-89e496433063

5.7.2.6 Generalized Linear Model

Los modelos lineales generalizados fueron formulados por John Nelder y Robert Wedderburn como una manera de unificar varios modelos estadísticos, incluyendo la regresión lineal, regresión logística y regresión de Poisson, bajo un solo marco teórico. Esto les permitió desarrollar un algoritmo general para la estimación de máxima verosimilitud en todos estos modelos. Esto puede extenderse de manera natural a otros muchos modelos. (15)

Ilustración 7: Ejemplo de Modelo Lineal Generalizado



Fuente: https://www.mathworks.com/help/stats/examples/fitting-data-with-generalized-linear-models.html

5.7.2.7 Random Tree

Es un clasificador de conjunto, los que se construyen juntando varios métodos clasificatorios y combinándolos de forma ponderada o no.

Los Árboles Aleatorios usan la misma lógica de un Árbol de Decisión, se hacen varias ramas de menor peso para crear un conjunto más potente, en donde se escoge de forma aleatoria el atributo y el corte de la rama que lo iniciará (16).

5.8 Eficiencia de un Aprendizaje

Lo que buscamos con los modelos creados, es poder aplicar ese aprendizaje a datos no observados anteriormente y obtener un resultado certero de clasificación.

Para los modelos supervisados, la eficiencia es parte del proceso de creación, ya que se reservan datos para una posterior validación del modelo, lo que significa que se compara la predicción del modelo sobre esos datos, con el dato real y así se obtiene un porcentaje de exactitud (*accuracy*).

Una forma de definir un buen nivel de *accuracy*, es utilizar un modelo sofisticado y conocido por su buen desempeño en modelos predictivos, como sería *Random Forest* o *Gradient Boosted Tree*. Ese resultado sería el *benchmark* para comparar con otros modelos más simples y rápidos de ejecutar, y escoger el que tenga el desempeño más cercano. (17)

Una herramienta fundamental para evaluar el desempeño de un modelo predictivo es la Matriz de Confusión. Con una estructura muy básica, entrega mucha información respecto al desempeño del modelo, cuantificando los eventos según su valor de predicción comparado con su valor real (18).

Ilustración 8: Matriz de Confusión

		Predicted	
		Negative	Positive
Actual	Negative	**a**	**b**
	Positive	**c**	**d**

Donde:

a = es la cantidad de predicciones correctas que un evento es negativo
b = es la cantidad de predicciones incorrectas que un evento es positivo
c = es la cantidad de predicciones incorrectas que un evento es negativo
d = es la cantidad de predicciones correctas que un evento es positivo

El *Accuracy*, es la razón entre las predicciones correctas y el total de predicciones, cuya fórmula sería: $$AC = \frac{a+d}{a+b+c+d}$$

Precision (precisión) es otro indicador del desempeño del modelo que nos entrega esta matriz, calculando la razón entre la predicción de eventos positivos correctos sobre el total de eventos positivos predichos. $$P = \frac{d}{b+d}$$

El *Recall o True Positive*, es la razón entre las predicciones positivas correctas y el total de eventos positivos, cuya fórmula sería: $$TP = \frac{d}{c+d}$$

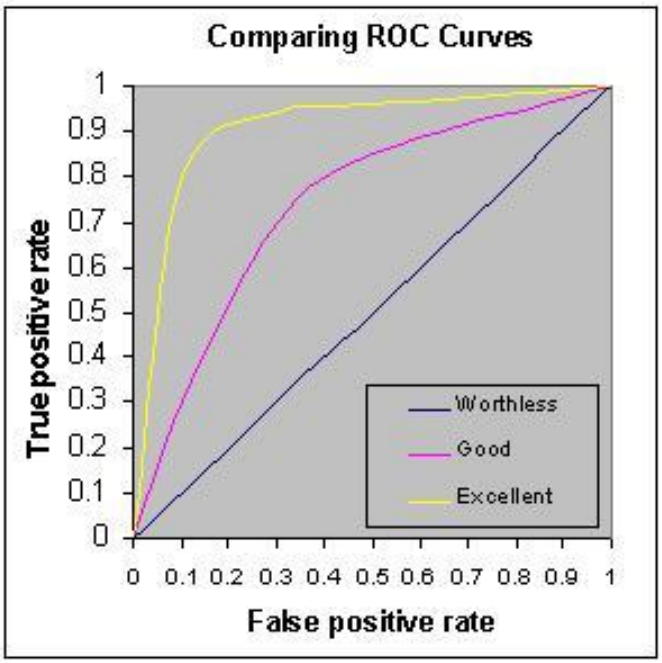
El *False Positive*, es la razón entre las predicciones negativas incorrectas y el total de eventos negativos, cuya fórmula sería: $$FP = \frac{b}{a+b}$$

De manera homóloga, tenemos las ecuaciones para *True Negative* y *False Negative* respectivamente: $$TN = \frac{a}{a+b} \qquad FN = \frac{c}{c+d}$$

Adicionalmente a la matriz de confusión, se puede usar las curvas ROC para medir el funcionamiento del modelo.

Son curvas que muestran la habilidad del clasificador para posicionar las instancias verdaderas respecto a las falsas (19) 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 área 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" (20)

Ilustración 9: Comparación de Curvas ROC



Fuente: http://gim.unmc.edu/dxtests/roc3.htm

6 Metodología

Basándonos en la definición de cliente fugado expuesta anteriormente, es que se formó la base de datos para el estudio, filtrando a los nuevos clientes ingresados durante los meses de septiembre a noviembre del 2017 y se analizó su estado en diciembre, para clasificarlo como fugado si no hubo actividad durante este último mes.

La base de datos inicial considera 9.744 nuevos usuarios en esos 3 meses, de los cuales 4.312 se clasificaron como fugados para el mes de diciembre, lo que entrega un nivel de fuga de 44,25% en 3 meses. Sin embargo, el primer mes es el de mayor impacto con un 56,92% de clientes fugados en 30 días o menos.

Ilustración 10: Tabla de Clientes Fugados por Mes

Mes de Registro	No Fugado	Fugado	% de Fuga
Sept.2017	2.232	1.091	32,83%
Oct.2017	1.819	1.396	43,42%
Nov.2017	1.381	1.825	56,92%
Total	**5.432**	**4.312**	**44,25%**

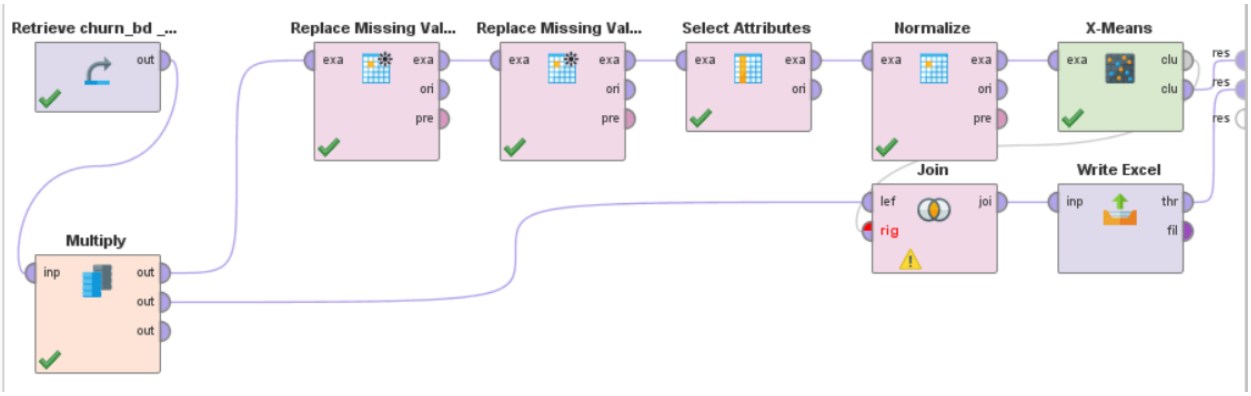
Fuente: Creación propia

Dentro de los nuevos usuarios detectados, encontramos la siguiente información:

- De los 9.744 usuarios, 5.830 son mujeres, 3.891 son hombres y 23 indeterminados.
- El ingreso de nuevos clientes es similar en los 3 meses evaluados, siendo el promedio de 3.248 personas.

Para esta primera etapa, se clasificaron los datos en 4 clusters, usando el programa RapidMiner™ y el proceso de X-Medias.

Ilustración 11: Modelo de Clasificación



Fuente: RapidMiner™

Para poder seleccionar los clusters, primero se reemplazaron valores faltantes de uso por “0”, indicando que si no había registro de actividad en esas categorías este es nulo.

El segundo reemplazo de valores faltantes es la edad de los usuarios, que luego de hacer varias pruebas, se comprobó que el uso de la mediana o la media no afectaba el resultado de los clusters, por lo que se dejó la media como parámetro.

Como segundo paso, se eliminaron los atributos que no aportaban información relevante, como variables con información duplicada por el reemplazo de un atributo nominal por varias variables dummy, atributos con la misma información para todos los usuarios o atributos relacionados a fechas.

Luego se procede a utilizar una función de transformación, para el caso del dataset, se utilizó la normalización estándar de manera que los atributos que poseen magnitudes diferentes no afecten en mayor medida el modelo.

Se usa X-medias para generar los clusters y se une la información a la base de datos original, formando un nuevo set de datos, que incluye la clasificación del cluster al que pertenecen cada uno de los clientes.

Para encontrar el modelo que mejor se adapte a las necesidades de una aplicación, es que se ha hecho una comparativa entre 7 modelos diferentes: Naive Bayes, Deep Learning, Logistic Regression, Decision Tree, Gradient Boosted Tree, Generalized Linear Model y Random Tree.
Esto se lleva a cabo usando el Auto Model de RapidMiner™, cargando la nueva data en el programa y seleccionando los atributos relevantes para la predicción, esto significa obviar data como "First Login" y otros atributos que tienen el mismo valor y no genera distinción entre un cliente u otro.

Siguiendo el método KDD, procedemos al análisis de los datos, adaptando la información disponible según las distintas necesidades de cada modelo.
La variable objetivo es el estatus de "Fugado" y se cuenta con 41 atributos que incluyen características demográficas, de uso, tiempo y software. Se utilizará un muestreo aleatorio para separar el 90% de la base de datos para la creación del modelo y se usará el 10% restante para validarlo.

Los datos seleccionados cuentan con 567 datos faltantes en 12 atributos de uso, lo que se puede atribuir a nula actividad y reemplazar por ceros.

Para cada uno de los modelos usados y comparados, se requieren tipos de datos distintos, por ejemplo, el modelo de Deep Learning, solo procesa datos de tipo Binomial, pero los Árboles de Decisión pueden usar datos del tipo Entero aparte de Binomiales, es por esto que cada uno de los modelos se estandarizan según sus capacidades y limitantes.

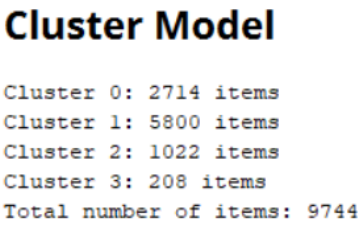
Basándonos en la Exactitud, el tiempo que demoran y las curvas ROC, se escogerá el modelo que será usado para la predicción.

Los modelos creados se prueban sobre la data reservada para validar el resultado, obteniendo distintos niveles de acierto en la predicción de fuga de los clientes.

7 Resultados

Los clusters creados identifican 4 tipos de clientes, la principal diferencia es el nivel de interacción con la aplicación, agrupando a aquellos clientes de bajo o nula interacción en el Cluster 0 y los de mayor interacción en el Cluster 3.

Ilustración 12: Detalle de Clusters



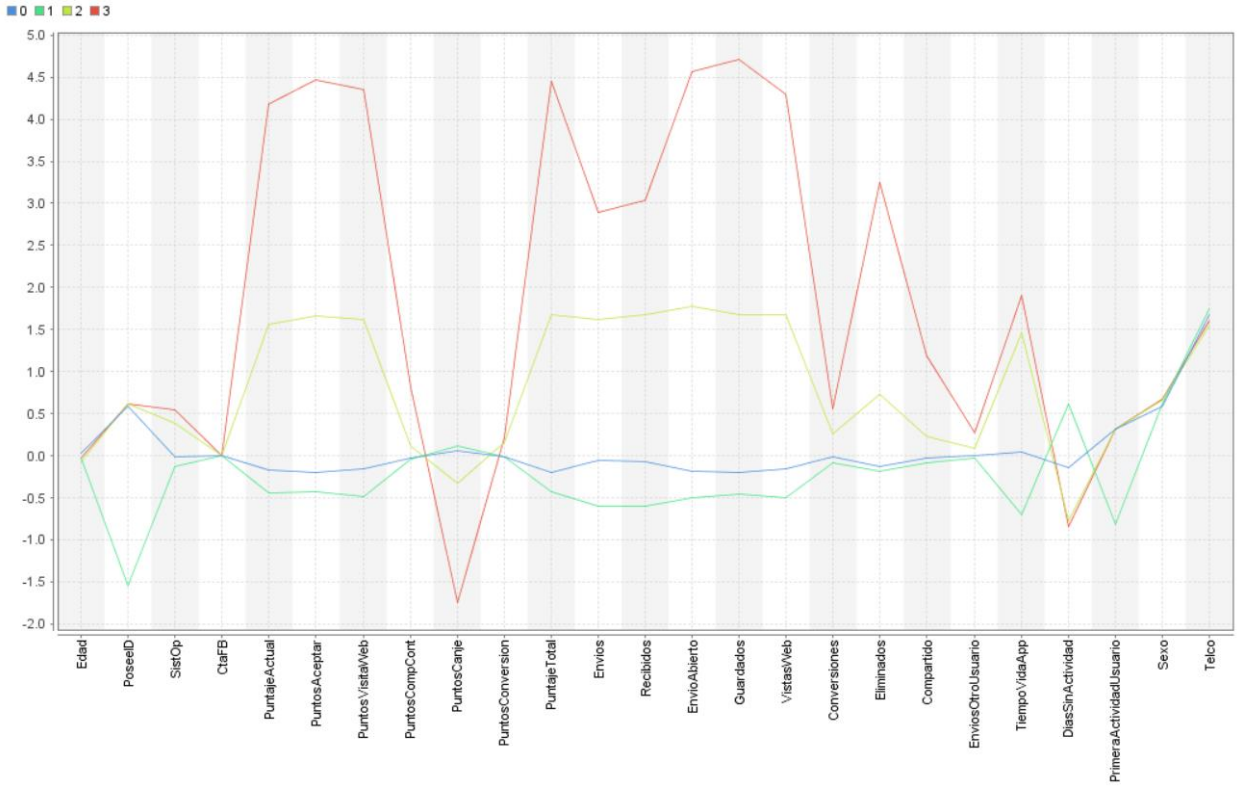
Fuente: RapidMiner™

Tabla 2: Tabla de Centroides

Atributo	Cluster_0	Cluster_1	Cluster_2	Cluster_3
Edad	0,024	-0,028	-0,072	-0,036
PoseeID	0,592	-1,543	0,614	0,616
SistOp	0,014	0,133	0,392	0,537
CtaFB	0,000	0,000	0,000	0,000
PuntajeActual	-0,169	-0,442	1,565	4,184
PuntosAceptar	-0,196	-0,432	1,656	4,467
PuntosVisitaWeb	-0,164	-0,480	1,613	4,351
PuntosCompCont	-0,026	-0,040	0,108	0,818
PuntosCanje	0,055	0,116	-0,325	-1,743
PuntosConversion	-0,020	-0,020	0,153	0,217
PuntajeTotal	-0,196	-0,436	1,674	4,443
Envios	-0,059	-0,602	1,623	2,889
Recibidos	-0,075	-0,596	1,676	3,029
EnvioAbierto	-0,182	-0,507	1,771	4,560
Guardados	-0,194	-0,462	1,681	4,703
VistasWeb	-0,160	-0,501	1,668	4,292
Conversiones	0,017	-0,087	0,263	0,553
Eliminados	-0,130	-0,193	0,736	3,234
Compartido	-0,032	-0,092	0,229	1,177
EnviosOtroUsuario	-0,008	-0,030	0,081	0,270
TiempoVidaApp	0,045	-0,705	1,460	1,914
DiasSinActividad	-0,138	0,610	-0,768	-0,848
PrimeraActividadUsuario	0,314	-0,813	0,314	0,314
Sexo	0,587	0,614	0,662	0,678
Telco	1,674	1,744	1,547	1,606

Fuente: RapidMiner™

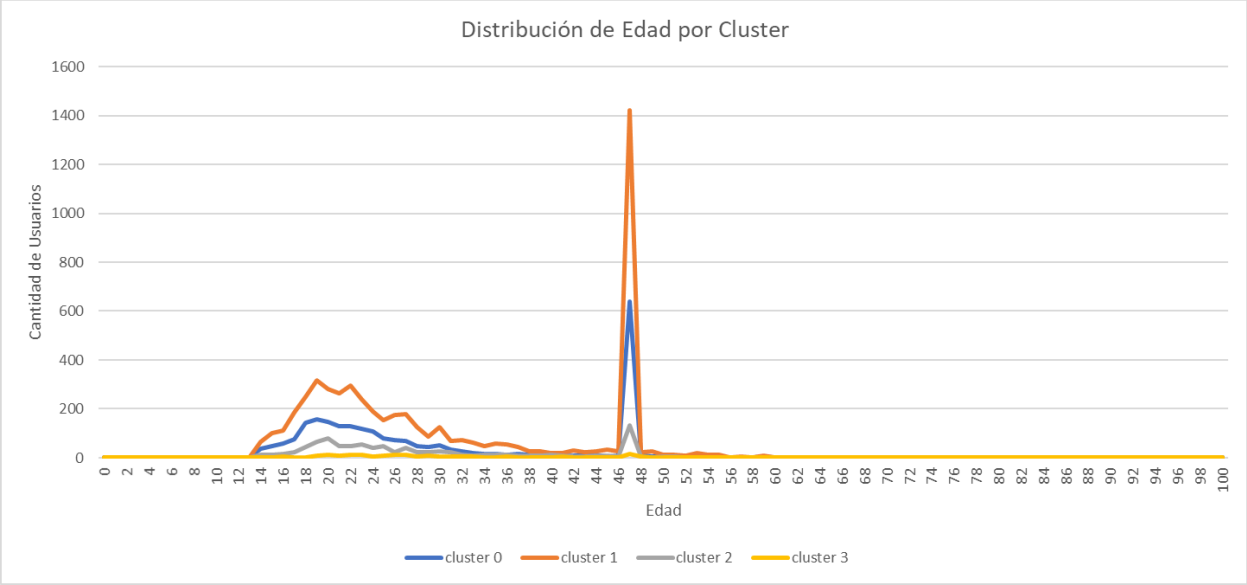
Gráfico 1: Gráfico de Centroides



Fuente: RapidMiner™

El gráfico de centroides es una buena herramienta para ver las variables que separan a cada cluster, por ejemplo “PuntajeActual”, “PuntosAceptar” y “PuntosVisitaWeb” están claramente separadas las líneas de cada cluster, pero si vemos “Sexo” y “Telco” están todas las líneas juntas, lo que implica que ese atributo en particular no es decisivo para la segmentación de los clientes.

Gráfico 2: Distribución de Edades por Cluster



Fuente: Creación propia.

Tabla 3: Resumen de Clusters
1: Cliente Fugado 0: Cliente Retenido

	cluster_0		cluster_1		cluster_2		cluster_3	
Values	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
Count of id_user	2.423	291	2.872	2.928	130	892	7	201
Average of PuntajeActual	1	28	82	247	914	1.234	2.231	3.008
Average of Envios	17	30	54	108	270	308	410	483
Average of Eliminados	0	0	1	2	23	17	3	75
Average of PuntajeTotal	0	0	81	114	918	880	2.231	2.219
Average of Recibidos	11	24	45	99	251	300	384	480
Average of PuntosVisitaWeb	0	0	4	7	37	37	96	92
Average of PuntosCompCont	0	0	0	0	3	1	85	7
Average of DiasSinActividad	44	13	37	4	19	1	9	0
Average of TiempoVidaApp	4	7	15	27	52	58	74	69
Average of PuntosAceptar	0	0	78	110	896	859	2.071	2.202
Average of Edad	31	33	31	32	28	29	26	31

Fuente: Creación Propia

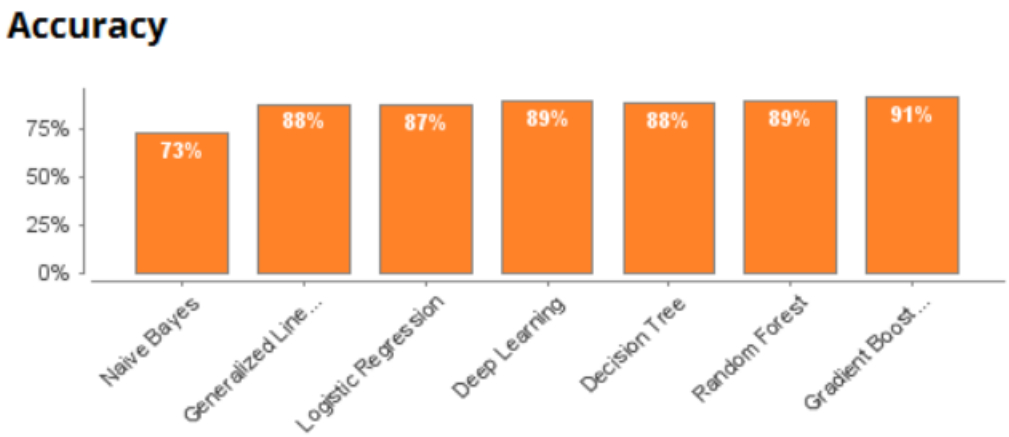
Tabla 4: Resumen de Fuga por Mes de Ingreso

Mes de ingreso	No Fugado	Fugado
Sept.2017		
Count of id_user	2.232	1.091
Average of DiasSinActividad	60,68	6,32
Average of TiempoVidaApp	16,69	69,17
Oct.2017		
Count of id_user	1.819	1.396
Average of DiasSinActividad	34,55	4,31
Average of TiempoVidaApp	9,92	38,64
Nov.2017		
Count of id_user	1.381	1.825
Average of DiasSinActividad	12,27	2,21
Average of TiempoVidaApp	2,48	10,10

Fuente: Creación Propia

Los 7 modelos creados, obtuvieron un rendimiento similar, ubicándose en el rango comprendido entre [73% - 91%] de *accuracy*, con tiempos de ejecución entre [355 ms y 2 min 43 s].

Gráfico 3: Accuracy por Modelo



Fuente: RapidMiner™

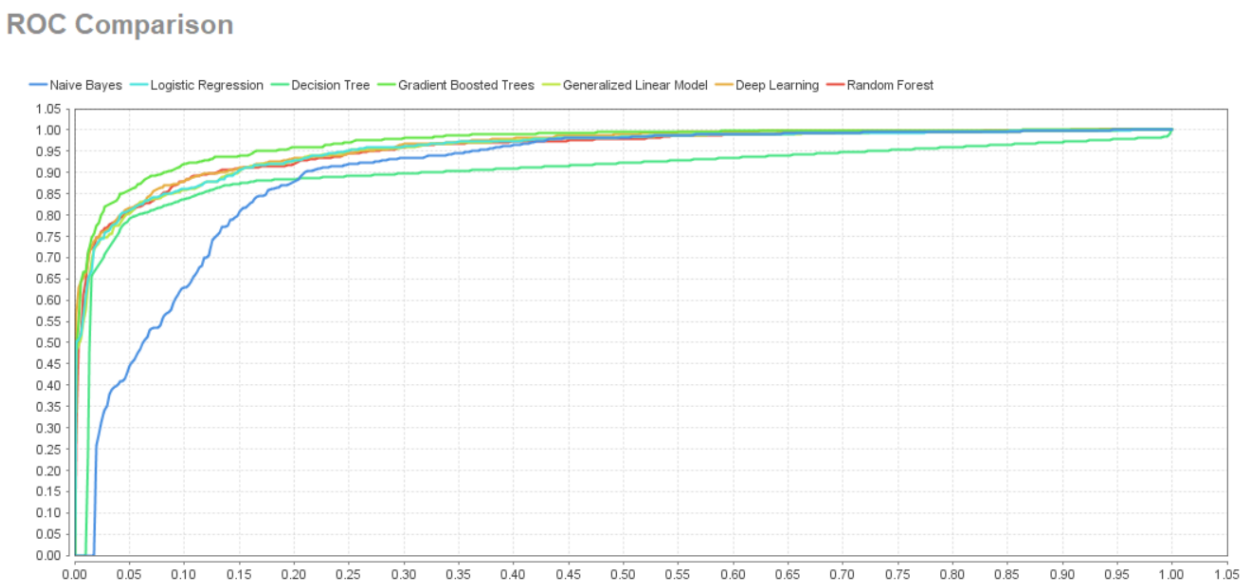
Tabla 5: Resumen Comparativo de Modelos

Model	Accuracy ↓	Runtime
Gradient Boosted Trees	91.1%	2 min 43 s
Deep Learning	89.2%	6 s
Random Forest	88.9%	1 min 31 s
Decision Tree	87.9%	3 s
Generalized Linear Model	87.5%	592 ms
Logistic Regression	87.3%	559 ms
Naive Bayes	72.5%	355 ms

Fuente: RapidMiner™

La información de los modelos representada en la Tabla N°5, nos indica que el nivel de exactitud que tienen es similar, el de peor rendimiento sería Naive Bayes.

Gráfico 4: Comparación de ROC



Fuente: RapidMiner™

De forma análoga al análisis anterior, las curvas ROC de cada modelo se observan bastante juntas, exceptuando Naive Bayes que se escapa del grupo.

Si bien todos los modelos están dentro de rangos aceptables para la predicción, el modelo escogido para predecir el comportamiento de los clientes es Deep Learning. Efectivamente el Árbol Potenciado tuvo el mejor rendimiento, pero el tiempo de ejecución es 27 veces el tiempo demora Deep Learning, por lo que usarlo de benchmark sería lo correcto.

Tabla 6: Matriz de Confusión Deep Learning

accuracy: 89.27%

	true range1	true range2	class precision
pred. range1	989	112	89.83%
pred. range2	97	750	88.55%
class recall	91.07%	87.01%	

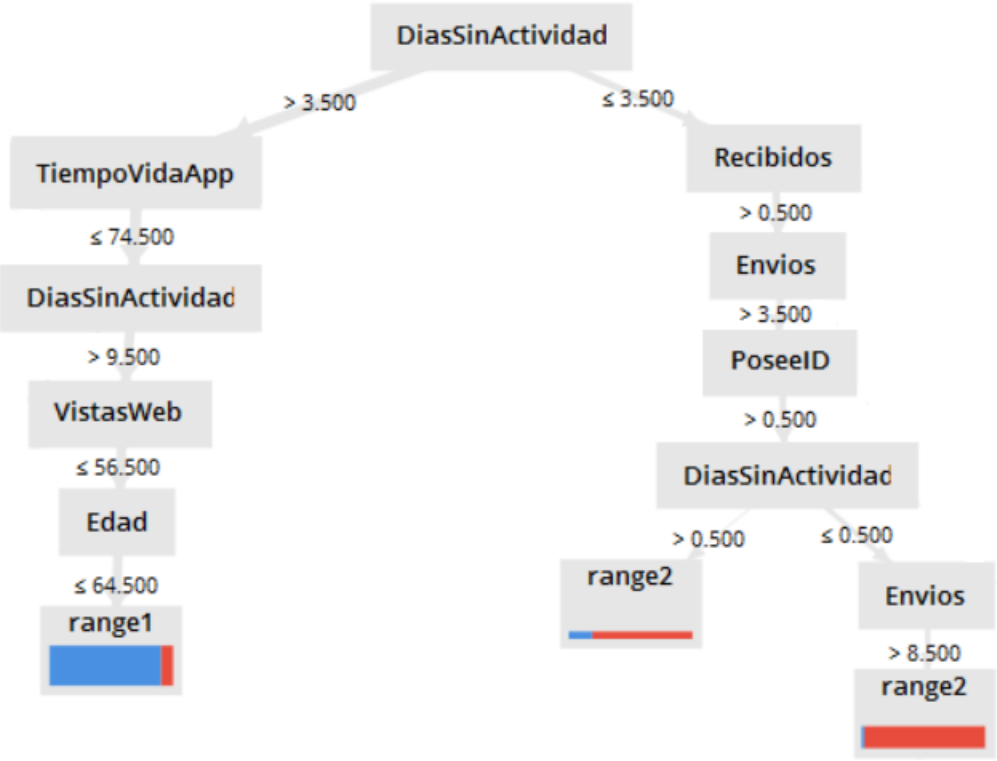
Fuente: RapidMiner™

La matriz de confusión se ve balanceada, los niveles de Precisión (87,01%) y Recall (88,55%) indican la capacidad de predicción correcta de fuga de clientes y están bastante cercanos a los valores de predicción de los no fugados, lo que compone un modelo balanceado y deseable para la predicción.

Como Deep Learning no entrega información de la relevancia de los atributos para el modelo, analizaremos el Árbol de Decisión, que, si bien tiene un 1% menos de exactitud que Random Forest, es de análisis más simple al ser una versión simplificada de modelamiento.

Con 3 ramas del árbol de decisión, podemos explicar el 69,50% de los datos y el 68,33% de los clientes fugados.

Ilustración 13: Muestra del Árbol de Decisión



Fuente: RapidMiner ™

En la segunda rama, concentramos el mayor porcentaje de clientes fugados, dónde el 48,26% de los clientes fugados borra la app en menos de 1 día después de recibir contenido. 5 de 7 atributos significativos para la clasificación son de uso, por lo que la interacción con el cliente es un atributo sensible para la retención, aunque a mayor interacción, mayor es la probabilidad de fuga, si observamos la segunda rama, recibir un mensaje, enviar más de 8 mensajes y poseer ID, son los atributos de mayor impacto para el fugador.

Para ver el árbol completo revisar Anexo N°2.

El modelo creado se puede revisar en el Anexo N°3

8 Conclusiones y Recomendaciones

Basándose en el análisis preliminar, podemos concluir lo siguiente:

- El nivel de fuga de la app es casi la mitad del promedio de fuga de aplicaciones estimado para 3 meses.
- La edad no es un factor determinante del uso de la *app*. Si observamos el Gráfico N°2 y la Tabla N°2, se evidencia que la distribución y el promedio de edad por cluster es muy similar.
- Los clientes se clasifican según el nivel de uso de la *app*. Como se observa en la Tabla N°2, "PuntajeActual", "PuntosAceptar" y "Envios", entre otros, el cluster_0 es el que tiene el menor nivel de actividad y el cluster_3 el mayor.
- Llama la atención, que el atributo "DiasSinActividad" es muy bajo en los clientes fugados, cosa que es contradictoria al sentido común. Se podría pensar, que los clientes no fugados, tienen un menor número de días entre el último contenido enviado y el último visto, pero podría explicarse por la eliminación de la app inmediatamente después de recibir contenido.
- Otro análisis relevante, es que los clientes fugados tienen un mayor nivel de actividad en la *app* que los clientes no fugados del mismo cluster. Analizando la Tabla N°3, tienen más bananas en promedio, más mensajes y descargas, por nombrar algunos.
- El primer mes de uso de la *app* es crucial en la retención de clientes, siendo el de mayor fuga y con un promedio de 10 días de uso como muestra la Tabla N°4, "TiempoVidaApp" para los fugados en noviembre.

Se recomienda las siguientes acciones:

1. Analizar en profundidad la razón de fuga de clientes. Una herramienta para obtener esta información es la Encuesta. Se puede contactar de forma directa a los usuarios o enviarles un email para responderla de forma online.

   Posibles razones pueden ser:

   a. Calidad del contenido: no es de interés del usuario la información enviada. En este caso, se recomienda aumentar la fuerza de venta, para tener mejores ofertas y mayor diversidad.
   b. Cantidad de contenido: podría percibirse como invasiva (spam) la app si el usuario se moviliza por zonas de interacción de forma regular. Una posible solución sería otorgar mayor control al usuario sobre el contenido que recibe, pudiendo limitar la cantidad o el tipo de información que se le envía.
   c. Beneficios insuficientes: la dificultad en el canje, disponibilidad y variedad de premios canjeables pueden ser factores que desincentiven el uso de la app. Otro factor relacionado, puede ser el tiempo que demora juntar bananas suficientes para canjear productos. La solución propuesta en el punto a. es efectiva para este problema también, mejorando la fuerza de venta se pueden conseguir mejores premios.
   d. Dificultad para interactuar con la app: hay muchos pasos intermedios entre comenzar a ganar bananas y lograr el canje de un producto, lo que puede generar desinterés en su uso. Para este caso, se podría tener una versión *Lite*, que involucre, por ejemplo, sólo ver videos promocionales a cambio de Bananas, simplificando el uso de la app.

2. Analizar qué tipo de cliente es el más rentable y estable para la *app*. Hay 5 formatos de interacción entre los clientes de la *app* y el usuario de la *app*. Con la información actual no es posible identificar las preferencias de los clientes cautivos o la proporción de preferencias de los fugados. Esto permitirá rentabilizar los esfuerzos de retención. Por ejemplo, si el costo fijo de mantener el formato “Videos” activo es superior al ingreso que generan, se debería evaluar el cierre del formato. Relocalizar el RR.HH podría generar un ingreso marginal mayor si se potencia otra área de la empresa o disminuir el costo fijo en caso de optar por la reducción de personal.

9 Referencias

01. (14 de Diciembre de 2008). *Janna Quitney Anderson, Lee Rainie.* Obtenido de Pew Research Center: http://www.pewinternet.org/2008/12/14/the-future-of-the-internet-iii/

02. (15 de Diciembre de 2015). *Walter van der Scheer*. Obtenido de Datafloq: https://datafloq.com/read/big-data-survey-2015-4-core-insights-success-data/1754

03. (22 de Marzo de 2018). *Justina Perro*. Obtenido de Localytics: http://info.localytics.com/blog/mobile-apps-whats-a-good-retention-rate

04. (08 de Octubre de 2018). Obtenido de Facebook: https://www.facebook.com/business/news/grow-your-business-with-more-engaged-app-users

05. (s.f.). *Khalid Saleh*. Obtenido de Invesp: https://www.invespcro.com/blog/customer-acquisition-retention/

06. (20 de Septiembre de 2016). *Andrés Zeledon.* Obtenido de Next U: https://www.nextu.com/blog/tres-principales-de-aplicacion-movil/

07. (s.f.). *Origen de las Aplicaciones Móviles.* Obtenido de https://appsmovilescavucm.wordpress.com/origen/

08. (3 de Noviembre de 2015). *Raposo, E.* Obtenido de http://www.ppmmobile.com/single-post/2015/11/03/Historia-y-evoluci%C3%B3n-de-las-APPs-m%C3%B3viles

09. (25 de Mayo de 2015). *Apps Made in Chile.* Obtenido de La Tercera: https://www.latercera.com/noticia/apps-made-in-chile/

10. (27 de Septiembre de 2016). *7 de cada 10 latinoamericanos disfrutan de la libertad de estar conectados...* Obtenido de Nielsen: https://www.nielsen.com/pr/es/insights/news/2016/7-

de-cada-10-latinoamericanos-disfrutan-de-la-libertad-de-estar-conectados-desde-cualquier-lugar-y-en-cualquier-momento.html

11. (11 de Junio de 2015). *so many apps so much more time for entertainment.* Obtenido de Nielsen: http://www.nielsen.com/us/en/insights/news/2015/so-many-apps-so-much-more-time-for-entertainment.html?afflt=ntrt15340001&afflt_uid=zhm6r8HxeBI.kmeCpa52n4aBmFPOSC5LjCML7rsy7qxa&afflt_uid_2=AFFLT_ID_2
12. (Octubre de 1996). *Data mining and knowledge discovery: making sense out of data.* Obtenido de U. M. Feyyad.
13. (24 de Septiembre de 2017). *Iñaki Ladero, Deep Learning: qué es y cómo se está usando.* Obtenido de Baoss: https://www.baoss.es/que-es-deep-learning-usos/
14. (s.f.). *Harry Zhang, The Optimality of Naive Bayes.* Obtenido de University of New Brunswick, Canada: http://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf
15. (1989). *Peter McCullagh; John Nelder, Generalized Linear Models.* Londres: Chapman and Hall.
16. (30 de Marzo de 2015). *Adele Cutler; Guohua Zhao.* Obtenido de ResearchGate: https://www.researchgate.net/profile/Adele_Cutler/publication/268424569_PERT-perfect_random_tree_ensembles/links/551940c00cf2d241f355ee7b/PERT-perfect-random-tree-ensembles.pdf
17. (20 de Abril de 2018). *Jason Brownlee*. Obtenido de Machine Learning Mastery: https://machinelearningmastery.com/how-to-know-if-your-machine-learning-model-has-good-performance/

18. (9 de julio de 2018). *Howard Hamilton, Confusion Matrix*. Obtenido de University of Uregina, Canadá: http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html

19. (2003). *John Wang , Data Mining: Opportunities and Challenges.* IGI Publishing.

20. (1996). Usama Fayyad, Gregory Piatetsky-Shapiro & Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases. *AI Magazine.* , Volume 17 Number 3.

10 Anexos

Anexo N°1

WAP:

Un sistema WAP consiste en tres partes principales:

- Una pasarela o gateway WAP
- Un servidor HTTP
- Un dispositivo WAP

Pasarela o gateway WAP

Esta pasarela actúa como un mediador un dispositivo celular y un servidor Web HTTP. Básicamente, enruta peticiones del cliente (teléfonos móviles) a un servidor HTTP (Web). Este gateway WAP puede estar localizado en una red una compañía telefónica o en un proveedor de servicios.

Servidor Web HTTP

Es el elemento que recibe la petición de la pasarela WAP, procesa dicha petición y finalmente vuelve a enviar la salida a la pasarela de nuevo. La pasarela entonces mandará la información al dispositivo WAP (teléfono móvil).

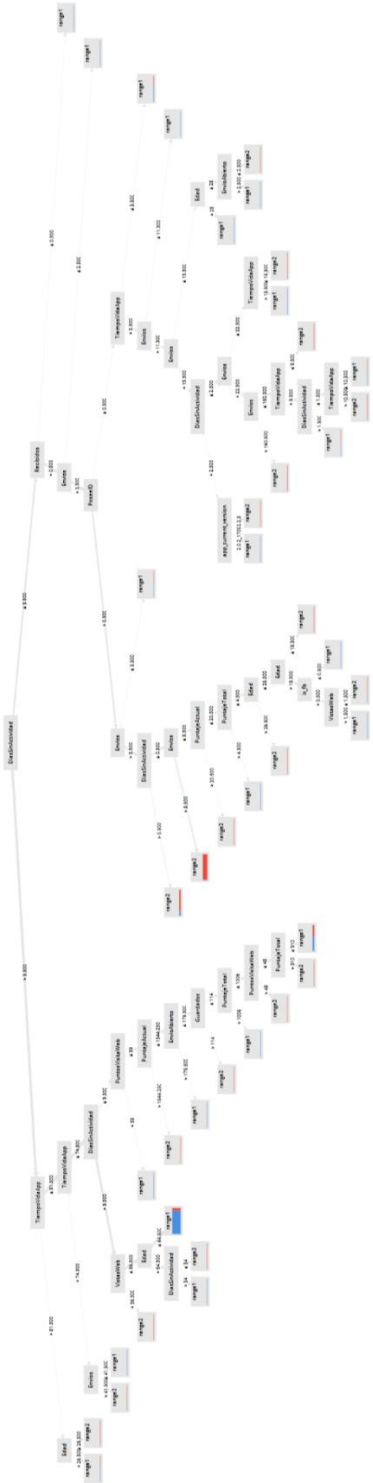
El dispositivo WAP

Son los teléfonos móviles, los cuales son partes de la red Wireless. Estos dispositivos envían la petición WAP a la pasarela WAP, la cual a su vez traduce dichas peticiones en un formato que

el servidor Web puede entender. Cuando la pasarela WAP vuelve a recibir la respuesta del servidor Web, lo vuelve a traducir en un formato WAP para que el dispositivo lo pueda interpretar.

Las páginas WAP son simples ficheros de texto con extensiones WML (*Wireless Markup Language*). Podemos definir WML como un lenguaje heredado de HTML, pero basado en XML, y que es usado para especificar contenido para dispositivos WAP. Se usa para poder crear páginas que pueden ser mostradas en un navegado WAP. WML usa lo que se llama WMLScripts para poder ejecutar códigos simples en el cliente. Se puede comparar de alguna manera con JavaScript, con la diferencia en que el consumo de memoria y CPU es bastante menor. (http://www.ordenadores-y-portatiles.com/wap)

Anexo N°2
Árbol de Decisión



Anexo N° 3

Modelo Árbol de Decisión