Índice

1. Definición del problema

En este capítulo se abordará de forma general el contexto del estudio realizado. En primera instancia se tratará el tema de los difusores de contenido audiovisual a través de internet, y cómo este concepto se ve relacionado con **Kikvi**, empresa de la cual se extrajeron los datos para los posteriores capítulos. En segundo lugar, se abordará el problema actual por el que pasa la empresa mencionada, y finalmente los objetivos, principales y secundarios, de este estudio. Se espera durante este capítulo dar un marco general del estudio e informar sobre los motivadores que llevaron a realizarlo.

1.1. Difusores de contenido audiovisual

Desde la masificación de las redes sociales, en el pasado con plataformas como MySpace, y hoy con plataformas muy masificadas como Facebook o Instagram, se ha buscado la forma de sacar provecho de esta interacción. Una gran cantidad de negocios se ha formado en torno a este concepto, principalmente relacionados con la publicidad. De la misma manera, la publicidad se ha ido adaptando a este nuevo entorno, haciéndose cada vez más sutil en algunos casos (*product placement*[1]), y aún más invasiva en otros, como es el caso de portales de contenido viral con *banners* y *popups*[2]. Con el paso del tiempo, los usuarios de internet se han vuelto reacios y menos susceptibles al segundo tipo de publicidad,

[1] Conocido en español como "publicidad por emplazamiento", consiste en la inserción de un producto, marca o mensaje dentro de la narrativa de un programa, en este caso en particular, fotografías o videos compartidos en redes sociales. Es un tipo de publicidad sutil indirecta. [Fuente: Wikipedia.org]

[2] Un banner es un formato publicitario característico de internet. Consiste en incluir una pieza publicitaria dentro de un sitio web, cuyo objetivo es atraer tráfico al vínculo correspondiente al banner. Un *popup* es una ventana emergente dentro de un sitio web, que por lo general cuenta con un *banner*, formulario o alguna forma de captura de información dentro de él. Su función final suele ser la misma que el *banner* pero de una forma más agresiva. – [Fuente: Wikipedia.org]

ocurriendo fenómenos como la llamada **ceguera del banner**[3], lo que hace necesaria una forma de publicidad más sutil, o más atractiva, que un *banner* tradicional. Es en torno a esto que se generan los difusores de contenidos.

Un difusor de contenido, como dice su nombre, cumple la función de **propagar y difundir contenido especializado a través de redes sociales**. La mayoría de los difusores aún enfocan su modelo de negocios en *banners*, usando la difusión en redes sociales de contenidos altamente atractivos para traer finalmente visitas a sus portales web. Kikvi es un difusor de contenidos audiovisuales, pero no enfoca su esfuerzo en *banners*; en él se explora la segunda posibilidad comentada de publicidad en internet, la publicidad sutil. A continuación, se revisará la historia y funcionamiento en detalle de Kikvi.

1.2. Kikvi

Kikvi nace como un proyecto de un grupo de estudiantes en la Feria de Creación de Software de la Universidad Técnica Federico Santa María en el año 2012. En aquel entonces, bajo otro nombre (privado), la idea del producto consistía en una red de usuarios dividida en 2 grupos principales: **creadores**, y **publicadores**. El primer grupo estaba compuesto por personas generadoras de contenido audiovisual, que habían invertido tiempo y dinero en esto, y necesitaban rentabilizarlo. Su desafío se presentaba al momento de difundir dicho contenido, tomando como vía principal las redes sociales, pero sin contar con el "*peso"*, suficiente para llegar a una masa suficientemente atractiva de público objetivo. Es aquí donde entraría en participación el segundo grupo usuario de la plataforma, los **publicadores**. Este conjunto estaba compuesto por una serie de personas

[3] Este fenómeno se hace presente con la masificación de los *banners*. El primer *banner* recordado de la historia de internet, tuvo una conversión del 44%, esto quiere decir que de cada 100 personas que visitaron el sitio del banner, 44 hicieron *click* sobre él. Hoy, la conversión, en el mejor de los casos, llega a 2%. [Fuentes: Red de *Display* de Google]

experimentadas en el uso de redes sociales y con alto índice de influencia en sus círculos. La función de los publicadores consistía en compartir los contenidos generados por los creadores en sus respectivas redes sociales, siendo dinero la motivación. En el ideal del modelo de negocios, el grupo de **creadores** compraría una cierta cantidad de vistas de "calidad", ya que se seleccionaba a los usuarios **publicadores** a través de un proceso de filtrado, pagando una cantidad de dinero fijo por vista. Esta cantidad se dividiría entre los **publicadores** que participaran de la campaña (consiguiendo parte de ese dinero por cada vista que consiguieran) y la empresa.

El concepto tuvo un éxito relativo en su fase inicial, consiguiendo financiamiento a través de la incubadora **3IE** de la UTFSM.

Poco tiempo después de este hito, hubo diferencias de opiniones entre los fundadores del proyecto, lo que resultó en la separación de los mismos. Parte de los involucrados siguieron con el proyecto inicial y el resto emprendió nuevos caminos, agregando nuevos integrantes al equipo y formando **Playgue**, plataforma que tenía la misma idea mencionada anteriormente. Luego de poco tiempo de funcionamiento, se hicieron claras las falencias del modelo de negocios, respaldándose además en el bajo éxito del proyecto seguido por el otro grupo de socios originales:

- La idea de "ganar dinero" en internet por poco esfuerzo fue una idea que se explotó mucho en el pasado, lo que hoy genera una fuerte desconfianza por parte del usuario.

- Los clientes (creadores) eran muy reacios sobre el origen de las vistas y su legitimidad. Esto se veía potenciado por servicios de países asiáticos que ofrecían una gran cantidad de vistas a muy bajo costo.

- El grupo de publicadores no tardó en encontrar la forma de optimizar su tiempo y sistema, generando grupos de “ayuda” en Facebook donde, entre ellos mismos, cada uno veía repetidas veces los videos publicados por el resto, generando ganancias para todos (menos para el objetivo real del negocio y los creadores).

A estas alturas se tomó la decisión de modificar la manera en la que se estaba abordando el negocio. El concepto de “dinero” en internet producía rechazo, por lo que se adoptó una metodología de puntos. Además, se decide cambiar la dinámica del contenido del sitio, complementando el contenido de marcas y clientes con videos altamente atractivos, pero cuya recompensa de puntos era considerablemente menor a la obtenida por los videos “auspiciados”. Es junto con estos cambios que se hace un fuerte trabajo de diseño de interfaces y la plataforma toma su nombre actual: **Kikvi**.

Los puntos obtenidos a través del portal podían utilizarse para canjear sobre un catálogo de productos, partiendo de cosas simples como entradas dobles al cine, y llegando hasta productos de alto valor como consolas de videojuegos y cámaras para deportes extremos. Este acercamiento provocó gran revuelo y consiguió la participación de muchos usuarios. El desafío consistía en mantener un catálogo de productos constante sin gastar más dinero del que ingresaba en la empresa.

Fue al poco tiempo después que se decidió incluir concursos en la plataforma, lo que en teoría solucionaría dos aristas en las que se estaba teniendo problemas:

- Los canjes solían ser por una gran cantidad de puntos, lo que desmotivaba fuertemente a los usuarios.

- Los canjes significaban una gran inversión de dinero (para ser atractivos).

La inclusión de concursos al sistema significó un seguimiento mucho más cercano de los usuarios a la plataforma. Interactuando de forma activa por períodos de tiempo (o al menos esto se creía). Al poco tiempo los concursos habían tomado gran fuerza en la plataforma, desplazando a los canjes inmediatos.

Surge a estas alturas la necesidad de entender de mejor manera el negocio, los usuarios y los procesos de la plataforma. Hasta el momento se estaba avanzando a ciegas: funcionando en base a prueba y error. El hecho de tener información de los procesos y funcionamiento se vuelve una herramienta atractiva y poderosa y se toma la decisión de explotarla.

1.3. El problema

Kikvi funcionó durante largo tiempo a ciegas, sin mucho conocimiento de un mercado muy poco explotado y sin respaldos ni casos de éxito cercanos para seguir. Entender el negocio, sus procesos y sus usuarios se convierte en un foco de atención, para poder mejorar la experiencia y percepción general sobre el producto.

Con el avance del tiempo se hace insostenible mantener una metodología de prueba y error, y se hace necesario tomar los pasos precisos en la dirección correcta. Además, es importante poder definir qué es lo que se considerará un caso de éxito en la plataforma, tanto en relación a un usuario como a un video en particular, para de esta forma poder potenciar y emular este tipo de comportamientos.

Kikvi comienza desde los cimientos sin financiamiento, lo que limita las posibilidades de contar con personal especializado para áreas como marketing, o análisis de datos. Esto lleva a que la plataforma funcione de acuerdo a estipulaciones e hipótesis, sin tener claro si el camino emprendido o la forma de abordar el

problema que se pretende resolver con Kikvi, mejorando la difusión de campañas audiovisuales a través de redes sociales, son los correctos.

En el escenario de hoy, la empresa se ve limitada al momento de comunicarse y trabajar con nuevos clientes, hay diversas interrogantes que son recurrentes en torno a esto, como, por ejemplo:

- **¿Qué define un caso de éxito?**: esta pregunta es, de forma implícita, recurrente al momento de comunicarse con nuevos clientes. El cliente quiere saber de casos de éxito anteriores, quiere saber si hay algún referente en la plataforma, algo que indique que su inversión va a dar frutos. Es entonces que surge la siguiente pregunta "*¿Podríamos ver casos de éxito?*", que no es posible responder si no se tiene una concepción de lo que define un caso de éxito dentro de la plataforma.

- **¿Cuántos usuarios hay?**: otra pregunta recurrente hace referencia a la cantidad de usuarios registrados en la plataforma. Aparentemente, su respuesta es simple, pues una consulta a la base de datos puede responderla sin dificultad. Si bien el número de usuarios se puede saber con total certeza, este número no es útil para el objetivo; lo que realmente importa es la cantidad de usuarios activos, o sea los usuarios que efectivamente se encuentran interactuando con la plataforma en un intervalo de tiempo. De nada sirve tener un sitio con cientos de miles de usuarios registrados, si sólo un porcentaje mínimo de ellos efectivamente es activo en la plataforma. Se hace necesario entonces saber reconocer un usuario activo, para así poder estudiar las variables del ambiente, y del usuario mismo, que lo hacen entrar en esta categoría.

Estas dos preguntas, totalmente válidas para un cliente que espera saber si vale la pena invertir o no parte de su presupuesto de marketing en **Kikvi**, actualmente se

encuentran sin respuesta. Es necesario entonces poder entender cómo se desenvuelven los usuarios en la plataforma, cómo interactúan con los videos, cuáles son las variables que hacen que un video sea exitoso, etc. El éxito de una campaña de un potencial cliente se ver estrictamente restringido por la respuesta de los usuarios de la plataforma ante ella, entonces, ¿cómo puede **Kikvi** apoyar esta campaña?

Es preciso que la plataforma se desarrolle de tal manera que optimice estos aspectos, que llame al usuario a mantenerse activo e interesado en las campañas existentes. Se deben descubrir los motivadores correctos y los escenarios ideales para obtener la mejor respuesta posible de la comunidad usuaria frente a las campañas de clientes. Además, es indispensable entender qué es lo que define y, aún más importante, cómo conseguir usuarios comprometidos con la plataforma, para así convertirla en una opción atractiva de inversión al momento de evaluar opciones de marketing digital.

1.4. Objetivos

Objetivo principal

Mejorar la percepción y experiencia usuaria de Kikvi para incrementar el éxito y penetración de campañas de clientes en un difusor de contenidos audiovisuales (Kikvi).

Objetivos específicos

Para poder lograr el objetivo principal propuesto es necesario, en primera instancia, realizar una serie de pasos relacionados con los datos:

- Hacer distinción de casos de éxito dentro de la plataforma, para así poder analizarlos y replicarlos, aumentando por un lado la satisfacción real del cliente, y por el otro la percepción del usuario.

- Descubrir qué indicadores son de interés para videos, usuarios y la plataforma en general, con el fin de tener una percepción de dónde enfocar esfuerzos y recursos para afectar de manera positiva al negocio. De la misma manera, ver cómo afectan estos indicadores a la percepción usuaria, para así poder mejorarla.

- Mejorar el porcentaje de rebote de usuarios en la plataforma; esto quiere decir que se pretende que los usuarios (visitas) no vengan con un objetivo específico a la plataforma y se vayan, si no que se distraigan, interactúen y exploren Kikvi. De esta manera, una visita no empieza y termina con la vista de un video, si no que puede significar un apoyo a otras campañas y, aún mejor, una adquisición de un nuevo usuario.

2. Estado del arte

En este capítulo se revisarán conceptos y conocimientos específicos que darán contexto a las herramientas y procedimientos utilizados durante este estudio.

Se espera entonces informar sobre las herramientas actuales para abordar esta clase de problemas, ahondando en las que fueron utilizadas para llevar a cabo este estudio.

2.1. Minería de datos

De forma general, la minería de datos consiste en el proceso de analizar datos[4] de múltiples fuentes, desde diferentes perspectivas, con el fin de resumirla en información[5] útil, o sea, la que pueda ser utilizada para aumentar ganancias, disminuir costos, mejorar procesos, etc. Entonces, un software de minería de datos es una herramienta analítica para datos.

Las herramientas de minería de datos permiten a sus usuarios analizar datos recopilados desde muchas dimensiones o ángulos diferentes, resumiéndolo todo en una serie de relaciones identificadas entre las variables estudiadas. Por lo general, la minería de datos se utiliza para encontrar correlaciones o patrones entre docenas de variables, o para encajar en el contexto, campos, de una gran base de datos relacional. A pesar de que la minería de datos es un término relativamente nuevo, la tecnología no lo es. Las compañías han utilizado por mucho tiempo computadores de alto rendimiento para iterar sobre grandes volúmenes de datos con el fin de generar reportes de interés para análisis durante años. Sobre estos escenarios, la innovación continua sobre herramientas computacionales como procesadores, discos de

[4] Hecho, número, o texto que puede ser procesado por un computador.

[5] Los patrones, asociaciones o relaciones entre datos pueden generar información. A diferencia de los datos, la información tiene uso, utilizad.

almacenamiento y software estadísticos, ha logrado incrementar dramáticamente la precisión de los análisis, mientras disminuyen los costos y tiempos de realizarlos.

De una forma muy simplificada, la minería de datos consiste en la identificación de patrones en conjuntos de datos, generalmente de grandes dimensiones, con el fin de adquirir algún conocimiento[6].

Los avances en métodos de captura de datos, procesamiento, transmisión de datos y almacenamiento, permiten hoy a las organizaciones integrar sus bases de datos en lo que se conoce como *data warehouses*. *Data warehousing* se define como el proceso de administrar y recuperar datos centralizados; representa la idea de mantener un repositorio central con todos los datos de una entidad. Esta práctica es necesaria para maximizar el acceso y posibilidades de análisis de los usuarios.

La minería de permite determinar relaciones entre tanto variables internas como externas de las compañías. Además, permite descubrir factores de retroalimentación como por ejemplo el impacto de una campaña en ventas, satisfacción de los clientes, etc. Con los resultados de un trabajo de minería de datos, un vendedor podría refinar el mercado objetivo de un producto para enfocar sus esfuerzos de campaña en esa dirección y lograr alta respuesta de clientes.

Un proyecto de minería está compuesto de 5 etapas principales:

- Extraer, transformar, y cargar datos en el *data warehouse (ETL[7])*.

- Almacenar y administrar los datos en un sistema de bases de datos multidimensional.

---

[6] La información puede ser transformada en conocimiento sobre patrones históricos o modas futuras.

[7] Por su sigla en inglés: *extract, transform, load.*

- Dar acceso a los datos a analistas del negocio y profesionales de TI.

- Analizar los datos con aplicaciones especializadas.

- Presentar la información en formatos útiles, como gráficos o tablas.

En relación a los niveles de análisis de la cuarta etapa mencionada, hay una serie de algoritmos y/o métodos utilizados comúnmente, como:

- Reglas de inducción: conjunto de reglas *if-then* útiles, basadas en significancia estadística.

- Árboles de decisión: estructuras en forma de árboles que representan una línea de diferentes resultados a través de una serie de decisiones. Estas decisiones generan reglas de clasificación para un conjunto de datos.

- Vecino más cercano: técnica que clasifica a cada registro en base a una combinación de sus *k* vecinos más cercanos[8].

- Redes neuronales artificiales: modelos predictivos no lineales que aprenden a través de entrenamiento.

- Algoritmos genéticos: técnicas evolutivas que usan procesos como combinaciones genéticas, mutaciones, y selección natural en un diseño basado en los conceptos de evolución natural.

[8] Su nombre en inglés es *k-nearest neightbours,* o por su sigla, *KNN*.

- Visualización de datos: interpretación visual de relaciones complejas en datos multidimensionales.

2.2. Procesos de minería de datos

En minería de datos hay variedad de procesos estándares para alcanzar los objetivos de la disciplina. A continuación, se revisarán los procesos más ampliamente utilizados.

Proceso de descubrimiento del conocimiento KDD[9]

Recibe este nombre el proceso que tiene por entrada la base de datos y sus versiones modificadas, y tiene como salida el subconjunto de patrones que se transformarán en conocimiento, luego de la aplicación de minería de datos.

De acuerdo con *[2]*, *KDD* es el proceso de usar métodos de minería de datos para extraer lo que es considerado conocimiento de acuerdo a una serie de medidas y umbrales, usando bases de datos en conjunto con cualquier pre-procesamiento necesario, extracción de muestras o transformación. El proceso cuenta con 5 fases fundamentales: Selección, Pre procesamiento, Transformación, Minería de datos, Interpretación/Evaluación.

En la **ilustración 1** se aprecian los 5 pasos del KDD. A continuación, se revisará más a fondo cada una de las etapas relacionadas con el proceso de descubrimiento del conocimiento.

El *KDD* es precedido por el desarrollo de un entendimiento del área de aplicación, cualquier conocimiento previo relevante y los objetivos del usuario final.

[9] *Knowledge discovery in databases*

Es un proceso iterativo e interactivo, e involucra numerosos pasos con muchas decisiones tomadas en el camino por el usuario.

1) **Selección**: esta etapa consiste en definir un conjunto de datos, o enfocar los esfuerzos en una serie de variables de los mismos. En esta, es fundamental contar con un conocimiento previo del negocio, que ayudará a definir cuáles variables son relevantes para el estudio y cuáles no lo son. Por ejemplo, si se desea descubrir qué clientes son más susceptibles a un esfuerzo de marketing, casi con certeza el nombre del cliente no será una variable importante para el estudio, pero sí el segmento económico o el nivel de ingresos del mismo.

2) **Pre procesamiento**: durante esta etapa se busca *limpiar* los datos. Este quiere decir que se tomará una serie de acciones para que los datos no cuenten con inconsistencias u observaciones faltantes/inválidas. Durante esta etapa se realiza una limpieza de los datos:

- **Faltantes**: en torno a esta situación se pueden tomar una serie de acciones, como ignorar datos con observaciones faltantes, llenarlos manualmente, usar una variable global para llenarlos (como N/A, -inf, etc), poner la media del atributo con respecto a todos los datos, usar la media del atributo considerando sólo los datos de la misma clase, o usar el valor más probable del dato.

- **Datos *ruidosos***: un dato ruidoso es una observación que tiene un error aleatorio en una variable medida.

- **Datos inconsistentes**: los datos inconsistentes se generan principalmente por variaciones al momento de ingresarlos, como el uso de diferentes capitalizaciones o faltas de ortografía. Una inconsistencia puede ser, por ejemplo, si en una observación de persona, su ciudad de residencia es “Santiago”, mientras que en otra es “stgo”, mientras que en otra es “Sanitago”.

Se entiende que todas las observaciones hacen referencia a la misma ciudad, pero por errores o decisiones humanas, tienen un valor diferente.

3) **Transformación**: en esta etapa se realizan todas las transformaciones necesarias a los datos para que puedan ser interpretados de mejor manera por los algoritmos de minería de datos. Dependiendo de los algoritmos a aplicar, se hace necesario aplicar uno o más tipos de transformación, siendo algunas de ellas:

- **Normalización**: consiste en representar los valores de las observaciones en un intervalo definido; por ejemplo, normalizar los datos para que sus valores estén dentro del rango [0,1]. Este método es de particular importancia cuando se planea utilizar técnicas de *clustering* basadas en distancia, ya que al no aplicarse, se desbalancea la importancia de diferentes variables por culpa de las unidades de medidas usadas. Por ejemplo, de distorsionará la distancia, dándole más importancia a una variable de mayor magnitud, como podría ser el ingreso per cápita de una base de datos de clientes (orden de los cientos de miles y millones) respecto de la edad.

- **Agregación**: utilizada cuando se desea agrupar variables. Por ejemplo, pasar una serie de registros de ingreso mensual a una cantidad más reducida de registros de ingreso anual.

- **Generalización**: como dice su nombre, consiste en generalizar variables. Se trata de reemplazar datos de variables de bajo nivel por un dato de niveles más altos. Por ejemplo, reemplazar datos de ciudades por regiones (Santiago a Región Metropolitana, Temuco a IX región, etc).

4) **Minería de datos**: esta etapa consiste en la búsqueda de patrones de interés en alguna forma particular de representación, dependiendo del objetivo final de la minería.

5) **Interpretación/Evaluación**: en esta etapa final, se interpretan y evalúan los patrones encontrados, con el fin de juzgar su utilidad para el objetivo final o negocio, además de su asertividad.

Ilustración 1: Proceso del descubrimiento del conocimiento (KDD)

(https://www.researchgate.net)

SEMMA (Sample, Explore, Modify, Model and Assess)

SEMMA es una serie de etapas secuenciales que guía a la implementación de aplicaciones de minería de datos. Su nombre es acrónimo para *Sample*, *Explore*, *Modify*, *Model & Assess*, lo que hace referencia a cada una de las fases del proceso:

1) *Sample* (**Muestreo**): consiste en la selección de un conjunto de datos para modelar. El desafío recae en que esta muestra debe ser lo suficientemente grande para que sea representativa, y lo suficientemente pequeña como para ser manejada de forma eficiente.

2) *Explore* (**Exploración**): durante esta fase se visualizan los datos, con el fin de entenderlos al descubrir relaciones, anticipadas como no anticipadas, entre las variables en ellos, además de la detección de anomalías.

3) *Modify* (**Modificación**): en esta etapa del proceso se realiza cualquier acción para seleccionar, crear y/o transformar datos con el fin de prepararlos para el modelo.

4) *Model* (**Modelado**): el objetivo de esta fase es aplicar varias técnicas de modelo sobre las variables preparadas con el fin de crear modelos que puedan posiblemente generar los resultados esperados.

5) *Assess* (**Evaluación**): última etapa de *SEMMA*, consiste en la evaluación de los modelos desarrollados, con el objetivo de juzgar si son suficientemente confiables y útiles.

Una crítica que se hace comúnmente a este proceso, es que deja los aspectos propios del negocio afuera del análisis, a diferencia de otros procesos como *CRISP-DM*[10], que cuentan con fases[11] enfocadas en estos aspectos.

[10] ***C**ross-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining*, o en español, Proceso estándar multi-industria para minería de datos.

[11] Fase *Business Understanding phase* (Fase de entendimiento del negocio) de *CRISP-DM*.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

CRISP-DM recibe su nombre del acrónimo en el título (en español, Proceso estándar multi-industria para minería de datos), y consiste en un ciclo compuesto de 6 etapas:

1) **Entendimiento del negocio**: en la primera etapa de *CRISP-DM*, se busca comprender los objetivos y requerimientos del proyecto desde el enfoque del negocio, para luego transformarlo en un problema de minería de datos y un plan preliminar para alcanzar los objetivos.

2) **Entendimiento de los datos**: comienza con un conjunto de datos inicial, y consiste en actividades con la finalidad de familiarizarse con los datos, identificar problemas de calidad en ellos, descubrir una primera mirada sobre los datos o bien descubrir subconjuntos interesantes para formular una hipótesis para información escondida.

3) **Preparación de los datos**: esta fase comprende todas las actividades necesarias para generar el set de datos final a partir de los datos en bruto.

4) **Modelo**: es la aplicación de varias técnicas de modelo, calibrando sus parámetros a valores óptimos.

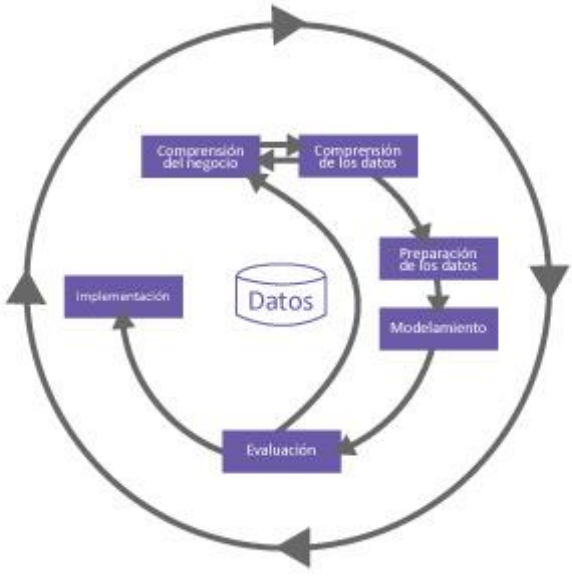
5) **Evaluación**: los modelos obtenidos son juzgados y los pasos para construirlos son evaluados con el fin de concluir con seguridad que efectivamente cumple con los objetivos del negocio.

6) **Despliegue**: el término del modelo por lo general no significa el fin del proyecto. El conocimiento obtenido luego debe ser organizado y desplegado de forma que el cliente final pueda utilizarlo.

De forma gráfica, se aprecia en la **ilustración 2** la serie de etapas que componen el proceso.

Ilustración 2: Ciclo de vida de CRISP-DM

(http://www.function1.com)



2.3. Tareas de minería de datos

En esta sección se revisarán los diferentes tipos de tareas de minería de datos. Además, se describirán cada una de las subcategorías pertenecientes a dichos tipos.

2.3.1. Tareas descriptivas

En este tipo de tareas el objetivo es describir los datos existentes. Busca proporcionar información entre las relaciones existentes en los datos y sus características. En el contexto, teóricamente se podría llegar a una afirmación como,

por ejemplo: el que un estudiante tenga actividades extra programáticas en el primer semestre, implica que también tendrá en el segundo.

Visualización

La tarea de visualización consiste en revisar los datos de forma mecánica, para revisar cualquier relación entre variables que se pueda apreciar en primeras instancias. Para facilitar esta tarea hay una gran cantidad de software en buenos estados de desarrollo, de donde destaca *Tableau*[12], que si bien se trata de una herramienta de procesamiento analítico de datos, puede ser utilizada también para visualización.

Correlaciones y factorizaciones

Esta tarea consiste en desplegar los datos y evaluar si se encuentra alguna correlación entre las variables pertenecientes al estudio. La correlación puede ser lineal, o pueden estar relacionadas de otra manera. Esta tarea solo puede ser realizada, por su naturaleza, sobre variables numéricas.

Asociación

La asociación es una tarea descriptiva, no supervisada, que hace referencia a reglas que son capaces de describir los datos en base a ocurrencias en las variables; en otras palabras, describe el comportamiento de una variable en base al de otra (u otras). Por ejemplo, una regla de asociación sería "si el año de ingreso de un estudiante es igual a 2008, ha tomado actividades extra programáticas y estudia arquitectura, entonces su colegio es subvencionado". Las reglas de asociación sólo pueden aplicarse sobre variables nominales (todas las involucradas). La asociación puede presentarse de dos maneras:

[12] *http://www.tableausoftware.com*

- Reglas de asociación: son asociaciones recíprocas, o sea, que hay una implicancia doble, describiendo cada una de las variables relacionadas a la asociación a la otra.

- Dependencias: a diferencia del caso anterior, este tipo de asociaciones son direccionales, o sea, el cumplimiento de una serie de condiciones implica que se cumplirán otras, y no al revés.

Los algoritmos de búsqueda de asociaciones tienen la particularidad de que la mayoría se puede descomponer en dos fases. La primera consta de la búsqueda de un conjunto de ítems frecuentes con un soporte[13] mayor o igual al deseado, o sea, que se buscan conjuntos de elementos que cuenten con cierto criterio establecido, sin separarlos aún. Luego, en la segunda fase, se hacen particiones de los conjuntos de ítems, calculando la confianza[14] de cada una, y reteniendo las reglas que tengan confianza mayor o igual a la deseada.

Segmentación (Agrupamiento)

Este tipo de tareas consisten en agrupar los datos en diferentes subconjuntos, o clases, de acuerdo a la relación entre ellos. Se busca que todos los elementos presentes en un grupo definido tengan propiedades parecidas, o sea, que sean similares entre sí y que sean diferentes a los elementos de otros grupos. La segmentación es una técnica de aprendizaje no supervisado, que utiliza el término de *distancia* para describir a los elementos; dos elementos tendrán poca distancia entre ellos si son parecidos (similares). Análogamente, tendrán harta distancia entre ellos si no son similares. En base a estos términos, la segmentación busca minimizar la

[13] Medida para cuantificar los casos en los que e antecedente se hace verdadero. Puede ser número de casos o porcentaje.
[14] Número de casos en que, habiéndose cumplido el antecedente de la regla, se cumple el consecuente.

distancia de los elementos pertenecientes a un mismo grupo y maximizar la distancia entre los grupos.

La definición de distancia puede variar de acuerdo a la fórmula que se utilice para calcularla:

- Distancia de Minkowski:

$$d_r(x,y) = \left(\sum_{j=1}^{J}\left|x_j - y_j\right|^r\right)^{\frac{1}{r}}, r \geq 1$$

- Distancia de Manhattan (Minkowaki, con r = 1)

$$d_1(x,y) = \sum_{j=1}^{J}\left|x_j - y_j\right|$$

- Distancia Euclideana (Minkowaki, con r = 2)

$$d_2(x,y) = \sqrt{\sum_{j=1}^{J}\left(x_j - y_j\right)^2}$$

- Distancia de Chebyshev (Minkowaki, cuando $r \rightarrow \infty$)

$$d_\infty(x,y) = \max_{j=1..J}\left|x_j - y_j\right|$$

Los algoritmos de segmentación pueden ser clasificados por diferentes tipos, los cuales se listan a continuación:

- Particionamiento: estos métodos construyen *k* particiones de un conjunto de datos, representando cada una de éstas un grupo. *K* debe ser menor o igual al número de elementos del conjunto de datos mencionado. Ejemplos: *K-Means, K-Medoids.*

- Jerárquicos: genera una descomposición jerárquica del conjunto de grupos, en otras palabras, crea una serie de subconjuntos de datos, en los que algunos engloban a otros. Ejemplos: *BIRCH, CURE*.

- Basados en densidad: de acuerdo a *[10]*, se define la pertenencia de cada elemento a un *cluster* si dicho elemento contiene una cantidad establecida de vecinos, dentro de una vecindad definida de radio mayor a 0. Ejemplos: *DBSCAN, OPTICS.*

- Basados en grilla: separa el espacio de datos en una grilla (finita), para luego realizar operaciones de agrupamiento sobre ella. Ejemplos: *STING, CLIQUE*.

- Basados en modelo: se utiliza un potencial de modelo para cada uno de los grupos, ajustando los datos a dichos modelos. Ejemplos: *COBWEB, CLASSIT* (estadísticos), mapas auto organizados (redes neuronales).

Detección de anomalías

La detección de valores e instancias anómalas es una tarea necesaria al momento de realizar minería de datos. En todo conjunto de datos se presentarán registros que se escapen de todo patrón o tendencia, y es importante poder reconocerlos para no considerarlos como un patrón común, si no como comportamientos anómalos como fraudes, fallas u *outliers*. Informalmente, un *outlier* es cualquier valor de dato que pareciera estar fuera de lugar con respecto al resto de los datos. De acuerdo a *[5]*:

“La definición intuitiva de un *outlier* sería 'una observación que se desvía tanto de las otras observaciones como para despertar sospechas de que fue generada por un mecanismo diferente'”[15]

2.3.2. Tareas predictivas

Las tareas predictivas, son problemas en los que se hace necesario predecir un (o varios) valores para un grupo de datos. La salida de una tarea predictiva es una categoría (a la que pertenece uno o más datos) o un valor numérico relacionado con el o los datos en cuestión. A continuación, se revisarán algunos de los tipos de tareas predictivas.

Clasificación

Las tareas de clasificación buscan definir un modelo que sea capaz de predecir la clase de un objeto que no la tiene definida. Además, pueden ser utilizadas para estimar un valor que se encuentre perdido o que no se tenga a priori. Estas tareas son supervisadas, por lo que se debe contar con un conjunto de datos de entrenamiento, que ya se encuentran clasificados. El proceso de generación de un modelo predictivo consta de 3 pasos:

1) División de los datos en dos conjuntos: entrenamiento y prueba.

2) Utilización del subconjunto de entrenamiento para la construcción del modelo.

[15] Textual en inglés: "The intuitive definition of an outlier would be 'an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism'"

3) Utilización del subconjunto de prueba para validar el modelo conseguido en el punto anterior, si el porcentaje de casos exitosos es aceptable, se valida el modelo (útil para clasificar otros casos).

A continuación, se revisarán algunos algoritmos de clasificación.

Árboles de clasificación

Los árboles de clasificación son un método en el cual se somete un dato a una serie de condiciones, que lo clasifican de acuerdo a los valores de las variables relacionadas con el mismo. Por ejemplo, se somete primero a un dato a la evaluación de una variable: "si el alumno tiene un promedio mayor a 55, entonces se cuestiona la variable *año de ingreso*; si no, se cuestiona la variable *plan de carrera"*, con el fin de predecir alguna variable en particular. Cabe destacar que no se trata de árboles binarios, si no que se pueden considerar numerosos intervalos o valores para cada variable para generar la clasificación.

Su estructura es similar a un diagrama de flujo, donde cada vértice simboliza una condición a la que se somete el dato a predecir. El último nivel del árbol, los nodos hoja, representan las clases. Su construcción suele llevarse a cabo con estrategias del tipo "dividir y conquistar"[16], empezando con todos los elementos del grupo de entrenamiento en la raíz, y continuando dividiéndolos en el atributo que se elija para ramificarlo.

Inducción de reglas de clasificación

Los métodos de inducción de reglas tienen las mismas propiedades que los métodos de árboles de decisión, describiendo una serie de condiciones *if-then* para llegar a la clasificación deseada. La obtención de dichas condiciones, o reglas, puede

[16] Conocido por su traducción en inglés *divide & conquer*.

ser a partir de un árbol de decisión, a través de algoritmos específicos como *STAR* o *Ripper*, o a partir de reglas de asociación. Además, es posible extraer reglas de clasificación desde una red neuronal, a través del algoritmo *MofN*, propuesto por *[7]*. En particular, permite la extracción de reglas desde una red neuronal multicapa, a través de los siguientes pasos de agrupamiento, extracción de reglas, agrupación de reglas y poda de reglas.

Métodos Bayesianos

Estas herramientas estadísticas son capaces de predecir las probabilidades de que un elemento en cuestión pertenezca a una clase en particular. Asumen que el valor de cada una de las propiedades es independiente de los valores de las otras (en un mismo elemento), llamada independencia condicional de clases. Los métodos Bayesianos se basan en el teorema de Bayes, y pueden ser utilizados tanto para fines descriptivos como predictivos. En el primer caso, se usan para descubrir relaciones de independencia y/o relevancia para poder realizar un estudio más profundo a través de inferencias estadísticas. En el segundo caso, se utilizan como clasificadores.

Métodos basados en casos y vecindad

Se caracterizan por utilizar el conjunto de entrenamiento para clasificar nuevos datos. En esta categoría hay presentes técnicas para segmentación, como *K-Means*, y para clasificación, como *LVQ*. Además, se utilizan métodos de ensamblaje, que combinan varios modelos con el objetivo de conseguir una mejor precisión final en el clasificador

Regresión estadística

A través de esta tarea, se busca generar una función matemática que sea capaz de estimar el valor de alguna variable de interés a partir del resto de las variables

relacionadas con un dato en particular. La regresión estadística puede ser utilizada únicamente para valores numéricos, y la función se puede calcular a través de interpolación, estimación o logística.

2.4. Herramientas de minería de datos

Hay una amplia gama de herramientas de minería de datos a disposición. Cada herramienta cuenta con implementaciones diferentes de una porción de los algoritmos más utilizados en los procesos de minería. Además, muchas de las herramientas cuentan con interfaces usuarias para facilitar el proceso de minería para quienes no tiene un conocimiento base de líneas de comando o programación. Algunas de las herramientas más utilizadas se listan a continuación.

- *Rapid Miner*: escrita en Java, esta herramienta de minería de datos funciona en torno a interfaces gráficas avanzadas, por lo que el usuario final requiere escribir muy poco código. Cabe destacar que esta herramienta se ofrece como servicio, más que como software local. Además, proporciona funcionalidades de pre-procesamiento y visualización, análisis predictivo, esquemas de aprendizaje y algoritmos de scripts de R. Esta potente herramienta es de código abierto, bajo licencia AGPL[17] (http://www.rapidminer.com).

- *Angoss*: enfocada principalmente para organizaciones involucradas con ventas, marketing y análisis de riesgo, esta herramienta cuenta con una interfaz gráfica avanzada además de un asistente amplio para sus procedimientos. Si bien las interfaces y asistente podrían ser restrictivos para usuarios avanzados, *Angoss* implementa soporte total de línea de comando en R, satisfaciendo así a los usuarios que prefieren personalización por sobre facilidad de uso. Una buena característica de esta herramienta es que tiene

[17] Para más información sobre la licencia de tipo AGPL, referirse a http://www.gnu.org/licenses/agpl-3.0.html.

una amplia gama de representaciones gráficas de datos. Si bien cuenta con una gama de implementaciones de los algoritmos más conocidos, no tiene suficientes herramientas para personalizar los procesos, por lo que no es la opción para quienes prefieran un ambiente fácilmente extensible (http://www.angoss.com).

- *KNIME (Konstanz Information Miner)*: esta herramienta nace como una solución para farmacéuticas a nivel empresarial. Los desarrolladores crearon un producto escalable, modular y de código abierto, teniendo la flexibilidad necesaria para adaptarse rápidamente a las demandas de un campo de estudio en crecimiento como es la minería de datos. Siguiendo su éxito en la industria farmacéutica, otras industrias siguieron la tendencia y utilizan *KNIME* para sus procesos de *CRM*[18] e inteligencia de negocios. Otra ventaja considerable que tiene esta herramienta, es que cuenta con una comunidad activa tanto de desarrolladores como de usuarios. No requiere conocimientos de programación para ser usada, ya que cuenta con interfaces intuitivas y de fácil uso (http://www.knime.org).

- *R:* más que una herramienta para minería de datos, R es un lenguaje de programación y ambiente para computación estadística y análisis. Es esta la razón que hace a R una potente herramienta de minería de datos. Bajo licencia GPL[19], y de código abierto, R puede ser personalizado abiertamente, sin restricción, lo que se traduce en una cantidad inigualable de algoritmos e implementaciones desarrolladas por usuarios alrededor del mundo, lo que se traduce en una herramienta flexible, escalable y extremadamente personalizable. Por otro lado, a pesar de que hay algunas interfaces para tratar con este lenguaje, se requiere conocimientos de programación (y del lenguaje

[18] *CRM*, o *Customer Relationship Management*, es el proceso que gira en torno a pulir y mejorar las relaciones con los clientes, a través de campañas de marketing específicas, servicio personalizado al cliente y gestión del equipo de ventas.
[19] Más información sobre licencia de R: https://www.r-project.org/COPYING

en sí) para sacar el máximo provecho de esta herramienta (http://www.r-project.org).

3. Diseño de la solución

Para este estudio se decide usar R, en primera instancia por las ventajas mencionadas en el párrafo anterior, y en segunda instancia porque la herramienta ya es conocida por quien realiza el estudio.

Para el proceso de desarrollo de la solución, se decide utilizar *CRISP-DM* debido a su integración y consideración de reglas del negocio.

Como primer acercamiento al alcance de los objetivos propuestos, se requiere identificar propiedades del negocio. Se ha detectado que el período de actividad de los usuarios no es satisfactorio para las proyecciones de funcionamiento de la plataforma. Se hace necesario entonces generar cambios que puedan mejorar la experiencia de los mismos y aumentar los niveles de interacción que tienen con el producto. Se abordarán en esta sección las dos primeras etapas de la metodología *CRISP-DM.*

3.1. Entendimiento del negocio

Se estudiará entonces cómo maximizar la cantidad de **usuarios activos** en un instante de tiempo, evaluando los escenarios en los que esta variable alcanza valores satisfactorios para poder replicar dichos escenarios. Por otro lado, se intentará también encontrar patrones e indicadores en los escenarios en lo que la variable mencionada se encuentre en sus valores más bajos, para así poder evitarlos.

La finalidad de buscar el aumento de usuarios activos en la plataforma recae en que, para los clientes del servicio, éste se vuelve más atractivo entre más usuarios activos posea. Es común que un negocio de este segmento cuente con una gran

cantidad de usuarios registrados, pero éstos no son los que tienen valor para el cliente final, ya que no necesariamente se encuentran actualmente en interacción con la plataforma. Por ejemplo, es más atractivo tener 10.000 usuarios, pero con un 50% de ellos activos (5.000 potenciales clientes), que tener 1.000.000 de usuarios, pero sólo con 1% de ellos activos (1.000 potenciales clientes).

Otra variable de particular interés para el negocio (y que se relaciona con la variable recién mencionada) es la **penetración** de un video en particular. Este indicador, a diferencia de usuarios activos que busca patrones de la plataforma en un instante de tiempo, se enfoca en el video mismo en cuestión (principalmente), como la duración. No se descarta que haya variables y patrones en la plataforma (externos al video mismo) que afecten en su penetración.

Una mayor penetración significa un mayor éxito en la campaña de los clientes de Kikvi, ya que se traduce en que el video de la campaña alcanzó una mayor cantidad de personas, que es justamente lo que ellos buscan en un servicio como este: llegar a la mayor cantidad de personas con la campaña en cuestión.

Tomando como enfoque principal los usuarios, se hace necesario poder identificar de forma eficiente la calidad de estos mismos, en base a su forma de interactuar con la plataforma en el tiempo. De esto se desprende una nueva variable de interés: **calidad usuaria**. Este indicador tiene un particular interés para quienes administran la plataforma, ya que el saber qué es lo que define a un usuario de calidad puede llevar al conocimiento de cómo se genera o consigue dicho usuario, para luego poder replicar el proceso y construir una base de usuarios de calidad para los objetivos finales del cliente, entregando finalmente un mejor servicio. Además, el entender cada una de las clases usuarias servirá para enfocar esfuerzos en la dirección correcta al hacer campañas de publicidad para conseguir usuarios nuevos.

Los indicadores que finalmente serán de interés para este estudio se pueden apreciar de manera resumida en la **tabla 1**.

Tabla 1: Indicadores de interés

Indicador	Descripción
Usuarios activos	Cantidad de usuarios activos (que han tenido actividad con la plataforma en la última semana) en un instante de tiempo.
Penetración	Proporción de usuarios activos que comparten un video en particular.
Calidad usuaria	Hace referencia a la frecuencia, dimensión y extensión de la forma en que un usuario interactúa con la plataforma.

Habiendo definido los indicadores de interés, se muestran en las **tablas 2 y 3** las posibles variables de entrada, cuya variación podría afectar o no a los indicadores de interés. Estas variables son extraídas de la base de datos de la plataforma, de las tablas de usuarios, videos y de sus interacciones)

Tabla 2: Variables de entrada, caso usuarios

Variable	Descripción
puntos_historicos	Variable que hace referencia a la cantidad de puntos totales que ha acumulado un usuario durante su actividad.
genero	Sexo del usuario.
puntos_gastados	Cantidad total de puntos que el usuario ha gastado en la plataforma.
shares_totales	Cantidad total de veces que el usuario compartió algún video a través de la plataforma.
recruitments	Cantidad de usuarios que fueron reclutados por el usuario en cuestión.
fecha_afiliacion	Fecha en la que un usuario se registró en la plataforma.
uni	Universidad en la que estudia (o estudió) un usuario.
nacimiento	Fecha de nacimiento de un usuario.
puntos_gastados	Cantidad de puntos gastados del usuario.
categoria_dominante	Categoría de videos preferida por usuario (en base a su interacción)
concursos_participados	Cantidad de concursos (diferentes) en los que participó un usuario.
premios_canjeados	Cantidad de premios canjeados por un usuario.
tickets_canjeados	Cantidad de tickets canjeados por un usuario.
difference_last_and_first_share	Diferencia de tiempo entre la primera y la última vez que un usuario compartió un video.
difference_last_raffle_first_share	Diferencia de tiempo entre la primera vez que un usuario compartió un video y la última vez que canjeó un ticket.

Tabla 3: Variables de entrada, caso videos.

Variable	Descripción
duración	Duración en segundos del video en cuestión
release_difference	Diferencia de lanzamiento entre el video en su fuente original (*youtube*) y su publicación en la plataforma.
total_views	Cantidad total de vistas que el video en cuestión consiguió durante el período de estudio.
shares_first_day	Cantidad de veces que un video fue compartido durante el primer día de su publicación.
shares_first_week	Cantidad de veces que un video fue compartido durante la primera semana de su publicación.
shares_first_month	Cantidad de veces que un video fue compartido durante el primer mes de su publicación.
total_shares	Cantidad total de veces que un video fue compartido durante el período de estudio.
active_raffles	Cantidad de concursos al momento de publicación del video en cuestión.
active_canjes	Cantidad de canjes al momento de publicación del video en cuestión.

3.2. Entendimiento de los datos

En esta sección se revisará cómo se distribuyen y relacionan en una primera instancia las variables mencionadas en el punto anterior. Se presentarán gráficos y descripciones sólo de variables relevantes para el estudio, es decir, que aporten a los objetivos planteados o sean razón de toma de decisiones.

Los datos de la plataforma estudiada se encuentran en una base de datos relacional detrás de un motor *MySQL*, cuyo modelo de datos parcial se puede apreciar en el **anexo 1**. Cabe destacar que se incluyeron sólo las tablas y variables más significativas en el diagrama.

La ventana de tiempo contemplada en el estudio es desde el 27 de junio 2013[20], hasta el 9 de enero 2015[21]. Se cuenta para este período con 2669 registros de usuarios y 921 registros de videos.

[20] Fecha de inicio de actividades de la plataforma.
[21] Fecha en que se comienza a realizar este estudio.

Una primera mirada a los datos deja en evidencia los siguientes problemas:

- Para el caso de datos de usuario, la variable **uni** (*universidad_id* en tabla *users*, referencia a la universidad en la que estudia o estudió el usuario en cuestión) tiene una gran cantidad de datos perdidos (nulos o vacíos); aproximadamente un 85% de los registros tiene falencias en este aspecto. Se decide no considerar esta variable para el estudio.

- La variable **release_difference** (o diferencia de lanzamiento desde la fecha en que se sube un video en su fuente original, y se publica en la plataforma), junto con otras variables[22] relacionadas con tiempo, se encuentran en unidades poco intuitivas para ser dimensionadas por el estudio. Se cambiarán estas variables a unidades más adecuadas en cada caso.

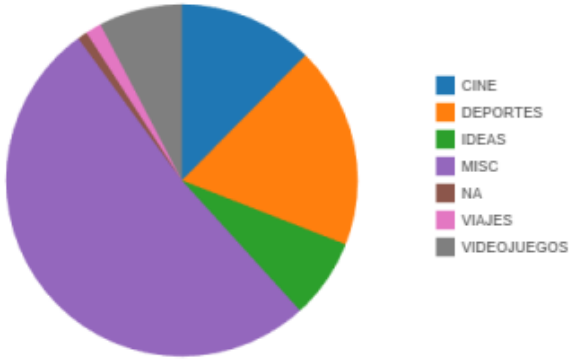
- Es necesario cambiar la fecha de nacimiento de los usuarios por una variable más comparable, o sea, por la edad de los mismos al momento del estudio.

Visualización de datos usando *Tableau* (OLAP)

A continuación, se revisará en primera instancia cómo se distribuyen y relacionan las variables de la plataforma. Para estos fines se utilizó la herramienta *Tableau*, principalmente por su variedad de posibilidades, facilidad de uso, y familiarización con quien realiza el estudio.

[22] Otras variables: *difference_last_and_first_share, difference_last_raffle_first_share*

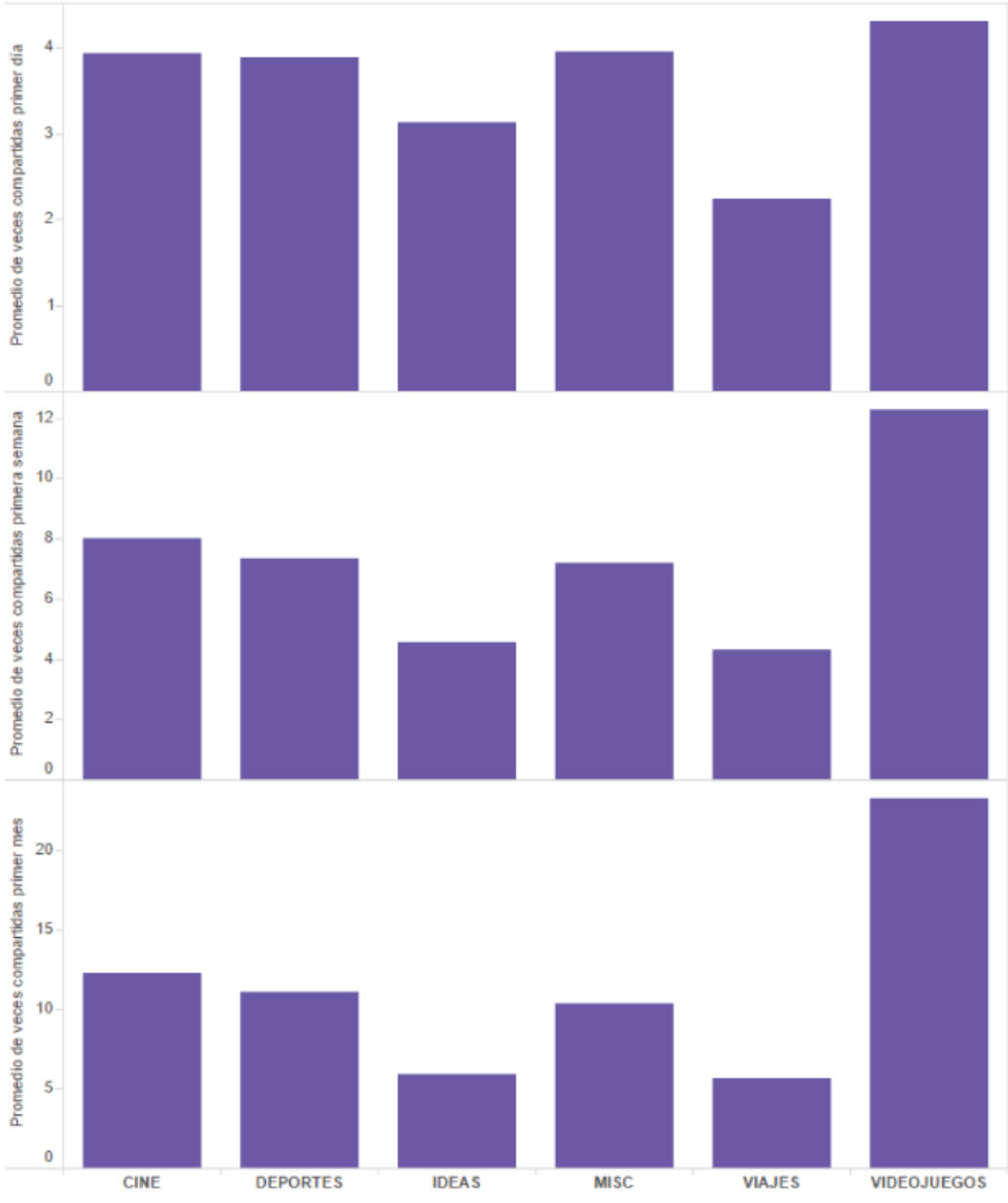
Ilustración 3: Distribución categorías de videos



En la **ilustración 3** se muestra la distribución de videos de acuerdo a su categoría. Aproximadamente un 50% de ellos pertenecen a *MISC*, que es una categoría relativamente general para definir a todos los videos que no pertenecen al resto de los grupos, lo que hace que le resta valor a esta variable en particular. En las siguientes ilustraciones se revisará cómo se comporta esta variable en relación a otras.

En las **ilustraciones 4a, 4b y 4c** se muestra cómo se relaciona la categoría de un video con la cantidad de veces que este se comparte el primer día, la primera semana, y el primer mes (luego de su publicación). Para el primer caso (4a), no se aprecia una diferencia muy significativa. Dejando de lado *viajes*, el resto parece comportarse de forma similar. Pero cuando se revisan los siguientes gráficos, que relacionan las categorías con las veces que se comparte en la primera semana (4b) y mes (4c) respectivamente, se aprecia que *videojuegos* tiende a mantener su popularidad, mientras que el resto de las categorías parecen dejar de ser interesantes para el usuario.

Ilustraciones 4a, 4b y 4c: Categoría de video vs promedio de veces que se comparte en intervalos



4a

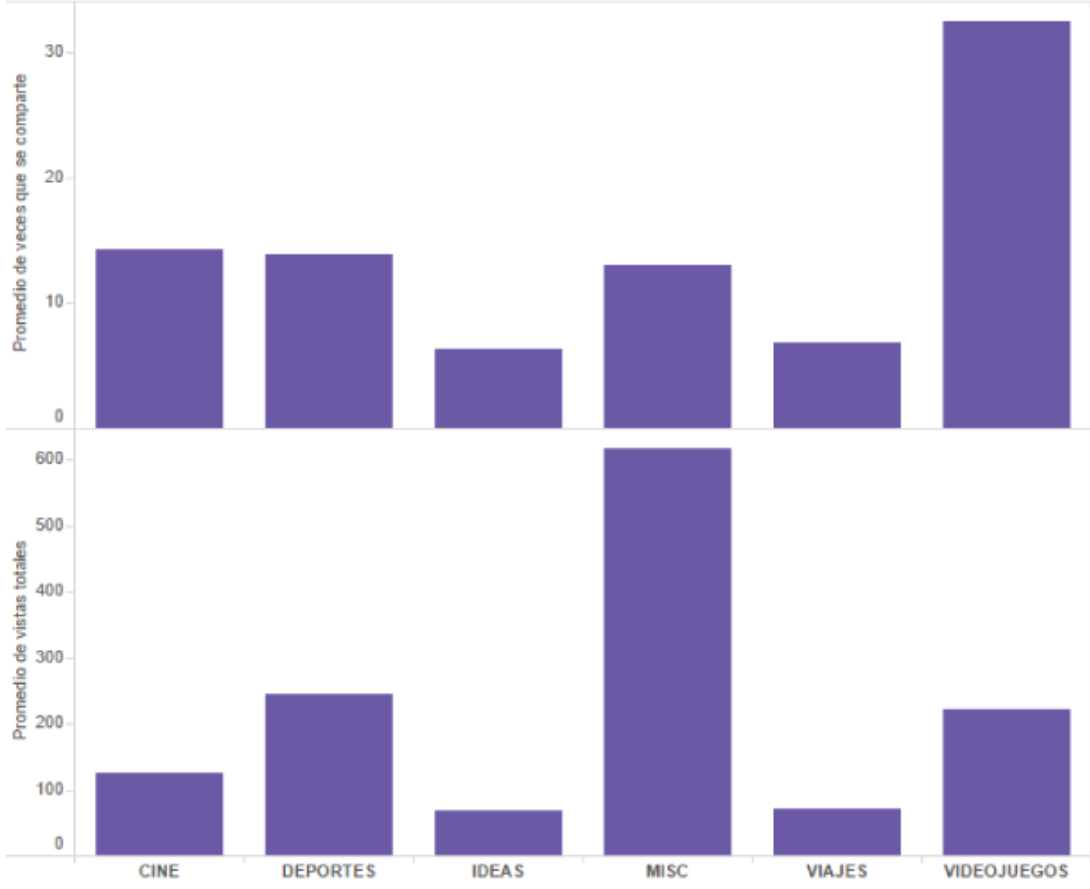
4b

4c

En base a esta potencial relación, se decide estudiar cómo se comportan cada una de las categorías a largo plazo en función de las veces que se comparten. Además, se decide estudiar si este comportamiento se ve reflejado o no en una mayor cantidad de vistas para dichas categorías al largo plazo. El resultado se aprecia en la **ilustración 5**, donde se confirma que la teoría es correcta y que en el tiempo la

categoría de *videojuegos* pareciera seguir siendo la más compartida. Por otro lado, al revisar la relación con el promedio del total de vistas, la categoría *misc* parece ser la que más destaca. Una mirada a los datos detrás de este gráfico permite ver que hay *outliers*[23] presentes en esta categoría en particular. Se prueba quitando estos datos para ver las nuevas proporciones, pero la tendencia se mantiene (aunque un poco menos marcada). Finalmente se concluye que *misc* parece ser la categoría más vista dentro de las planteadas, aunque no la más compartida.

Ilustración 5: Categoría de video vs Promedio de vistas y veces que se comparte

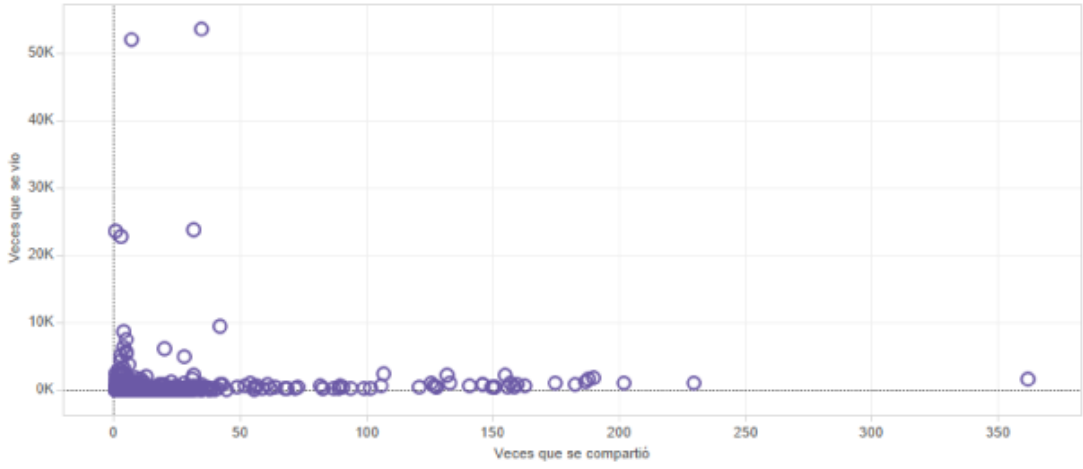


En base al comportamiento indicado por las ilustraciones 4a, 4b y 4c; se decide revisar si existe alguna relación entre la cantidad de vistas que consigue un

[23] En este caso, un video con más de 50.000 vistas, y tres videos con más de 20.000.

video y la cantidad de veces que se comparte. Esta relación se aprecia en la **ilustración 6**, donde se ve que una gran cantidad de los videos se encuentran en el cuadrante de menores valores del gráfico, por lo que se decide estudiar en detalle el cuadrante definido por videos con hasta 10.000 vistas, y hasta 200 veces compartido. Este cuadrante se aprecia en la **ilustración 7** donde, nuevamente, no se aprecia ningún tipo de tendencia o relación entre la cantidad de veces que se comparte un video y la cantidad de vistas que consigue. Iteraciones siguientes de filtrado no reflejaron cambios significativos, por lo que se decide no incluir los gráficos correspondientes.

Ilustración 6: Veces que se comparte vs Veces que se ve



Para la base de datos de usuarios, se presenta en la **ilustración 8** la distribución de las observaciones en relación a la edad. Se deduce de este gráfico que la mayor cantidad de usuarios de la plataforma tiene entre 20 y 26 años. También se aprecia la presencia de muy pocos usuarios sobre 30 años de edad.

Ilustración 7: Veces que se comparte vs veces que se ve (filtrado)

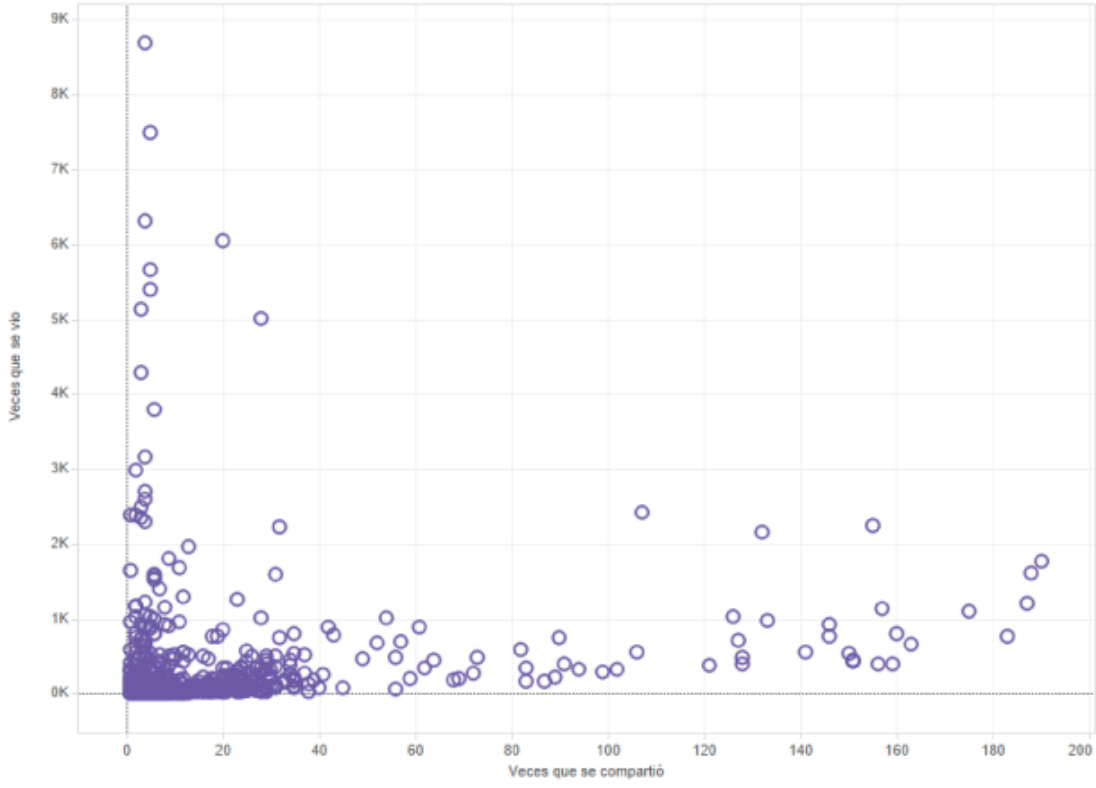
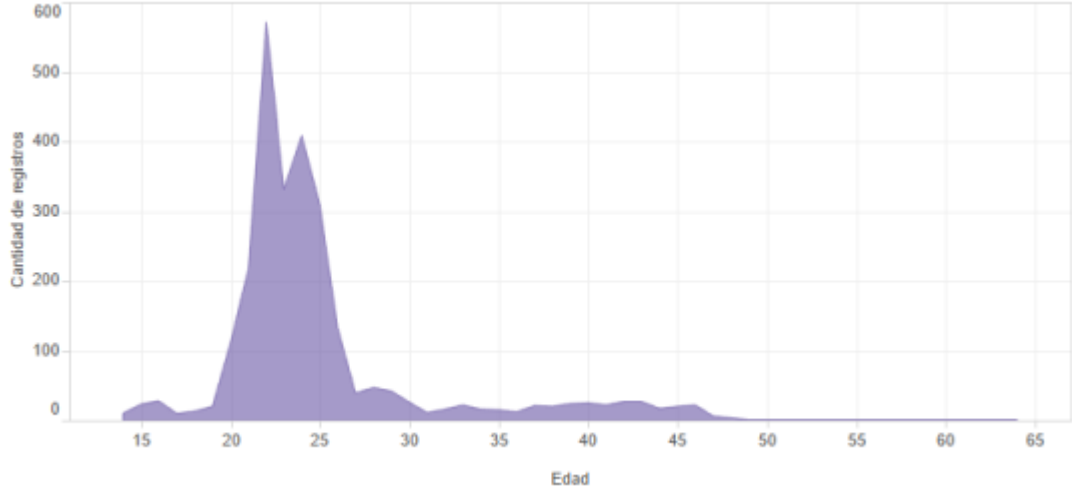


Ilustración 8: Distribución etaria de grupo usuario



A modo de continuación del análisis en torno a la edad de los usuarios de la plataforma, en la **ilustración 9** se aprecia la relación entre esta variable y la cantidad total de veces que un usuario compartió videos. Rápidamente salta a la vista la presencia de *outliers*: una observación de 31 años de edad con más de 500 videos compartidos, y registros con edades superiores a 45 años de edad con muy pocos videos compartidos. En la **ilustración 10** se presenta nuevamente la relación mencionada, esta vez sin los *outliers* identificados.

Ilustración 9: Edad vs Veces que compartió

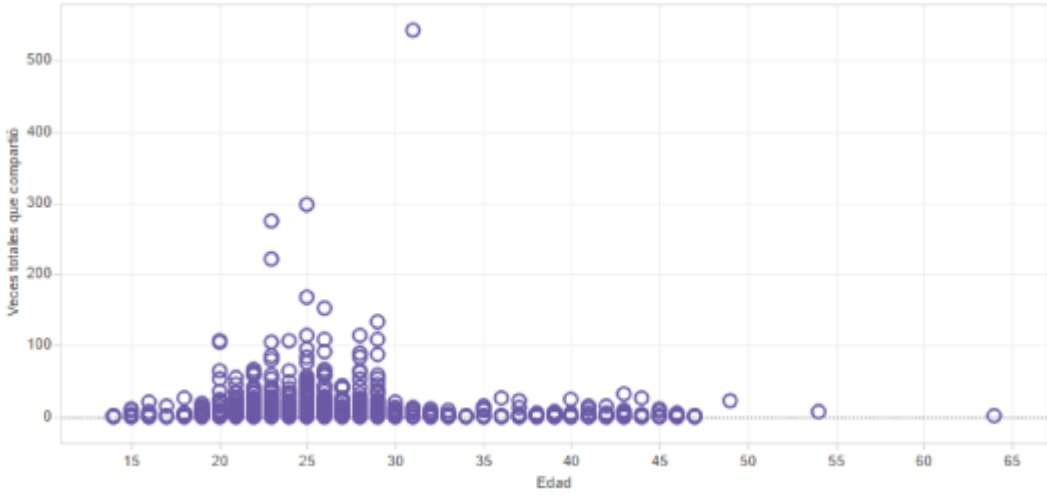
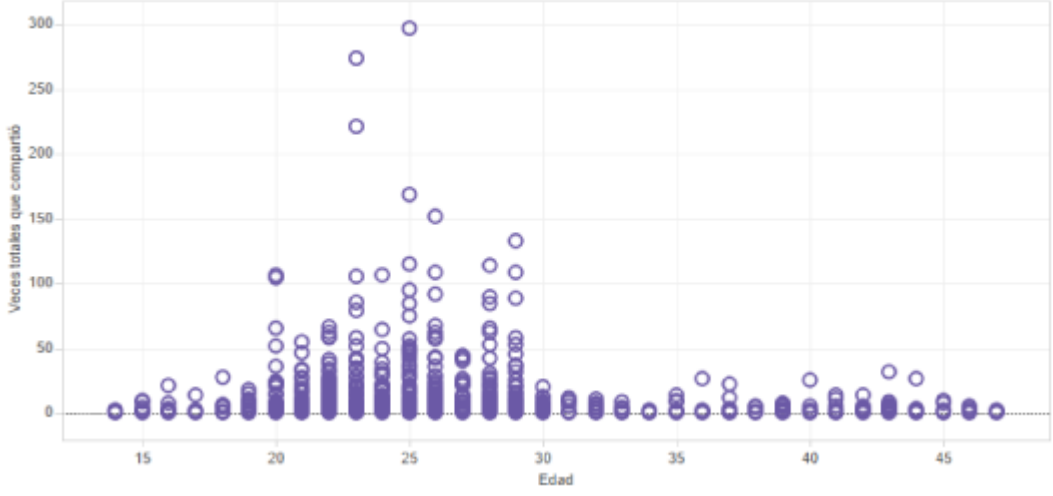


Ilustración 10: Edad vs Veces que compartió (filtrado)



En la ilustración 10 aún se aprecian registros que se escapan de la norma, pero se puede apreciar que el grupo etario más activo en relación a compartir videos se encuentra entre los 20 y los 29 años.

Como último análisis en torno a la edad del grupo usuario de Kikvi, se presenta en la **ilustración 11** la relación entre esta variable y la frecuencia promedio en días con la que los usuarios de dicha edad comparten contenido de la plataforma. Para fines del negocio, un valor menor de frecuencia es mejor, ay que significa que el usuario comparte más veces en un período de tiempo definido. Esto genera un problema ya que hay una gran cantidad de registros que tienen frecuencia con valor 0, que en la práctica quiere decir que no compartieron contenido de la plataforma en ningún momento, pero que en el gráfico (y al promediarse con otros valores) se verá reflejado como algo positivo (disminuyendo la frecuencia promedio en la edad a la que pertenezcan); por esta razón se decide filtrar los registros cuya frecuencia sea igual a 0. El resultado de este proceso se aprecia en la **ilustración 12**, de donde se puede deducir que los usuarios que comparten más frecuentemente se encuentran entre las 14 y los 28 años de edad.

Ilustración 11: Edad vs Frecuencia en que comparte

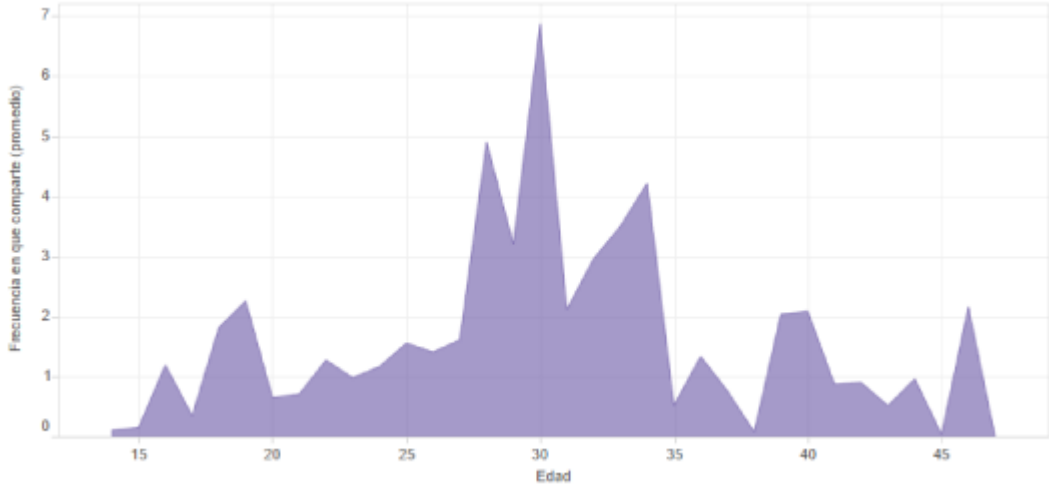
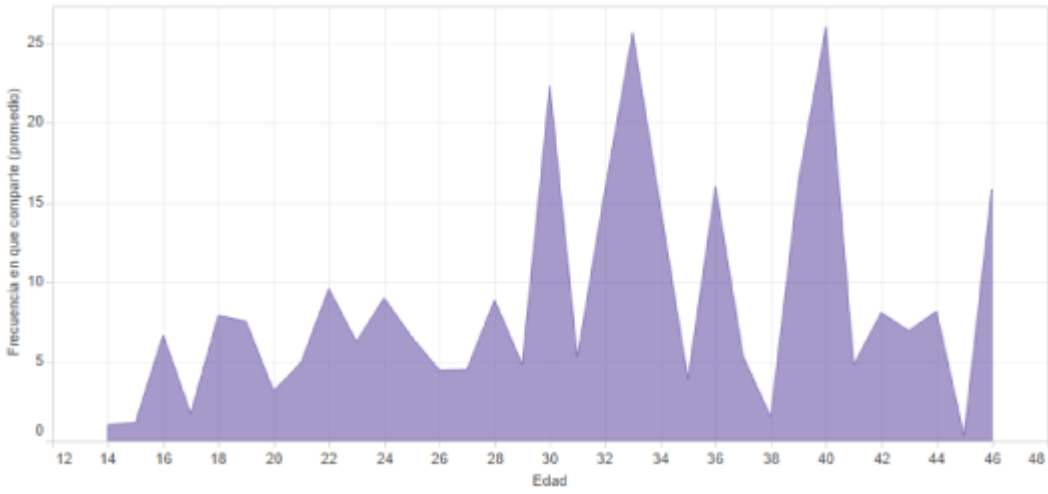


Ilustración 12: Edad vs Frecuencia en que comparte (filtrado)



4. Desarrollo de la solución

En este capítulo se abordará el resto de las etapas pertenecientes al proceso *CRISP-DM*. En una primera instancia se aplicarán acciones sobre los datos, con el fin de adecuarlos para procesos posteriores. Finalmente se modelarán, evaluarán y desplegarán los resultados obtenidos.

4.1. Preparación de los datos

Durante esta etapa se realizó una serie de operaciones sobre los datos, con el fin de mitigar por un lado la falta de datos mencionada en el punto anterior, y por otro lado descubrir variables que no se encuentran inicialmente en el set de datos, pero que pueden ser calculadas en base a la existentes.

Con el fin de mitigar la ausencia de datos de usuarios y de videos que podrían ser interesantes para el estudio, se utilizaron las herramientas (*API*) de *Facebook* y de *YouTube* para buscar información faltante de los usuarios y videos respectivamente. El proceso se automatizó utilizando los *scripts* en *php* y *ruby on rails* presentes en los **anexos 2 y 3** respectivamente.

Para poder satisfacer necesidades del negocio tratadas en puntos anteriores, se hace necesario definir una nueva variable, **calidad usuaria**; hace referencia a qué tan bueno es un usuario para el objetivo final de la plataforma, y qué tanto aporta este mismo en el cumplimiento de este objetivo. Teniendo como consideración que el enfoque actual de Kikvi es la difusión de contenidos, se definen 8 niveles de calidad usuaria, usando como parámetros de entrada la frecuencia en la que el individuo comparte, la cantidad de veces que comparte, y el período total de actividad del

mismo. En la **tabla 4** se muestran los criterios para asignación de clases de esta variable.

Tabla 4: Calidad usuaria

Calidad	Descripción
No interesado, No entendió	Usuario se registró en la plataforma, pero nunca tuvo interacción[24] con la misma
No capturado	Usuario se registró e interactuó en la plataforma durante un único día.
Perdido	Usuario se registró en la plataforma e interactuó con ella por un período menor a una semana.
Diario semanal	Usuario se registró en la plataforma e interactuó con ella diariamente, por un período de entre 7 y 29 días.
Diario mensual	Usuario se registró en la plataforma e interactuó con ella diariamente, por un período mayor a 30 días.
Diario constante	Usuario se registró en la plataforma e interactuó con ella diariamente, por un período mayor a 60 días.
Semanal mensual	Usuario se registró en la plataforma e interactuó con ella semanalmente (al menos), por un período de entre 1 y 2 meses.
Semanal constante	Usuario se registró en la plataforma e interactuó con ella semanalmente (al menos), por un período mayor a 2 meses.

Habiéndose definido el primero de los indicadores clave para el negocio mencionados en la **tabla 1**, en la **tabla 5** se explica cómo fueron calculados los dos restantes.

Tabla 5: Cálculo de indicadores clave

Variable	Descripción y cálculo
Usuarios activos	Cantidad de usuarios activos al momento en que un video fue publicado en la plataforma. Se define un usuario activo como un usuario que interactuó con la plataforma durante una semana definida [9]. Para este caso en particular, una interacción se entiende como la acción de compartir un video, realizar un canje, reclutar otro usuario, ver un video o participar en un concurso. Entonces, en un momento dado, la cantidad de usuarios activos estará dada por el número de usuarios que interactuaron con la plataforma en los últimos 7 días.
Penetración	Porcentaje de usuarios activos que compartieron el video en cuestión. Para este cálculo, se utilizó la variable active_users en el tiempo, y se comparó con la cantidad de veces que se compartió un video en el mismo intervalo.

[24] Se define interacción como cualquiera de las siguientes acciones: compartir un video, visualizar un video, participar en un concurso, realizar un canje de algún producto.

Para poder tener un mejor entendimiento del grupo usuario, se definió y/o calculó una serie de variables que podrían ser significativas para explicar el comportamiento y calidad usuaria, listadas en la **tabla 6**.

Tabla 6: Nuevas variables para usuarios

Variable	Descripción
edad	Edad (en años) del usuario. Esta variable se encontraba presente en el grupo de datos como fecha de nacimiento. Posibles valores: Número entero, entre 14 y 64.
sistema_registro	Hace referencia al sistema a través del cual se registró el usuario Posibles valores: *Facebook campaign*: hace referencia a los usuarios que se registraron al sitio a través de una campaña en Facebook. Cabe destacar que no se trata del uso de "Facebook Ads", si no que de un post/campaña a través de la Fan Page de Kikvi. *Normal sign in*: usuarios que se registraron al sitio sin ninguna referencia registrada. En otras palabras, llegaron al sitio y se registraron haciendo *click* en el botón "Registrarse". *Physical campaign – fair*: registros durante la ejecución de la feria de software en la que la empresa participó en sus inicios, bajo otro nombre. *Physical campaign – flyers*: hace referencia a los usuarios que se registraron luego de llegar al sitio a través de una campaña física de *flyers* realizada en universidades de Santiago. *Recruited*: usuarios adquiridos a través del sistema de reclutamiento establecido en el sitio[25]. *Through video display*: registros a través de links dispuestos en la "vitrina"[26]. Esta es la clase más común para esta variable.
calidad_videos	Calidad de videos en el tiempo comprendido entre una semana antes del registro del usuario y una semana después. Posibles valores: "*High*": alta calidad de videos "*Somewhat high*": calidad de videos relativamente buena "*Regular*": videos regulares "*Somewhat low*": videos de relativamente baja calidad "*Low*": videos de baja calidad

[25] El sistema de reclutamiento funciona de la siguiente manera: Un usuario puede motivar a sus contactos a registrarse en la plataforma a través de un *link* personal. Esto tiene como consecuencia que, una vez que la persona reclutada junte 300 puntos o más, se regalan al reclutador 300 puntos.
[26] Vitrina es el nombre utilizado para referirse a la página en la que se muestra un video.

densidad_concursos	Considera la densidad de concursos de las dos semanas siguientes al registro de un usuario. Posibles valores: "*High*": alta calidad de videos "*Somewhat high*": calidad de videos relativamente buena "*Regular*": videos regulares "*Somewhat low*": videos de relativamente baja calidad "*Low*": videos de baja calidad
densidad_videos	Variable que hace referencia a la cantidad de videos en las dos semanas siguientes al registro de un usuario. Posibles valores: "*High*": alta calidad de videos "*Somewhat high*": calidad de videos relativamente buena "*Regular*": videos regulares "*Somewhat low*": videos de relativamente baja calidad "*Low*": videos de baja calidad

Además, se cambiaron las unidades de una serie de variables con el fin de ser más útiles dentro de su contexto al momento de visualizar los datos.

Durante la etapa de modelado se aplicarán diferentes algoritmos de minería de datos. Algunos de estos algoritmos requieren que las variables de ingreso sean exclusivamente discretas, por lo que, en algunos casos, se hace necesaria la categorización de ellas. Para conseguir este objetivo, se utiliza la función *discreize* del paquete de *arules.* Se decide utilizar esta función ya que tiene la posibilidad de no dividir la información en rangos fijos, si no que considerando su posición en la escala a la que corresponde y separándolos en *clusters*. Por un lado, esto significa que algunos de los rangos generados podrían estar considerablemente mejor representados que otros, pero además tiene como consecuencia que los valores que se escapan mucho del resto de los datos quedan en sus propios rangos (debido a la separación en *clusters*, lo que podría servir para explicar o descartar de mejor manera relaciones encontradas. En las **tablas 7 y 8** se presentan las variables categorizadas, con sus valores posibles.

Tabla 7: Variables categorizadas para usuarios

Variable	Cambio aplicado	Posibles valores
puntos_historicos	Se categoriza en diferentes rangos.	*"[0, 4651)"* *"[4651, 18212)"* *"[18212, 46406)"* *"[46406,118093)"* *"[118093,299581]"*
shares_totales	Se categoriza en diferentes rangos.	*"[0.0, 17.0)"* *"[17.0, 68.4)"* *"[68.4,184.3)"* *"[184.3,403.0)"* *"[403.0,542.0]"*
recruitments	Debido a los bajos valores de usuarios reclutados, se decide utilizar esta variable como un indicador binario.	*"1"*: El usuario reclutó 1 o más usuarios. *"0"*: El usuario no reclutó a otros usuarios.
concursos_participados	Se categoriza en diferentes rangos.	*"[0.00, 1.21)"* *"[1.21, 3.37)"* *"[3.37, 5.56)"* *"[5.56, 8.72)"* *"[8.72,16.00]"*
premios_canjeados	De la misma forma que para la variable **recruitments**, se decide utilizar como un indicador binario.	*"1"*: El usuario canjeó 1 o más premios. *"0"*: El usuario no canjeó ningún premio.
tickets_canjeados	Se categoriza en diferentes rangos.	*"[0.00, 7.37)"* *"[7.37, 27.94)"* *"[27.94, 66.36)"* *"[66.36,209.18)"* *"[209.18,459.00]"*
edad	Se categoriza en diferentes rangos.	*"[14.0,21.0)"* *"[21.0,23.6)"* *"[23.6,27.8)"* *"[27.8,36.3)"* *"[36.3,64.0]"*

Tabla 8: Variables categorizadas para videos

Variable	Cambio aplicado	Posibles valores
active_raffles	Se categoriza en diferentes rangos.	*"2 o menos"* *"Entre 3 y 4 "* *"Entre 5 y 6"* *"Entre 7 y 9"* *"10 o más"*
active_users	Se categoriza en diferentes rangos.	*"[0, 40]"* *"(40, 80]"* *"(80, 120]"* *"(120, 160]"* *"Más de 160"*
duracion	Se categoriza en diferentes	*"30 o menos"*

	rangos.	*"(30, 60]"* *"(60, 120]"* *"(120, 180]"* *"(180, 240]"* *"(240, 300]"* *"(300, 600]"* *"600 o más"*
release_difference	Se categoriza en diferentes rangos.	*"Menos de 6 horas"* *"Entre 6 horas y 1 día"* *"Entre 1 y 3 días"* *"Entre 3 días y 1 semana"* *"Entre 1 y 2 semanas"* *"Más de 2 semanas"*
penetracion	Se categoriza en diferentes rangos.	*"Menos del 10%"* *"Entre 10% y 20%"* *"Entre 20% y 30%"* *"Entre 30% y 40%"* *"Entre 40% y 50%"* *"Entre 50% y 60%"* *"Entre 60% y 70%"* *"Entre 70% y 80%"* *"Entre 80% y 90%"* *"Más del 90%"*
total_views	Se categoriza en diferentes rangos.	*"[1, 386)"* *"[386, 1484)"* *"[1484, 4417)"* *"[4417,20842)"* *"[20842,53557]"*
shares_first_day	Se categoriza en diferentes rangos.	*"[0.00, 4.13)"* *"[4.13, 8.14)"* *"[8.14,12.62)"* *"[12.62,19.38)"* *"[19.38,31.00]"*
shares_first_week	Se categoriza en diferentes rangos.	*"[1.00, 7.08)"* *"[7.08,16.19)"* *"[16.19,29.44)"* *"[29.44,47.96)"* *"[47.96,76.00]"*
shares_first_month	Se categoriza en diferentes rangos.	*"[1.00, 5.98)"* *"[5.98, 15.70)"* *"[15.70, 33.73)"* *"[33.73, 70.85)"* *"[70.85,173.00]"*
total_shares	Se categoriza en diferentes rangos.	*"[1.00, 7.52)"* *"[7.52, 19.94)"* *"[19.94, 54.34)"* *"[54.34,123.42)"* *"[123.42,362.00]"*
active_canjes	Se categoriza en diferentes rangos.	*"0.000"* *"[0.693,2.193)"* *"[2.193,3.612)"* *"[3.612,5.112)"* *"[5.112,6.000]"*

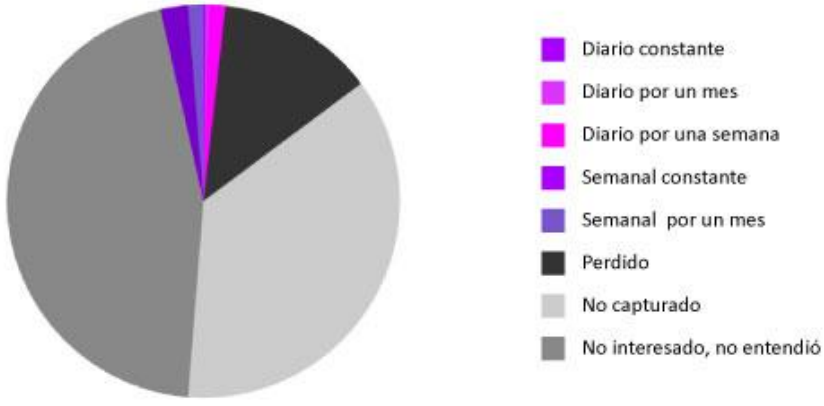
4.2. Modelado

En esta etapa de *CRISP-DM* se procede a aplicar algoritmos y modelos ya existentes y conocidos para extraer información que pueda ser valiosa para el negocio a partir de los datos tratados en el punto anterior. Además, se utilizan herramientas de procesamiento analítico para generar gráficos que pudiesen aportar valor al estudio.

Análisis de datos tratados, usando Tableau (OLAP)

En esta sección, se presentan algunos gráficos de interés generados con esta herramienta, que a primera vista parecen ser importantes para conocer el negocio. Los gráficos mencionados buscan entender de mejor manera como se distribuyen y comportan las variables mencionadas en la **tabla 1.**

Ilustración 13: Distribución de la calidad usuaria



En la **ilustración 13** se muestra cómo se distribuyen los registros de usuarios en referencia a su calidad. Para facilitar su entendimiento, se han dispuesto tonalidades de morado para las calidades usuarias consideradas positivas, y

tonalidades de gris para las que son consideradas negativas. Se puede apreciar en la ilustración mencionada que la proporción de usuarios de calidad positiva respecto de calidad negativa es preocupante[27]. Esta distribución extremadamente dispareja también significará que cualquier regla de asociación que se pueda extraer de los datos que haga referencia a calidades positivas de usuarios, tendrá un soporte extremadamente bajo (independiente de la confianza que tenga).

Para poder utilizar técnicas de clasificación sobre este indicador se hace necesario hacer *undersampling*[28] o en su defecto *oversampling*[29] de los datos. Se decide proceder con este último ya que para algunos escenarios la cantidad de registros es muy pequeña, y una submuestra no generará suficientes datos para el estudio. Como criterio de creación de datos, se utilizan todos los valores existentes en registros conocidos de cada clase, y se asigna aleatoriamente uno de ellos al nuevo registro.

Es interesante en el desarrollo de este estudio considerar las variables del ambiente (de la plataforma) en el momento en que se registran los usuarios, para así ver si hay alguna relación, patrón o tendencia marcada para cualquiera de las clases. A continuación, se presenta la relación de dichas variables con indicadores de interés relacionados.

En la **ilustración 14** se compara la calidad usuaria con la calidad de los videos (recientes) al momento que dichos usuarios se registraron. En la tabla, un color más intenso hace referencia a una mayor cantidad de registros, lo que además se puede apreciar por el número en cada celda. Debido a la baja cantidad de registros para las calidades positivas, no es posible inferir a simple vista si hay alguna clase de patrón

[27] De forma más exacta, hay 2527 (94.68%) de registros de usuario negativos contra 142 (5.32%) positivos.
[28] Submuestrear, consiste en el proceso de eliminar registros del set de datos con el fin de que todas las clases objetivo tengan una cantidad similar de ocurrencias.
[29] Sobremuestrear, análogo a submuestrear, consiste en generar datos ficticios, acorde a un criterio definido, con el fin de que todas las clases objetivo tengan una cantidad similar de ocurrencias.

con referencia a la calidad de los videos al momento de registro. Por otro lado, para las calidades negativas (últimas 3 columnas), pareciera haber mayor peso en los valores de calidades negativas (últimas 2 filas).

Ilustración 14: Calidad usuaria vs calidad de videos

Calidad Videos	Diario constante	Diario por una semana	Diario por un mes	Semanal por un mes	Semanal constante	Perdido	No capturado	No interesado, no entendió
High	1			6	7		45	74
Somewhat High		8	2	5	16	18	117	123
Regular		21	4	16	26	7	173	241
Somewhat Low	2	5	1	5	8	20	444	519
Low	1	2	1	2	3	382	194	258

En el caso de los usuarios perdidos (sexta columna), para la gran mayoría de los casos la calidad de los videos al momento de registro es mala (*Low*). Además, para las clases de calidad usuaria "No capturado" y "No interesado, no entendió" (últimas 2 columnas) se ve que los valores más comunes corresponden a las calidades de videos regular, relativamente mala y mala, con mayor peso en relativamente mala.

Ilustración 15: Calidad usuaria vs densidad de concursos

Densidad Concursos	Diario constante	Diario por una semana	Diario por un mes	Semanal por un mes	Semanal constante	Perdido	No capturado	No interesado, no entendió
High	1	6	2	7	7	60	19	260
Somewhat High	1	4	3	6	10	51	21	192
Regular		6		3	11	34	19	107
Somewhat Low	2	11	2	11	14	117	213	364
Low		9	1	7	18	85	701	284

En la **ilustración 15**, siguiendo los niveles y características visuales de la tabla anterior, se presenta la distribución de calidad usuaria en comparación a la densidad

de concursos al momento de registro. Una vez más, no parece haber una tendencia en las calidades de usuario positivas, mientras que en las calidades negativas parece hacer peso por bajas densidades de videos, en especial en el caso de la clase “No capturado”, donde una notable mayoría se registró en un período de baja densidad de concursos (o al menos, relativamente baja). Para el caso de la clase “No interesado, no entendió” la distribución de valores de densidad de concursos parece ser bastante pareja.

Ilustración 16: Calidad usuaria vs densidad de videos

En la **ilustración 16** se aprecia la relación entre la calidad usuaria y la densidad de videos en las 2 semanas siguientes a su registro. De forma diferente a las relaciones que se habían revisado hasta el momento, en este caso parece haber una tendencia en las calidades usuarias positivas. Para las 5 primeras columnas (clases consideradas positivas) pareciera que los valores de densidad de videos se concentran en los 2 valores más altos: Alta y Relativamente alta. Este patrón parece además reflejarse en la calidad usuaria “Perdido”, aunque en este caso también hay peso considerable, e incluso mayor, en los valores más bajos de densidad de videos. Para el caso de usuarios de clase “No capturado” se ve una fuerte tendencia a los valores más bajos de densidad de videos. Por otro lado, para la clase “No interesado, no entendió” no pareciera haber una relación a primera vista.

En la **ilustración 17** se aprecia la relación entre calidad usuaria y sistema de registro. Cabe destacar que, para esta variable, los valores de las clases no tienen ningún tipo de escala. Salta a la vista que para la clase "No interesado, no entendió" el sistema de registro más común es a través de la "vitrina", lo que parece repetirse, acompañado por el sistema de registro normal, para la clase "No capturado". Otra relación que se ve notoriamente es la que comprende a las clases de calidad usuaria "Perdido" y de registro a través de "Reclutamiento". Por otro lado, para las clases positivas de calidad usuaria pareciera haber un peso mayor para los valores de sistema de registro de campaña de Facebook y campaña física de *flyers*, aunque esta tendencia no está muy marcada.

Ilustración 17: Calidad usuaria vs sistema de registro

Sistema Registro	Diario constante	Diario por una semana	Diario por un mes	Semanal por un mes	Semanal constante	Perdido	No capturado	No interesado, no entendió
Facebook campaign	1	15	4	18	43	11	193	28
Normal sign in	1	2	1		2	20	258	78
Physical campaign: fair					1	29	113	48
Physical campaign: flyers	2	16	1	15	13	1	101	15
Recruited		2				273	57	24
Through video display		1	2	1	1	13	251	1 014

Hasta el momento se han comparado diferentes variables con la variable de interés **calidad usuaria**. A continuación, se realizarán comparaciones en búsqueda de patrones que pudiesen influir en las variables **penetración** y **usuarios activos**, respectivamente.

En la **ilustración 18** se aprecia la distribución de valores presentes para penetración. En el eje horizontal se presentan los valores de penetración aproximada a dos decimales, mientras que en el vertical se refleja la cantidad de registros que cuentan con ese valor. A primera vista la variable parece estar bien distribuida en sus valores posibles, teniendo una baja considerable en sus valores más altos (80% de penetración o más).

Ilustración 18: Distribución de videos de acuerdo a su penetración

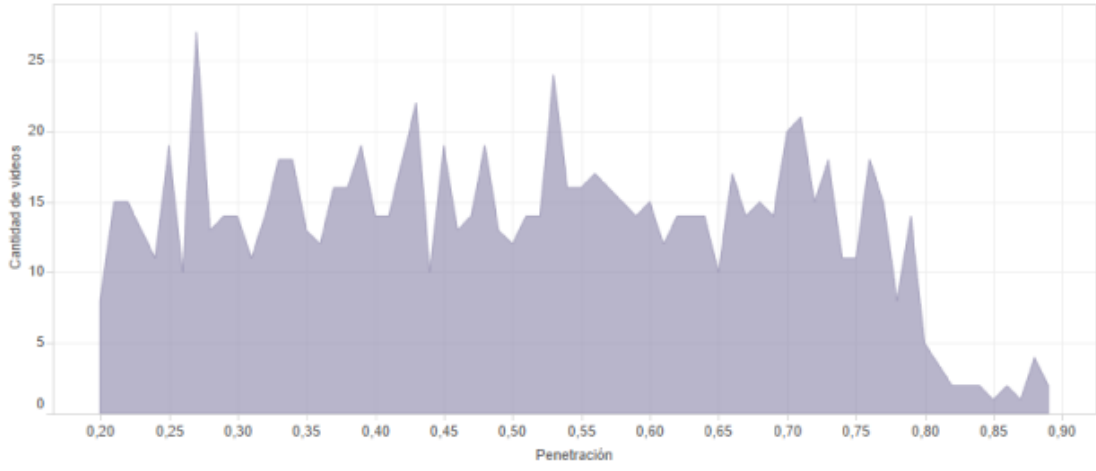
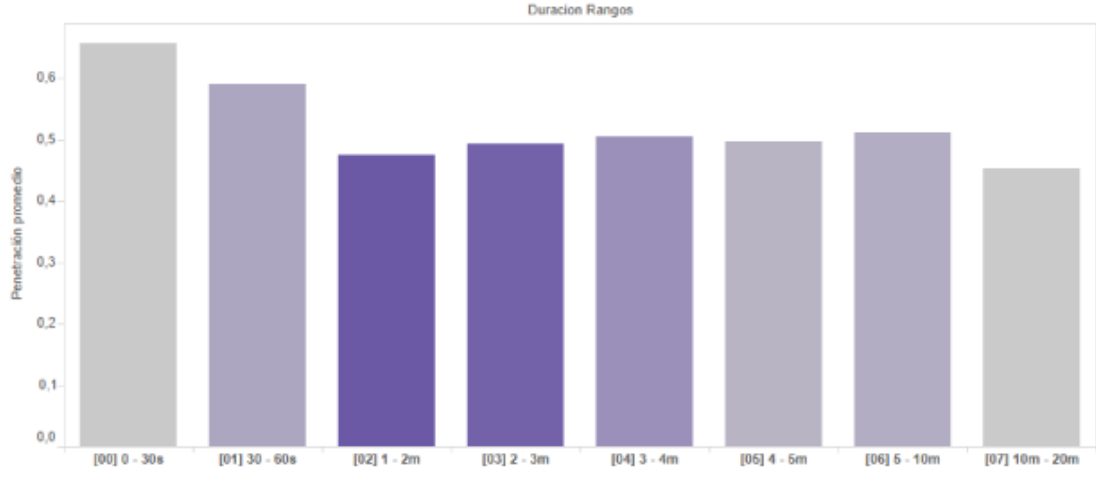


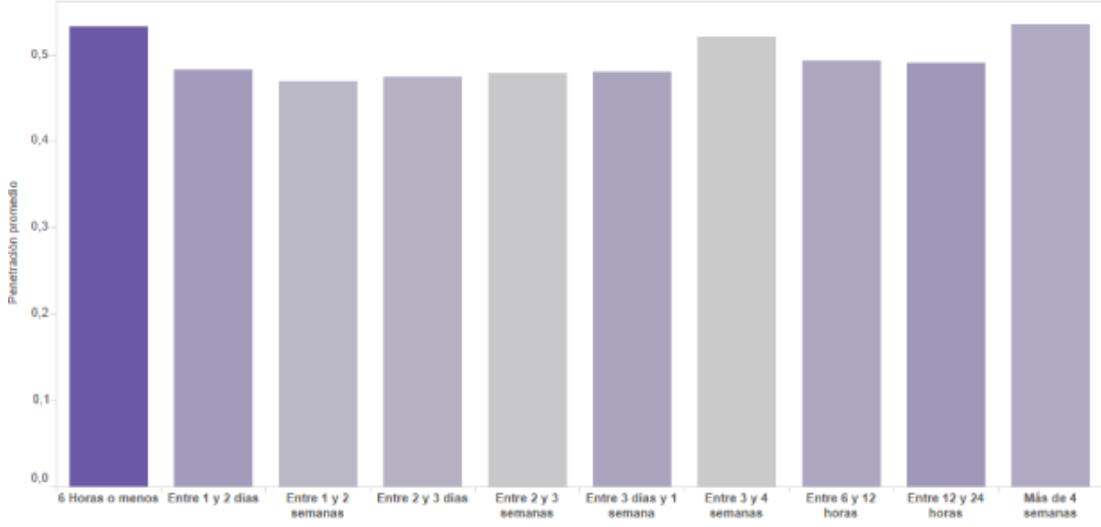
Ilustración 19: Penetración vs duración



En la **ilustración 19** se refleja la relación entre la duración de un video en rangos y la penetración promedio del mismo en ese rango. El color representa la cantidad de registros que ese rango representa (a mayor saturación, mayor cantidad de registros). A pesar de que la tendencia no parece muy marcada, pareciera que un video tiene mejor penetración a menor duración.

Se teoriza que es importante ser de las primeras plataformas en difundir un contenido en particular para conseguir un buen recibimiento del grupo usuario, traducido en una penetración alta. En la **ilustración 20**, se busca entender una posible relación entre la diferencia de lanzamiento[30] y la penetración de un video. El eje horizontal representa la diferencia de lanzamiento como variable independiente, y el vertical la penetración promedio de ese segmento en particular. A simple no pareciera haber una diferencia muy marcada entre los diferentes rangos de lanzamiento, aunque el primer rango se nota considerablemente mejor representado que el resto de los rangos (la opacidad del color del gráfico hace referencia a la cantidad de datos que pertenecen en cada segmento).

Ilustración 20: Penetración contra diferencia de lanzamiento



Como exploración adicional, se decide estudiar conjuntamente la influencia de ambas variables en la penetración.

En la **ilustración 21** se aprecia la exploración propuesta. En una primera interpretación pareciera que los videos en el primer rango de diferencia de

[30] Diferencia de tiempo entre que el contenido es publicado en YouTube en comparación a su publicación en la plataforma.

lanzamiento y primeros rangos de duración alcanzan una mejor penetración. Llama la atención la penetración considerablemente marcada del primer rango de duración (entre 0 y 30 segundos) y la diferencia de lanzamiento *"Entre 1 y 2 semanas"*, pero al revisar los datos detrás de estos rangos se descubre que hay un único registro en ellos, concluyéndose que es probable que se trate de un *outlier*.

Ilustración 21: Relación penetración vs diferencia de lanzamiento y duración

Duracion Rangos	6 Horas o menos	Entre 6 y 12 horas	Entre 12 y 24 horas	Entre 1 y 2 dias	Entre 2 y 3 dias	Entre 3 dias y 1 semana	Entre 1 y 2 semanas	Entre 2 y 3 semanas	Entre 3 y 4 semanas	Más de 4 semanas
[00] 0 - 30s	0,7317	0,5600	0,5650	0,3900		0,6350	0,8000			
[01] 30 - 60s	0,8097	0,3775	0,4750	0,4900	0,5817	0,4354	0,3800	0,5200	0,4900	0,5100
[02] 1 - 2m	0,4793	0,4935	0,4793	0,4463	0,4506	0,4583	0,4633	0,4100	0,6450	0,5158
[03] 2 - 3m	0,5020	0,4333	0,4728	0,5187	0,4645	0,4616	0,4123	0,5550	0,5000	0,5766
[04] 3 - 4m	0,5279	0,5568	0,5016	0,4540	0,4908	0,5119	0,4982	0,5180	0,4900	0,4525
[05] 4 - 5m	0,4878	0,4850	0,5143	0,5067	0,4267	0,6233	0,4600		0,4600	0,5550
[06] 5 - 10m	0,4656	0,5356	0,5287	0,5464	0,4633	0,5500	0,5317	0,2850	0,5800	0,5475
[07] 10m - 20m	0,5060	0,2700		0,4100				0,5100		

A continuación, se busca encontrar alguna relación entre la cantidad de usuarios activos y variables. Para esto, se considera cada publicación de video como un instante de estudio (efectivo debido a que la publicación de videos durante la ventana de tiempo correspondiente a la base de datos de estudio es periódica[31]). De esta base, se ve en las **ilustraciones 22 y 23**, la relación entre la cantidad de usuarios activos promedio para estos instantes de tiempo en comparación a la cantidad de concursos[32] y canjes[33] activos, respectivamente. En ambos casos se puede apreciar

[31] Se publican videos repetidas veces a la semana, sin dejar pasar múltiples días sin nuevos videos (excepto en excepciones pequeñas)

[32] Un concurso es un proceso en el que se pueden cambiar puntos adquiridos en la plataforma por *tickets* que dan la posibilidad de ganar algún premio en particular. Se sortea el premio entre todos los *tickets* comprados para ese evento en particular, lo que significa una mayor probabilidad entre mayor sea la cantidad de tickets canjeados por un usuario. El costo en puntos de un concurso suele ser bajo.

[33] Un canje (o canje directo) es un intercambio de puntos adquiridos en la plataforma por un producto o servicio en particular. El cambio es inmediato y la entrega de productos dentro de una

una posible relación directa entre la cantidad de usuarios activos y las variables en cuestión, siendo aún más notoria esta posible relación para el caso de los concursos activos.

Ilustración 22: Usuarios activos vs concursos activos

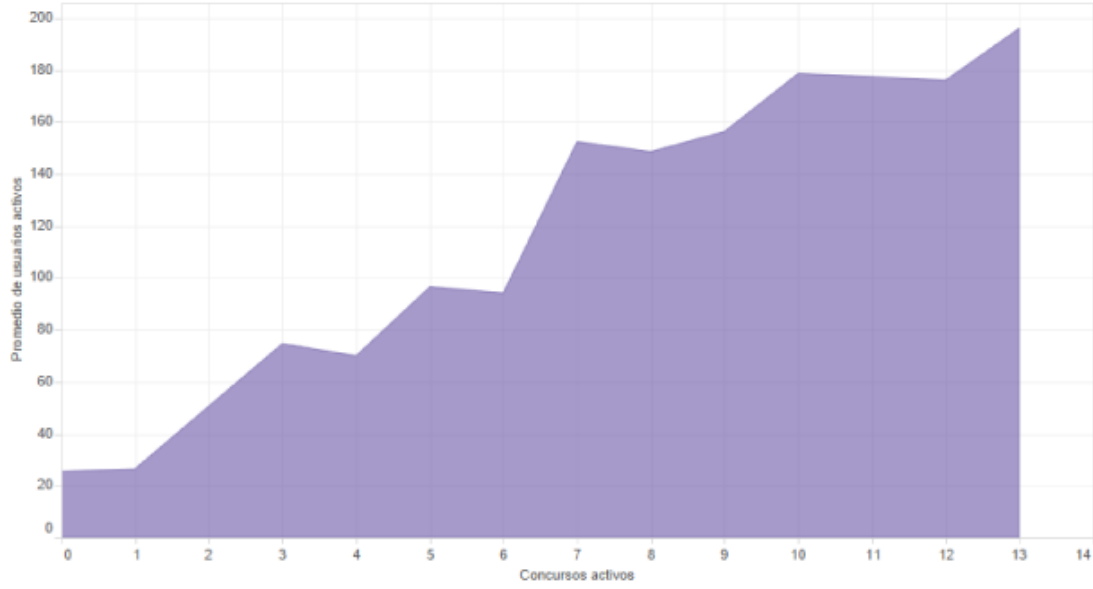
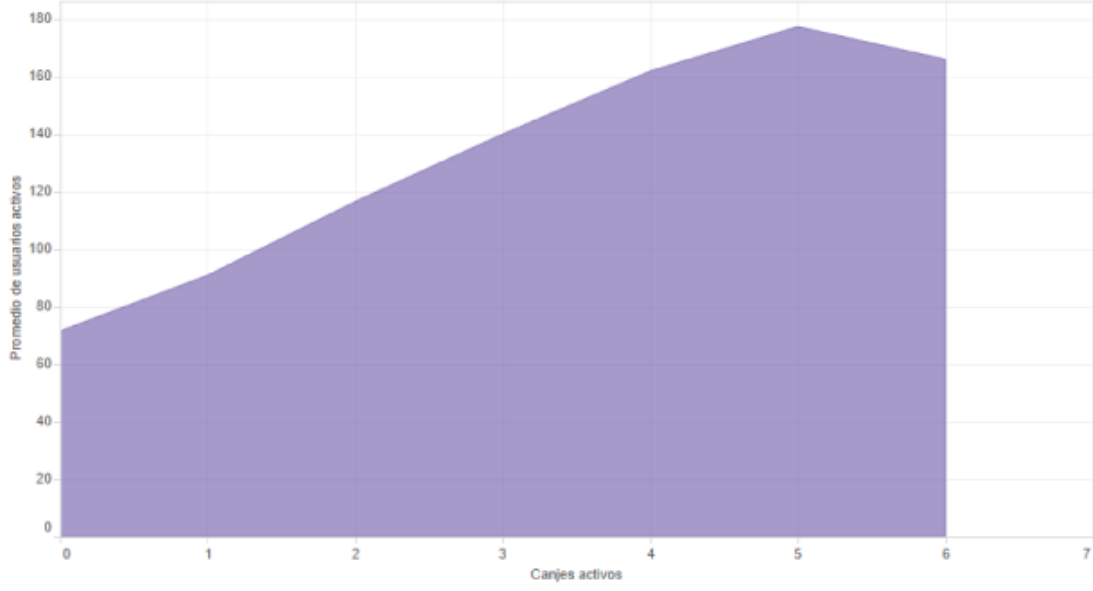


Ilustración 23: Usuarios activos vs canjes activos



semana de tiempo. El costo de un canje es alto en comparación a la ratio de adquisición de puntos posibles en la plataforma, considerablemente más alto que el costo de participar en un concurso.

Reglas de asociación

A modo de primer acercamiento, se decide utilizar *Apriori*, debido a su simpleza, y a que suele dar un buen punto de partida para saber qué variables podrían estar influyendo directamente en los indicadores de interés de este estudio. Como se mencionó en capítulos anteriores, se usa *R* para este objetivo, utilizando la implementación del algoritmo del paquete *arules*.

Ya que las variables fueron discretizadas en el punto anterior, se procede a aplicar el algoritmo mencionado. Para disminuir la cantidad de resultados, se decide utilizar el parámetro *appearance*. Este parámetro limita la cantidad de variables y valores que son considerados en uno o ambos lados de las relaciones encontradas, y así reducir considerablemente las reglas resultantes de la ejecución para facilitar su análisis. Cabe destacar que esta limitante no se realiza durante la ejecución, si no únicamente al momento de mostrar los resultados. En otras palabras, se filtran los resultados automáticamente para conseguir las relaciones que tengan en la mano derecha a las variables de interés mencionadas en capítulos anteriores. La lista de *appearance* se define como:

```
apriori_users_appeareance_list = list(
  rhs = c(
    “<indicador>=<clase 1>“,
    “<indicador>=<clase 2>“,
    …
    “<indicador>=<clase n>“,
  ),
  default = “lhs”
)
```

Esto quiere decir que se mostrarán relaciones que tengan a mano derecha (*rhs: righ hand side*) los diferentes *valores* ***clase*** para el indicador *indicador*, mientras que a mano izquierda (*lhs: left hand side*) se mantendrán todos los resultados encontrados (funcionamiento por defecto). Al aplicar esta limitante con respecto a lo que se busca en la mano derecha de la regla, esta pasa a ser una regla de clasificación.

Como se revisó anteriormente en la **ilustración 8**, sólo el 5.32% de los datos corresponden a registros de calidades usuarias positivas, por lo que ninguna regla de asociación que pudiese encontrarse tendrá mayor soporte a ese valor. Entrando en detalle en estos registros, la cantidad total de datos de cada clase positiva de calidad usuaria se aprecia en la **tabla 9**. De acuerdo a esta información, se sabe de antemano que no se encontrarán reglas con soporte mayor al 2.25%. Se decide finalmente descartar el uso de *apriori* para encontrar reglas de asociación para estas clases.

Tabla 9: Cantidad de registros de calidades usuarias positivas

Clase	Cantidad de registros (porcentaje de la muestra total)
Daily, for a month	8 registros (0.3%)
Daily, for a week	36 registros (1.35%)
Daily, constant	4 registros (0.15%)
Weekly, for a month	34 registros (1.27%)
Weekly, constant	60 registros (2.25%)

Por otro lado, la cantidad que cuentan con calidades usuarias negativas es del 94.68%. En la **tabla 10** se presenta el total de registros de cada clase negativa.

Tabla 10: Cantidad de registros de calidades usuarias negativas

Clase	Cantidad de registros (porcentaje de la muestra total)
Not interested/Didn't get it	1207 registros (45.22%)
Not captured	973 registros (36.46%)
Lost	347 registros (13%)

Ya que la mayoría de las variables hace referencia a la actividad que los usuarios tienen con la plataforma, se espera poca información de las reglas cuya mano derecha contenga el valor *"Not interested/Didn't get it"*, ya que estos usuarios, por definición, no interactuaron con la plataforma. Se decide entonces buscar reglas de asociación que contengan en su mano derecha sólo las clases *"Not captured"* y *"Lost"*. En la **tabla 11** se presentan las reglas encontradas para estas clases; cabe destacar que éstas no son las únicas reglas resultantes de la ejecución del algoritmo, pero el resto de ellas eran combinaciones de las diferentes variables dependientes de la interacción, en las que en todos los casos hacían referencia a los primeros intervalos de valores, lo que es de esperarse de grupos de usuarios no capturados y perdidos.

Tabla 11: Reglas de asociación para clases *Not captured* y *Lost*

Clase	Mano izquierda	Mano derecha	Soporte	Confianza	Lift
Not captured	densidad_videos=Low	quality=Not captured	25.48%	63.31%	1.7
	densidad_concursos=Low	quality=Not captured	26.26%	63.44%	1.7
Lost	sistema_registro=Recruited	quality=Lost	10.29%	76.69%	5.9
	premios_canjeados=0, sistema_registro=Recruited	quality=Lost	9.98%	76.08%	5.9
	recruitments=0, sistema_registro=Recruited	quality=Lost	10.04%	76.35%	5.8
	recruitments=0, premios_canjeados=0, sistema_registro=Recruited	quality=Lost	9.7%	75.73%	5.8

Estas reglas encontradas concuerdan con la información visualizada en la primera etapa del proyecto, por lo que no se descubre nuevo conocimiento a durante este proceso.

Luego se procede a aplicar este algoritmo sobre la base de datos de videos. De la misma forma que para la de usuarios, se busca las reglas cuya mano derecha contenga valores relacionadas con los indicadores de interés. Lamentablemente, no se encuentran reglas que aporten valor a los objetivos del estudio.

Técnicas de clasificación

En segunda instancia, se decide utilizar técnicas de clasificación, para así poder definir un modelo que se adecue de la mejor manera posible a cada uno de los indicadores claves del estudio. Se implementan y comparan los algoritmos: árboles de clasificación[34], *random forest*[35], bayes ingenuo[36] y máquinas de vectores de soporte[37]. Finalmente se compara el desempeño de cada algoritmo utilizando curvas matrices de confusión y curvas ROC.

Durante la generación de clasificadores se decide separar los datos en subgrupos de entrenamiento y prueba: 70% de los datos se utilizó para entrenamiento y un 30% para pruebas en cada uno de los casos.

Clasificación de registros con objetivo de usuarios activos

Como el indicador de usuarios activos es numérico, se hace necesario definir previamente qué valores de éste serán considerados exitosos. Al no existir referencias previas para el dominio del negocio, se decide definir sobre 150 usuarios activos como un caso de éxito. Esta métrica dependerá de la situación actual del negocio.

En la **ilustración 24** se muestra el árbol resultante al utilizar el algoritmo de generación de árboles de clasificación *rpart*. En cada una de las hojas del árbol, el primer número representa la probabilidad de que un registro en ella no sea exitoso, o sea, que tenga menos de 150 usuarios activos. Análogamente, el segundo valor representará la probabilidad de éxito en dicha hoja. Finalmente, el porcentaje en

[34] Paquete *rpart*
[35] Paquete *randomForest*
[36] Paquete *e1071*
[37] Paquete *e1071*

estos nodos hace referencia a la cantidad de datos que este representa. En la **ilustración 25** se muestra gráficamente una explicación de la información presente en las hojas del árbol de decisión.

Ilustración 24: Árbol de clasificación para usuarios activos

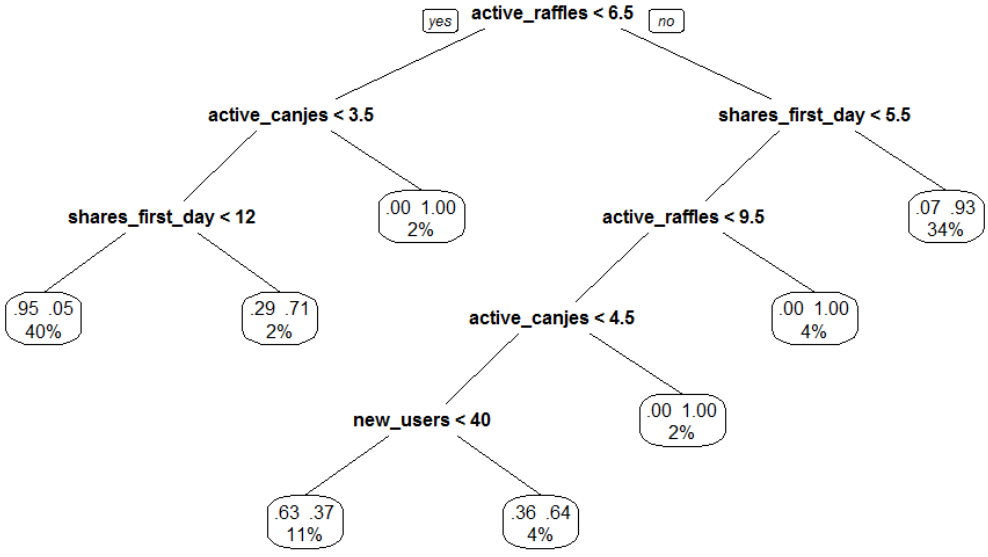
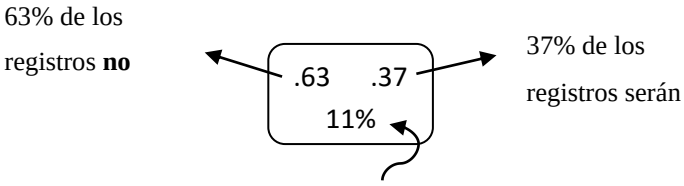


Ilustración 25: Explicación ejemplificada de árbol de decisión



El 11% de los datos cumplen con las condiciones que terminan en esta hoja

De acuerdo al árbol encontrado, las variables *active_raffles*[38], *shares_first_day*[39], *active_canjes*[40] y *new_users*[41] parecen ser decisivas al momento de determinar la cantidad de usuarios activos de la plataforma en un momento dado.

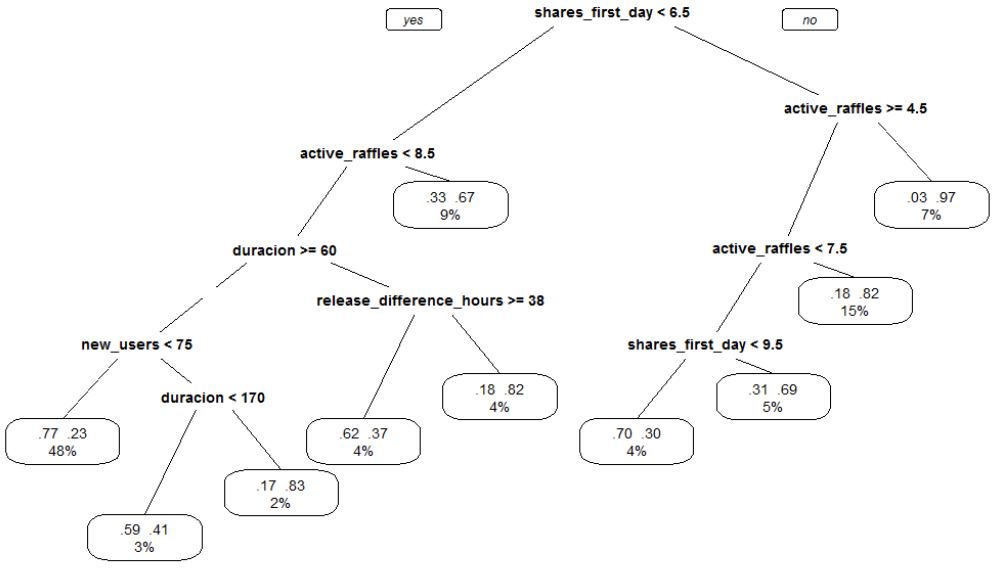
[38] Cantidad de concursos activos
[39] Cantidad de veces que se comparte un video en su primer día
[40] Cantidad de canjes activos

Luego se aplican el resto de los algoritmos mencionados, los que no generan apoyo visual del clasificador, pero cuya efectividad será comparada en el siguiente capítulo.

Clasificación de registros con objetivo de penetración

Continuando con los indicadores de interés, se repite el proceso de generación de clasificadores. Nuevamente se trata de un indicador numérico, por lo que se separa en categorías que serán consideradas provechosas o no para el negocio. En este caso, se define un caso de éxito como un video que alcanza más del 70% de penetración.

Ilustración 26: Árbol de clasificación para penetración



En la **ilustración 26** se presenta el árbol de decisión resultante para el caso del indicador de penetración de video. Para este caso, las variables que influyen de forma

[41] Cantidad de usuarios nuevos

directa con la clase objetivo son *active_raffles*, *shares_first_day*, *duración*, *new_users* y *release_difference*.

Al igual que para el indicador interior, el resto de los algoritmos no generan información relevante para esta sección del estudio, por lo que se revisarán más a fondo en secciones posteriores.

Clasificación de registros con objetivo de calidad usuaria

El último de los indicadores a abordar con técnicas de clasificación es el de calidad usuaria. En este caso, la variable objetivo cuenta con 7 valores posibles, por lo que el árbol resultante no es tan claro como en los casos anteriores. En las hojas se presenta la probabilidad de que un registro perteneciente a esta sea de cada una de las clases, además del porcentaje total de observaciones que forman parte de ella.

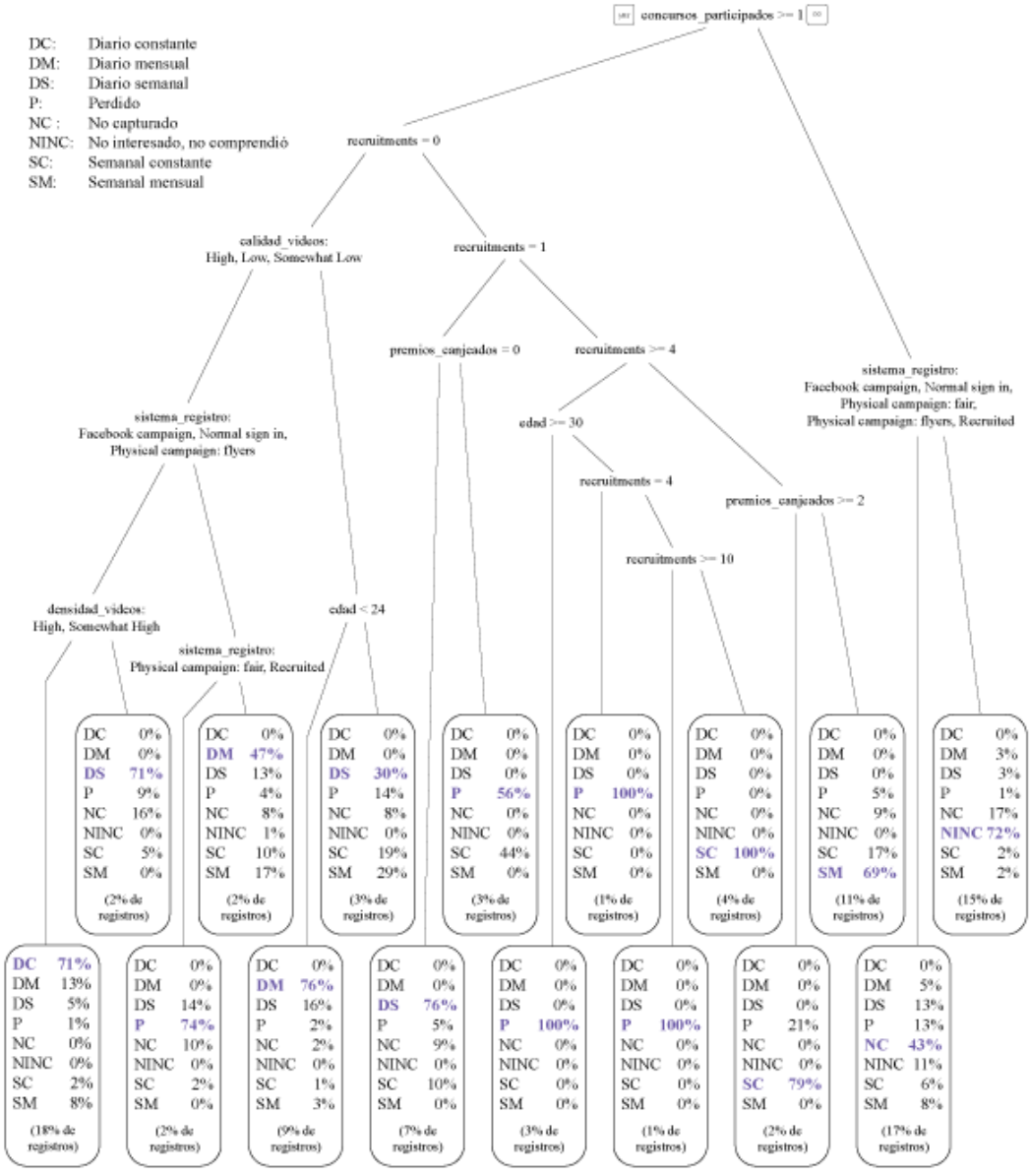
En la **ilustración 27** se muestra el árbol resultante para el indicador de calidad usuaria. A primera vista, las variables que influyen en él son *caldiad_videos*, *recruitments*, *concursos_participados*, *premios_canjeados*, *edad*, *sistema_registro* y *densidad_videos*.

A modo de simplificación para análisis, se decide agrupar las calidades usuarias en calidades positivas y negativas. El agrupamiento se realiza de la siguiente manera:

- Clases positivas de calidad usuaria: diario constante, diario semanal, diario mensual, semanal mensual y semanal constante.

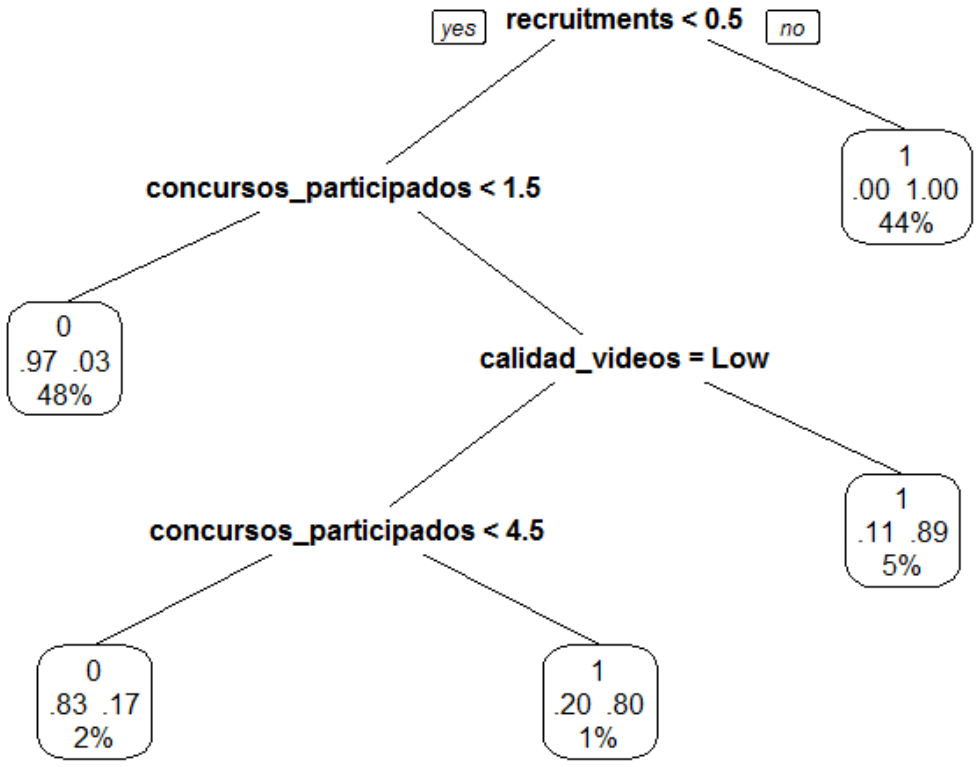
- Clases negativas de calidad usuaria: perdido, no capturado, no interesado/no comprendió.

Ilustración 27: Árbol de clasificación para calidad usuaria



En base a esta nueva forma de clasificación, se genera nuevamente el árbol de clasificación para el indicador. En la **ilustración 28** se presenta el nuevo árbol obtenido, esta vez sólo con dos clases posibles en el indicador objetivo. En este caso, se ve una fuerte dependencia de las variables *recruitmets*, *concursos_participados*, y *calidad_videos*.

Ilustración 28: Árbol de decisión (*rpart*) para calidad usuaria simplificada



4.3. Evaluación

En esta etapa de *CRISP-DM* se procede a evaluar los modelos obtenidos en la etapa anterior, compararlos entre sí y concluir en base al modelo que sea más adecuado para cada indicador.

Reglas de asociación

En la **tabla 11**, se presentaron las reglas de asociación encontradas para las calidades usuarias *No capturado* y *perdido*. De ellas se extrae que, para la primera de estas clases, malos valores de densidad de videos y concursos serán una característica

común, con un soporte de 25% y confianza de 63%, aproximadamente. Por otro lado, para el caso de los usuarios perdidos (clase *Perdido*), los soportes son preocupantemente bajos para concluir decisivamente, (alrededor del 10% para cada una de las 4 reglas encontradas), aunque, por otro lado, cada una de estas reglas se encuentra relacionada con una o más de las mismas combinaciones variable=valor (*sistema_registro=Recruited*, *premios_canjeados=0*, *recruitments=0*), y todas ellas tienen una confianza superior al 75%.

Técnicas de clasificación

En esta sección se evaluar individualmente cada uno de los clasificadores generados en la etapa anterior. Se separarán los resultados y comparaciones por indicador objetivo, ya que la efectividad de cada algoritmo depende del escenario en el que se esté aplicando.

Clasificación de indicador de usuarios activos

Para la ejecución del algoritmo ***rpart***, es importante, antes de concluir en torno a un árbol de decisión, saber qué tan bueno es prediciendo resultados de la clase objetivo. En este marco, se presenta en la **tabla 12** la matriz de confusión generada al evaluar el árbol relacionado con usuarios activos con el grupo de pruebas para este caso.

Tabla 12: Matriz de confusión para árbol de decisión (*rpart*) de usuarios activos

	No exitoso, predicho	Exitoso, predicho
No exitoso, original	215	43
Exitoso, original	13	191

Se concluye que el árbol generado es un método predictivo suficientemente bueno, teniendo aproximadamente un 83% de predicciones correctas para casos no exitosos, y un 94% de predicciones correctas para los casos exitosos.

Teniendo en consideración las variables que, para el árbol dispuesto en la **ilustración 19**, definen al indicador de usuarios activos, es interesante definir en qué magnitud influyen en el resultado final de dicho indicador. Para este objetivo se utiliza una herramienta presente en el paquete utilizado, llamada *variable.importance*, que da un acercamiento a qué tan importante es cada variable en comparación al resto. El resultado de esta función se aprecia en la **tabla 13**, donde a mayor valor, mayor la importancia al momento aplicar el predictor.

Tabla 13: Importancia de variables en árbol (*rpart*) de usuarios activos

Variable	Importancia relativa
active_raffles	259.13
active_canjes	146.42
shares_first_day	113.45
new_users	68.75
release_difference	36.88
duracion	3.79

Como era de esperarse por visualizaciones anteriores, la cantidad de concursos activos (*active_raffles*) juega un papel importante al momento de definir la cantidad de usuarios activos en un momento dado en la plataforma.

Siguiendo con la evaluación de los algoritmos aplicados, se presenta en la **tabla 14** la matriz de confusión para la ejecución de ***randomForest***. Para este caso, se aprecian un 87% de predicciones correctas en el caso de escenarios no exitosos, y un 89% en caso de escenarios exitosos, aproximadamente. Dela misma forma que para el clasificador anterior, se presenta en la **tabla 15** la lista de importancia de cada variable en este escenario. La columna "Clase: no exitoso", hace referencia a qué tan importante es esta variable al momento de predecir un registro de clase "no exitoso", o sea que, a mayor valor, más importante es la variable. Análogamente, la columna "Clase: exitoso" hará referencia a la clase "exitoso". *Mean Decrease Accuracy* hará referencia a qué tanto debe considerarse esta variable en el predictor con el fin de disminuir errores de clasificación. Finalmente, *Mean Descrease Gini* hace referencia a cómo disminuye la inequidad en la clasificación, valores más altos son mejores, ya que una inequidad baja significa una variable que en particular juega un papel más importante al dividir los datos en clases definidas. A modo de resumen, para la tabla de importancia de variables de *randomForest*, mayores valores se traducen en una variable más importante al momento de clasificar. Nuevamente *active_raffles* destaca como la variable más significativa.

Tabla 14: Matriz de confusión para clasificador de *randomForest* de usuarios activos

	No exitoso, predicho	**Exitoso, predicho**
No exitoso, original	202	30
Exitoso, original	26	204

Tabla 15: Importancia de variables en clasificador de *randomForest* de usuarios activos

Variable	Clase: no exitoso	Clase: exitoso	*Mean Decrease Accuracy*	*Mean Decrease Gini*
active_raffles	48.37	23.86	46.87	180.3
active_canjes	26.61	14.19	26.59	86.71
shares_first_day	22.84	11.27	23.89	105.32
new_users	14.19	0.62	13.19	69.53
release_difference	7.53	4.26	7.96	51.09
duracion	2.3	2.68	3.74	39.08

El siguiente clasificador evaluado es el generado por la ejecución del algoritmo *naiveBayes*. En la **tabla 16** se presenta la matriz de confusión relacionado con este clasificador. Se desprende de ella que un 83% de las predicciones de casos no exitosos son efectivas, mientras que un 91% de ellas lo son para el caso de la clase exitosa, aproximadamente. El algoritmo utilizado para la generación del clasificador de Bayes ingenuo no cuenta con herramientas incorporadas para definir la importancia de las variables en cada caso.

Tabla 16: Matriz de confusión para clasificador de *naiveBayes* de usuarios activos

	No exitoso, predicho	Exitoso, predicho
No exitoso, original	209	42
Exitoso, original	19	192

Para el caso de máquinas de vectores de soporte (*SVM* por su sigla en inglés[42]), se generaron dos clasificadores diferentes: uno utilizando como núcleo separadores lineales, y otro usando separadores radiales. Ese parámetro tiene influencia directa en el clasificador final, variando de caso a caso qué núcleo representa de mejor manera al conjunto de datos en cuestión.

Para el caso de *SVM lineal*, se presenta su matriz de confusión en la **tabla 17**. De ella se desprende que este clasificador es efectivo en el 85% de los casos para la clase no exitosa, y en un 92% para la clase exitosa, aproximadamente. Luego, en la **tabla 18**, se presenta la matriz de confusión para el caso de *SVM radial*, de donde se calcula un 85% de predicciones correctas en el caso de la clase no exitosa, y un 93% para la clase exitosa, aproximadamente. La ejecución de este algoritmo no genera elementos de importancia de variables.

Tabla 17: Matriz de confusión para clasificador de *SVM lineal* de usuarios activos

	No exitoso, predicho	Exitoso, predicho
No exitoso, original	211	38
Exitoso, original	17	196

Tabla 18: Matriz de confusión para clasificador de *SVM radial* de usuarios activos

	No exitoso, predicho	Exitoso, predicho
No exitoso, original	214	37
Exitoso, original	14	197

[42] Support vector machines

Finalmente se presenta en la **tabla 19** la **precisión**, **sensibilidad** y **especificidad** para el indicador de usuarios activos para cada uno de los clasificadores diferentes generados. La precisión se define como la probabilidad de que una predicción sea correcta; la sensibilidad, como la probabilidad de detectar un registro de clase exitosa en un universo de registros exitosos y la especificidad como la habilidad de detectar un registro no exitoso en un universo de datos no exitosos. Cada una de estas métricas se calcula de acuerdo a:

$$Precisión = \frac{(Verdaderos\ positivo + Verdaderos\ falsos)}{(Universo\ positivo + Universo\ negativo)}$$

$$Sensibilidad = \frac{Verdaderos\ positivos}{Universo\ positivo}$$

$$Especificidad = \frac{Verdaderos\ negativos}{Universo\ negativo}$$

Tabla 19: Métricas de clasificadores para caso de usuarios activos

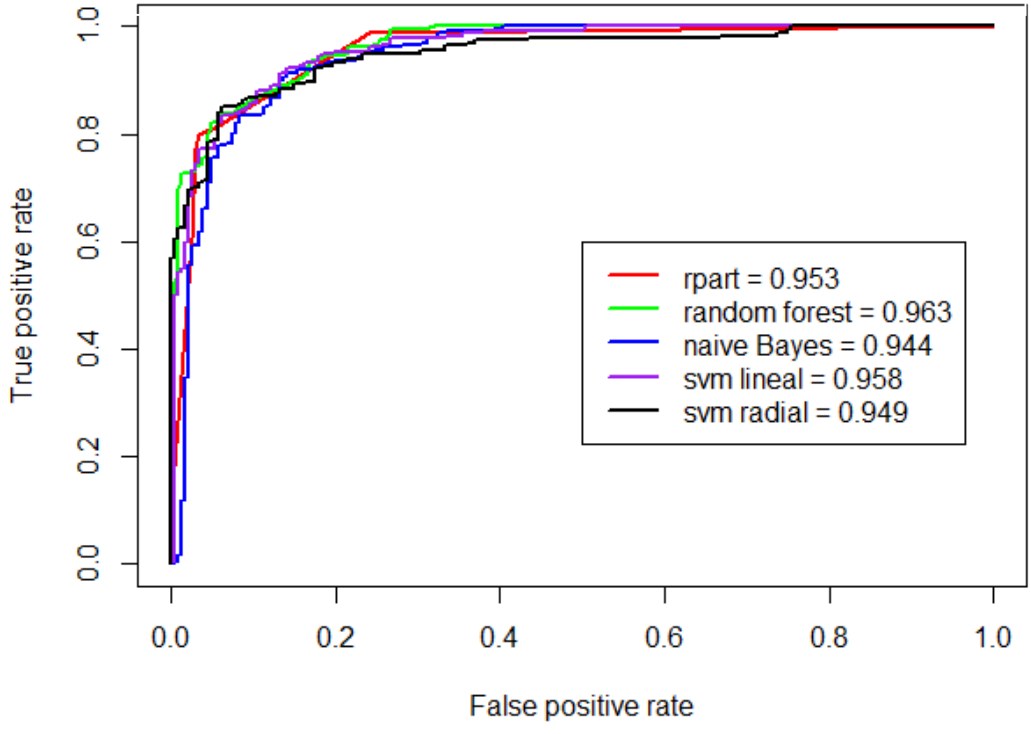
Clasificador	Precisión	Sensibilidad	Especificidad
Árbol (rpart)	87.88%	93.62%	83.33%
Random Forest	87.88%	88.7%	87.07%
Naive Bayes	86.8%	91%	83.27%
SVM lineal	88.1%	92.02%	84.74%
SVM radial	88.96%	93.36%	85.26%

Estas métricas dan un buen primer acercamiento para poder comparar los clasificadores generados, pero sólo cubren algunos de los casos. Como es propuesto

por *[11]*, las curvas *ROC*[43] son una buena forma de modelar cómo se comportan la sensibilidad y especificidad (cada una en relación a la otra).

Finalmente, el área bajo la curva del gráfico de sensibilidad (verdaderos positivos) contra 1-especifidad (falsos positivos) será un buen indicador de qué tan bueno es un clasificador. En el caso óptimo, esta curva tendrá un área de 1 (considerando los porcentajes de 0 a 1), por lo tanto, entre mayor sea la el área bajo una curva *ROC*, mejor será el predictor. En la **ilustración 28** se presenta la curva de cada uno de los clasificadores generados, y en la leyenda, el área bajo la curva respectiva. Se concluye finalmente que el mejor clasificador para este indicador en particular (usuarios activos) es el generado por el algoritmo *randomForest*.

Ilustración 28: Curvas *ROC* y áreas bajo estas para usuarios activos



[43] Acrónimo in inglés de *Receiver-operating characteristic.*

Clasificación de indicador de penetración

Para el clasificador del algoritmo *rpart* se presenta en la **tabla 20** la matriz de confusión del árbol generado del indicador de penetración. Se desprende que el árbol clasificará de forma correcta un video no exitoso (en relación a penetración) en un 71% de los casos, y uno exitoso en un 78% de ellos, aproximadamente.

Tabla 20: Matriz de confusión para clasificador de *rpart* de penetración

	No exitoso, predicho	Exitoso, predicho
No exitoso, original	182	76
Exitoso, original	44	152

Siguiendo con el análisis, se presenta en la **tabla 21** la importancia de las variables relacionadas con el árbol de clasificación de penetración. En ella se aprecia que *shares_first_day*, o sea, la cantidad de veces que se comparte un video en su primer día de publicación, es más importante al momento de clasificar un registro bajo este indicador. Además, se reitera la importancia de *active_raffles*, o sea, la cantidad de concursos activos al momento de la publicación del video.

Tabla 21: Importancia de variables en árbol (*rpart*) de penetración

Variable	Importancia relativa
shares_first_day	85.49
active_raffles	43.48
duracion	18.9
release_differende	13.83
new_users	9.11
active_canjes	5.1

El siguiente algoritmo aplicado para este indicador fue el del *randomForst*. En la **tabla 22** se presenta la matriz de confusión para este caso. El clasificador será efectivo en un 80% de los casos "no exitosos", y en un 83% de los casos que si lo son, aproximadamente.

Tabla 22: Matriz de confusión para clasificador de *randomForest* de penetración

	No exitoso, predicho	Exitoso, predicho
No exitoso, original	189	47
Exitoso, original	37	181

Siguiendo con el orden de este estudio, en la **tabla 23** se aprecia la importancia de cada variable para el caso del clasificador generado con *randomForest* para indicador de penetración. De ella se puede concluír que no todas las variables son importantes en iguales escenarios. Por ejemplo, al tratarse de

predecir un video exitoso a nivel de penetración, la cantidad de usuarios nuevos (*new_users*), la cantidad de usuarios activos (*active_users*) y la cantidad de canjes activos (*active_canjes*) no son importantes, pero sí lo son al momento de clasificar un video de clase no exitosa.

Tabla 23: Importancia de variables en clasificador de *randomForest* de penetración

Variable	Clase: no exitoso	Clase: exitoso	*Mean Decrease Accuracy*	*Mean Decrease Gini*
shares_first_day	24.63	7.98	25.39	92.54
active_raffles	24.87	2.35	25.63	80.69
new_users	19.24	-3.19	14.21	69.49
active_users	18.6	-4.19	16.95	71.23
duracion	10.38	8.31	13.03	83.98
active_canjes	10.62	-1.57	9.5	31.46
release_difference	9.22	3.87	9.6	71.42

En relación al clasificador generado con el algoritmo *naiveBayes*, se presenta en la **tabla 24** su matriz de confusión para el indicador de penetración. Una primera mirada revela que no parece ser de los mejor clasificadores de este estudio, ya que será efectivo para el caso de los registros no exitosos sólo en un 62% de las pruebas, y para los exitosos en un 76% de ellas, aproximadamente. Como se mencionó anteriormente, el algoritmo de Bayes ingenuo utilizado en este estudio no brinda información con respecto a la importancia de las variables.

Tabla 24: Matriz de confusión para clasificador de *naiveBayes* de penetración

	No exitoso, predicho	Exitoso, predicho
No exitoso, original	193	122
Exitoso, original	33	106

Los últimos clasificadores a evaluar son los generados por el algoritmo *SVM*, con núcleo lineal y radial. En las **tablas 25 y 26** se presentan las matrices de confusión para las ejecuciones de los algoritmos *NVM lineal* y *radial* respectivamente, de ellas se concluye que el clasificador en su versión *radial* es más efectivo al momento de clasificar registros tanto no exitosos como exitosos que su contraparte en versión lineal (69% contra 62% en casos no exitosos, y 82% contra 78% en casos exitosos, aproximadamente).

Tabla 25: Matriz de confusión para clasificador de *SVM lineal* de penetración

	No exitoso, predicho	Exitoso, predicho
No exitoso, original	196	120
Exitoso, original	30	108

Tabla 26: Matriz de confusión para clasificador de *SVM radial* de penetración

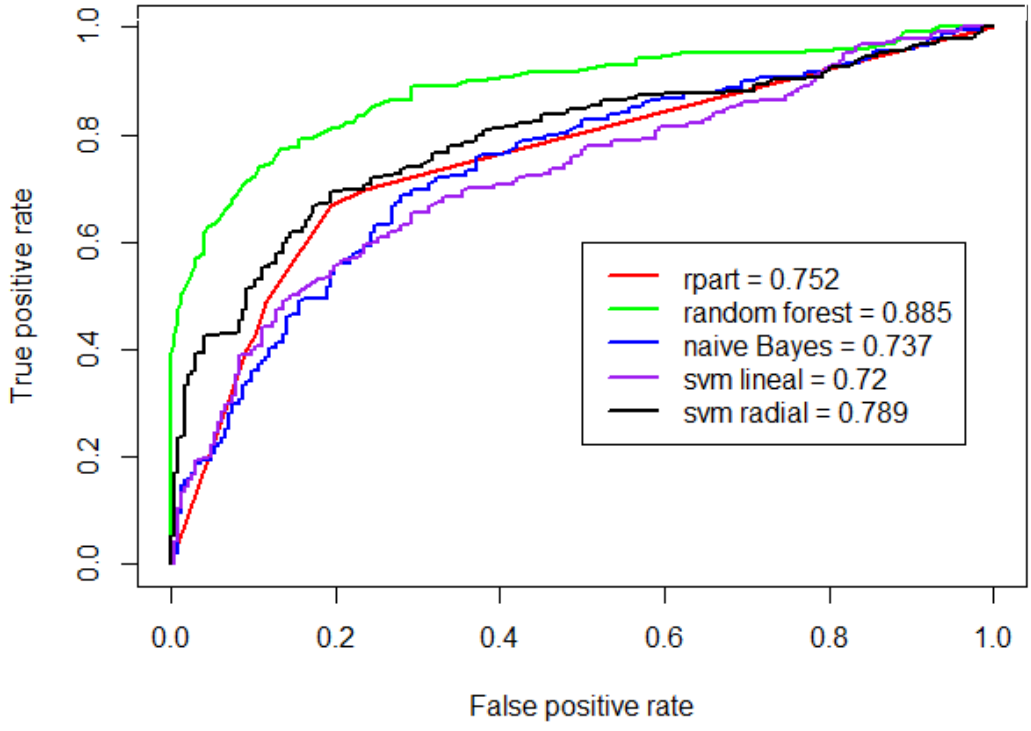
	No exitoso, predicho	Exitoso, predicho
No exitoso, original	192	87
Exitoso, original	32	141

A modo de resumen, se comparan en la **tabla 27** las métricas de precisión, especificidad y sensibilidad para cada uno de los clasificadores utilizados para el indicador de penetración.

Tabla 27: Métricas de clasificadores para caso de penetración

Clasificador	Precisión	Sensibilidad	Especificidad
Árbol (rpart)	73.57%	77.55%	70.54%
Random Forest	81.5%	83.03%	80.08%
Naive Bayes	65.86%	76.26%	61.27%
SVM lineal	66.96%	78.26%	62.03%
SVM radial	73.67%	81.5%	68.82%

Ilustración 29: Curvas *ROC* y áreas bajo estas para penetración



Finalmente, en la **ilustración 29** se presenta la curva de cada uno de los clasificadores generados, y en la leyenda, el área bajo la curva respectiva. Se concluye que el mejor clasificador para este indicador en particular (penetración), por una diferencia considerable, es el generado por el algoritmo *randomForest*.

Clasificación de indicador de calidad usuaria (siplificada)

Por simplicidad, posibilidad de evaluación y comparación, y necesidades del negocio, se decide no utilizar el modelo de calidades usuarias generado para los 8 valores posible, si no el simplificado. Para el objetivo de este estudio, no es necesario ahondar en una calidad usuaria positiva, ya que cualquiera de estas es satisfactoria para el objetivo del negocio.

En la (tabla) se presenta la matriz de confusión relacionada con el árbol de clasificación generado con *rpart* para para el indicador de calidades usuarias (simplificadas). De ella se desprende que el modelo es satisfactorio en el 98% de los casos para calidades usuarias positivas, y en el 97% para negativas, aproximadamente; por lo que, a primera vista, pareciera ser un muy buen modelo para estos escenarios.

Tabla 28: Matriz de confusión para clasificador de *rpart* de calidades usuarias simplificadas

	Calidad mala, predicho	Calidad buena, predicho
Calidad mala, original	739	20
Calidad buena, original	16	695

Luego se procede a evaluar las variables incluidas en el árbol, cuya importancia se ve representada en la **tabla 29**. Como era de esperarse al ver el árbol

relacionado, la cantidad de reclutamientos tiene una importancia resaltante en el modelo, junto con concursos participados.

Tabla 29: Importancia de variables en árbol (*rpart*) de calidades usuarias simplificadas

Variable	Importancia relativa
recruitmets	1329.17
concursos_participados	1152.92
premios_canjeados	843.66
calidad_videos	252.68
densidad_videos	233.11
edad	229.97
sistema_registro	14.21
densidad_concursos	1.04

Del segundo modelo aplicado al indicador de calidad usuaria simplificada, *randomForest*, se muestra la matriz de confusión en la **tabla 30**. De ella se calcula que el 99% de las clasificaciones negativas fueron acertadas, así como el 93% de las positivas, aproximadamente. Luego, en la **tabla 31**, se presenta la lista de importancia de variables de este modelo.

Tabla 30: Matriz de confusión para clasificador de *randomForest* de calidades usuarias simplificadas

	Calidad mala, predicho	Calidad buena, predicho
Calidad mala, original	749	9
Calidad buena, original	6	706

Tabla 31: Importancia de variables en clasificador de *randomForest* de calidades usuarias simplificadas

Variable	Clase: mala	Clase: buena	*Mean Decrease Accuracy*	*Mean Decrease Gini*
recruitments	48.98	27.35	50.38	719.09
concursos_participados	46.25	17.44	47.67	618.28
premios_canjeados	27.48	16.13	28.26	266.21
sistema_registro	20.81	14.12	25.68	71.49
densidad_videos	11.29	9.11	11.44	51.33
edad	6.84	7.42	9.7	32.25
calidad_videos	12.84	5.38	13.92	33.53
densidad_concursos	7.99	0.72	6.58	18.21
genero	-1.74	-0.76	-1.9	4.35

El siguiente algoritmo aplicado, al igual que en los indicadores anteriores, es *naiveBayes*. Del modelo generado de su ejecución se desprende la matriz de confusión presentada en la **tabla 31**. Se calcula que el modelo es efectivo para el 98% de los casos de calidad positiva y para el 96% de las negativas, aproximadamente.

Tabla 31: Matriz de confusión para clasificador de *naiveBayes* de calidades usuarias simplificadas

	Calidad mala, predicho	Calidad buena, predicho
Calidad mala, original	744	34
Calidad buena, original	11	681

El último de los algoritmos para modelar clasificadores para el indicado de calidad usuaria es *SVM*. Al igual que en casos anteriores, se aplica con núcleo *lineal* y *radial*, y sus matrices de confusión resultantes se aprecian en las **tablas 32 y 33**. De ellas se desprende que ambos modelos son buenos para clasificar tanto un usuario malo como bueno, con una efectividad del 98% y 99% de manera aproximada, respectivamente (en ambos casos).

Tabla 32: Matriz de confusión para clasificador de *SVM lineal* de calidades usuarias simplificadas

	Calidad mala, predicho	Calidad buena, predicho
Calidad mala, original	751	14
Calidad buena, original	4	701

Tabla 33: Matriz de confusión para clasificador de *SVM radial* de calidades usuarias simplificadas

	Calidad mala, predicho	Calidad buena, predicho
Calidad mala, original	751	18
Calidad buena, original	4	697

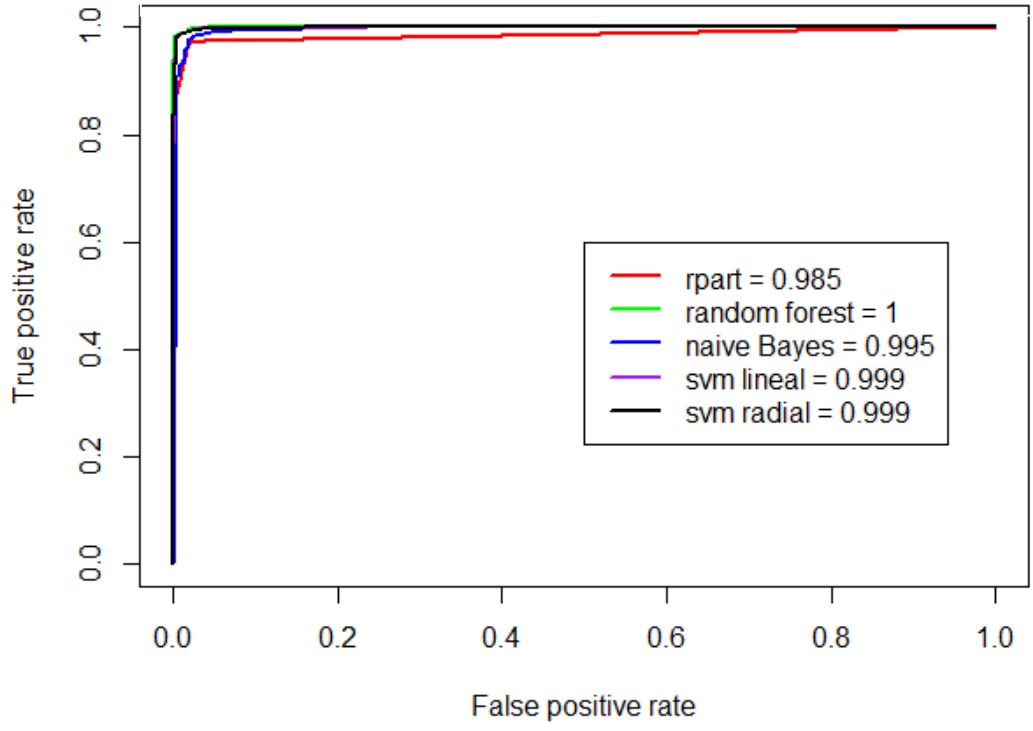
De manera resumida se presentan en la **tabla 34** las diferentes métricas de calidad de los modelos construidos para el indicador de calidad usuaria simplificada,

y en la **ilustración 30**, sus curvas *ROC* junto con los valores de área bajo la curva en cada caso.

Tabla 27: Métricas de clasificadores para caso de calidades usuarias simplificadas

Clasificador	Precisión	Sensibilidad	Especificidad
Árbol (rpart)	97.55%	97.75%	97.36%
Random Forest	98.98%	99.16%	98.81%
Naive Bayes	96.94%	98.41%	95.63%
SVM lineal	98.78%	99.43%	98.17%
SVM radial	98.5%	99.43%	97.66%

Ilustración 29: Curvas *ROC* y áreas bajo estas para penetración



4.4. Despliegue (PENDIENTE)

En esta, la última etapa del proceso *CRISP-DM*, se despliega de forma resumida toda la información relevante obtenida en las fases anteriores. Es dentro de este objetivo que se dividirá esta sección en tres, abordando en cada una toda la información rescatable referente a los indicadores propuestos para este estudio.

Despliegue para usuarios activos

Es importante para el negocio saber en qué condiciones la plataforma ha tenido cúspides de actividad usuaria, para así poder replicar dichas condiciones, ver cómo influyen, e incluso mejorarlas. En base a este principio, se presentan en la **tabla 19** la lista de factores que influyen a esta variable para el contexto de la plataforma del estudio.

Tabla 19: Variables que influyen sobre usuarios activos

Factor o variable	Descripción de relación	Respaldo
Concursos activos	Los concursos afectan de forma directa a la cantidad de usuarios activos en un momento dado en la plataforma, esto es, a mayor cantidad de concursos activos, mayor cantidad de usuarios activos.	Visualización de datos utilizando herramientas OLAP (*Tableau*), árbol de clasificación.
Canjes activos	De la misma manera que los concursos, los canjes activos se relacionan directamente con la cantidad de usuarios activos.	Visualización de datos utilizando herramientas OLAP (*Tableau*), árbol de clasificación.
Usuarios nuevos	Como es de esperarse, una mayor cantidad de usuarios nuevos tiene como consecuencia una mayor cantidad de	Árbol de clasificación.

	usuarios activos.	
Cantidad de veces que se comparte un video en su primer día	Si bien el estudio muestra una relación de esta variable con la cantidad de usuarios activos, conocimiento del negocio hace ver que, si bien la relación existe, si dirección es hacia el otro sentido. Son los usuarios activos los que comparten videos en sus primeros días, y es por esto que, por definición, a mayor cantidad de usuarios activos, más probabilidad hay que un video sea ampliamente compartido.	Árbol de clasificación.

Despliegue para penetración

Un video con buena penetración tiene como consecuencia una mayor propagación del contenido del cliente, tanto en la plataforma como en las redes sociales de sus usuarios, además de una ayuda evidente a la plataforma como marca, al compartirse su imagen y dominio junto con el video. En la **tabla 20** se presentan las variables que influyen sobre la penetración de un video.

Tabla 20: Variables que influyen sobre penetración

Factor o variable	Descripción de relación

Diferencia de lanzamiento y duración	Se agrupan estas variables ya que la relación encontrada no se limita exclusivamente a ninguna de ellas (al menos durante el proceso en que se observa dicho comportamiento). A menores rangos de duración, y cuando la diferencia de lanzamiento del contenido original y su publicación en la plataforma es de 6 horas o menos, se aprecia un máximo de penetración con respecto al resto de combinaciones de intervalos. Esta relación se ve reflejada además en el árbol de ckasificación generado, donde dos nodos consecutivos, de duración y diferencia de lanzamiento respectivamente, terminan en una hoja con 82% de probabilidad de éxito (70% de penetración o más, para videos de menos de 60 segundos y diferencias de lanzamiento de menos de 38 horas). Respaldo: Visualización de datos utilizando herramientas OLAP (*Tableau*), árbol de clasificación.
Cantidad de veces que se comparte un video en su primer día	Al analizar el árbol de decisión generado, se aprecia que, en todos los casos, mientras más veces sea compartido un video en su primer día, mayores probabilidades de alcanzar valores altos de penetración (sobre 70%) tendrá. Respaldo: Árbol de clasificación.
Concursos activos	La cantidad de concursos activos al momento de lanzamiento de un video afecta directamente a su penetración. El óptimo encontrado para este caso se fija en 9 o más concursos. Respaldo: Árbol de clasificación.

Despliegue para calidad usuaria

La calidad usuaria es un factor fundamental para entender a la comunidad de la plataforma, y poder potenciar aquellos factores que atraen, o potencian, las calidades provechosas para conseguir los objetivos del negocio. En la **tabla 21** se presentan las variables encontradas que afectan a las diferentes clases de calidad usuaria.

Tabla 21: Variables que influyen sobre calidad usuaria

Factor o variable	Descripción de relación
Calidad de videos	Durante el proceso de visualización de datos con herramientas OLAP, se encuentra una relación entre la calidad de videos y las clases negativas de calidad usuaria. Para estas tres clases, la calidad de videos fluctúa entre *Low* y *Somewhat Low*, que son las dos categorías peor valoradas en la escala. Respaldo: Visualización de datos utilizando herramientas OLAP (*Tableau*), árbol de clasificación.
Densidad de concursos	Para la clase usuaria *No capturado*, hay un peso notorio en la densidad de concursos *Low*. Esto quiere decir, que al momento en que se registró un usuario (y la semana siguiente a este registro), la densidad de concursos fue baja. En otras palabras, en la primera semana de la mayoría de los usuarios *No capturados*, la densidad de concursos fue baja. En muy pocos casos, fue *Regular* o mejor. Esta relación se confirma luego con la generación de reglas de asociación. Respaldo:

	Visualización de datos utilizando herramientas OLAP (*Tableau*), reglas de asociación, árbol de clasificación.
Densidad de videos	Nuevamente para la clase usuaria *No capturado*. De acuerdo a lo observado en *Tableau*, Para la gran mayoría de los casos de usuarios *No capturados*, la densidad de videos al momento de registro y su semana siguiente es baja, y en una segunda mayoría, relativamente baja. Esta relación se confirma luego con la generación de reglas de asociación. Por otro lado, al tratarse de las calidades usuarias positivas, para todas estas, la densidad de videos se encontraba en valores altos o relativamente altos al momento de registro. El análisis del árbol de clasificación confirma esta variable como la más importante al momento de definir la calidad usuaria, y confirma que a valores altos hay probabilidades altas (sobre el 90%) de que un usuario sea de buena calidad. Respaldo: Visualización de datos utilizando herramientas OLAP (*Tableau*), reglas de asociación, árbol de clasificación.

Sistema de registro	El sistema de registro más común (por una amplia mayoría) para el caso de usuarios *No interesado/No comprendió* es a través de la vitrina de videos. Este sistema de registro también destaca para la clase usuaria *No capturado*. Por otro lado, se aprecia una relación entre el sistema de registro de reclutamiento y la calidad usuaria *Perdido*. La que se confirma luego por la generación de reglas de asociación. En referencia a las calidades usuarias positivas, la mayoría de los registros se realiza a través de campañas de Facebook, y campañas de *flyers*. Durante el análisis del árbol de clasificación, se confirma que el sistema de registro es un buen clasificador. Respaldo: Visualización de datos utilizando herramientas OLAP (*Tableau*), reglas de asociación, árbol de clasificación.
recruitmets	La cantidad de usuarios que recluta un usuario dado tiene influencia directa sobre su clasificación. Esto se ve reflejado tanto en reglas de asociación como en árboles de decisión. Esto quiere decir que los usuarios de buena calidad suelen reclutar a otros usuarios, pero no se trata de una variable que pueda ser manejada por quien administra la plataforma. Cabe destacar que se descubrió anteriormente que los usuarios reclutados suelen terminar siendo de la clase *Perdido*. Respaldo: Reglas de asociación, árbol de clasificación.

Conclusiones

Luego de este estudio se tiene un panorama más completo del funcionamiento de Kikvi y de sus usuarios. Cabe destacar que todas estas conclusiones aplican para el contexto del estudio realizado, y que deben ser comprobadas caso a caso, aunque se recomienda tomarlas como referencia o punto de partida en servicios similares.

Una primera mirada a los datos reveló que la mejor categoría de videos para ser compartida por la comunidad es la de **videojuegos**. Sobre esto, se plantean la siguiente consideración: los videos de videojuegos en la plataforma comprenden principalmente *teasers* y *trailers* de nuevos lanzamientos, y no videos de *streaming*[44] ni *gameplays*[45]. Esta notoria ventaja de los videos de videojuegos no se ve traducida en mayor cantidad de vistas. De la misma manera, no se encuentra una relación visible entre la cantidad de veces que se comparte un video y la cantidad de vistas que consigue.

En relación a acciones a tomar en la plataforma o en el servicio relacionado con la misma, se concluye:

- **El sistema actual de reclutamiento no genera valor para el negocio.** Si bien los usuarios de buenas calidades reclutan a otros usuarios, estos nuevos reclutas tienen una altísima probabilidad de convertirse en usuarios perdidos. Se recomienda la eliminación del sistema de reclutamiento, o realizar estudios siguientes para idear uno que capture usuarios de buena calidad.

- **El sistema de registro influye directamente en la calidad que tendrá un usuario a futuro**. Se descubrió que las formas más efectivas de conseguir

[44] Video generalmente en vivo de una persona jugando algún videojuego en particular.
[45] Video que muestra fragmentos de una persona jugando un videojuego en particular.

usuarios de calidad son a través de campañas en Facebook (a través del *fanpage*) y campañas de *flyers*. Por otro lado, la forma menos efectiva de conseguir usuarios es a través del registro en vitrina. Esto se puede deber a que las personas que se registran a través de este medio no tienen información del sitio ni de la plataforma, y al tratarse de un sistema bloqueante[46], el usuario no sabe en qué se está registrando. Se cree que además esto puede producir roce para usuarios potencialmente buenos, y que no da tiempo de mostrar lo atractivo de la plataforma. Se recomienda liberar las secciones principales de la plataforma, y motivar al usuario a registrarse cuando este ya tenga un entendimiento mayor de la misma, y se haya familiarizado con el contenido y la comunidad.

- **Es fundamental mantener un buen ambiente en la plataforma**. Se ha descubierto que el ambiente al momento de registro tiene influencia en la calidad que tendrá un usuario a futuro, además de su permanencia en el sistema. Las consideraciones más importantes en torno a este punto se relacionan por un lado con la cantidad de concursos y canjes, y por otro lado a la densidad de videos. A mayor cantidad de concursos, más usuarios activos habrá en la plataforma, y los usuarios que se registren en este período tendrán mayor probabilidad de pertenecer a las calidades usuarias positivas. En relación a la densidad de videos, a mayor cantidad de videos (nuevos) al momento de registro de un nuevo usuario, es más probable que este se mantenga interactuando con la plataforma. La calidad de los videos también juega un papel en el ambiente al momento de registro, aunque su importancia es menor a las de las variables ya mencionadas.

- **Los videos más cortos, y más nuevos, tienen mejor aceptación**. Un video de menor duración tiene mejor aceptación por la comunidad de la plataforma.

[46] Actualmente, sólo se puede ver la vitrina de un video y la página principal (sin contenido atractivo) de Kikvi siendo un usuario no registrado.

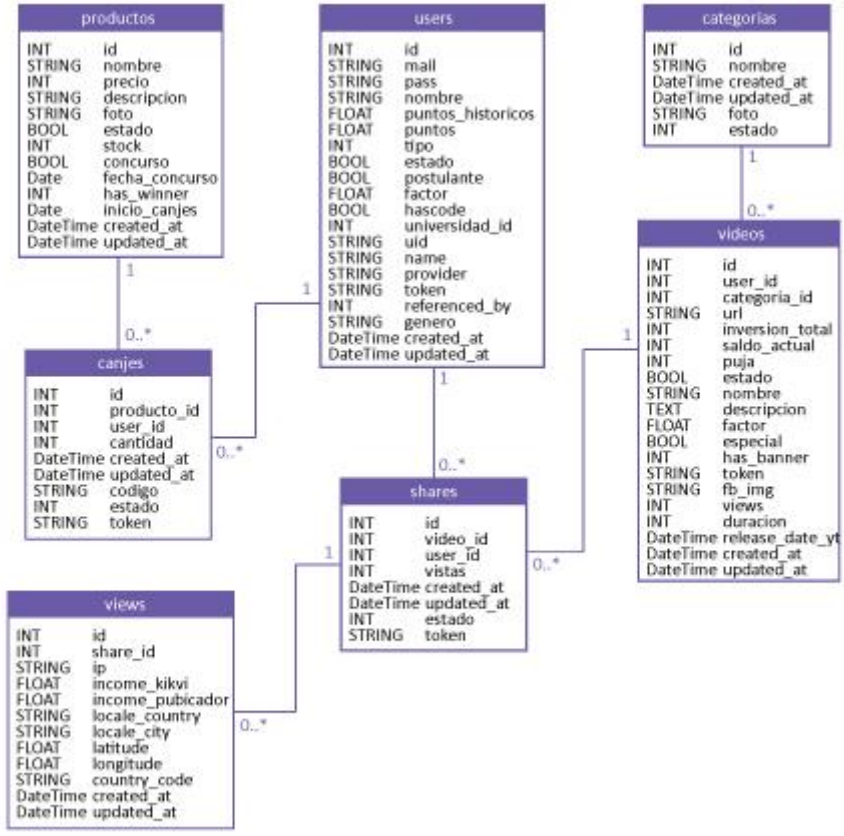
Además, es fundamental que el lanzamiento de dicho video sea rápido en comparación a otras fuentes de publicación. Se recomienda traspasar esta información al cliente para que lo tenga en consideración al momento de generar nuevos contenidos. Además, se descubre que la cantidad de veces que un video se comparte en el primer día de su publicación influye directamente en la penetración que conseguirá finalmente, por lo que se recomienda potenciar los videos en sus inicios, y así generar esta relación.

Referencias bibliográficas

[1] ***UCLAAnderson*** *- school of management, http://www.anderson.ucla.edu -*
[2] ***KDD, SEMMA AND CRISP-DM: A PARALLEL*** *OVERVIEW - Azevedo & Santos - 2008*
[3] ***Advanced Tech Computing Group UTPL*** *- http://advancedtech.wordpress.com/ - última visita en mayo 2016*
[4] ***OLAP TOOLS*** *- http://www.informationbuilders.com/olap-online-analytical-processing-tools - última visita en mayo 2016*
[5] ***Identification of Outliers*** *-- Douglas M. Hawkins, 1980*
[6] ***Indicadores Claves de Desempeño o Key Performance Indicator*** *-- http://www.profitline.com.co/BPO/BusinessProcessOutsourcing/182/indicadores-claves-de-desempeno-o-key-performance-indicator.html - última visita en mayo 2016*
[7] ***Refinement of approximate domain theories by knowledge-based neural networks****, G. Towell, J. Shavlik, and M. Noordewier, 1990.*
[8] ***Rule extraction from artificial neural networks****, H. MAYER, Huber, Rohde, and Tamme, Universitsat Salzburg, 12th October 2006 2006.*
[9] ***Active Users: Measure the Success of Your Business****, Claudiu Murariu, 2014 - https://blog.innertrends.com/active-users-2/748 - última vista en mayo 2016*
[10] ***Algorithms and Applications for Spatial Data Mining*** **-** Martin Ester, Hans-Peter Kriegel, Jörg Sander (University of Munich), 2001.
[11] **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** - Hanley, J. A. and B. J. McNeil, 1982.

Anexo 1: Modelo de datos relacional

Tablas relevantes al estudio.



Anexo 2: Script API Facebook PHP

```
<?php
   require(“fbSDK/autoload.php”);
   use Facebook\FacebookSession;
   use Facebook\FacebookRequest;
   use Facebook\GraphUser;
   use Facebook\FacebookRequestException;
   use Facebook\FacebookRedirectLoginHelper;
   $facebook = FacebookSession::setDefaultApplication(<app_id>,
<app_token>);
   $session = FacebookSession::newAppSession();
   $sql = “SELECT
           users.*
         FROM
           users
         WHERE
           users.id != 0
           AND users.id != 8
           AND users.tipo = 1
           AND users.estado = 1
           AND users.uid IS NOT NULL”;
   $db = mysql_connect(“localhost”,<db_user>,<db_pass>);
   $selected = mysql_select_db(<db_name>);
   $rs = mysql_query($sql);
   $getFromFacebook = array();
   while($row = mysql_fetch_assoc($rs)){
      if(!empty($row[“uid”])){
         $tmp = new StdClass();
         $tmp->id = $row[“id”];
         $tmp->fbid = $row[“uid”];
         array_push($getFromFacebook, $tmp);
      }
   }
   mysql_close();
   $total = sizeof($getFromFacebook);
   $current = 1;
   $updateQuery = ““;
   foreach($getFromFacebook as $user){
      print_r(“Fetching... “.$current.”/”.$total.”\n”);
      print_r(“/”.$user->fbid.”\n”);
      $request = new FacebookRequest($session, 'GET', '/'.$user->fbid);
```

```
     try{
        $response = $request->execute();
        $graphObject = $response->getGraphObject();
        $gender = $graphObject->getProperty('gender');
        $gender = strtoupper(substr($gender,0,1));
        $updateQuery .= “UPDATE users SET genero = '$gender' WHERE id =
'$user->id';”;
     }
     catch (Exception $e){
        print_r(“Error: perfil borrado”.”\n”);
     }
     $current++;
  }
  $file = “sqlQuery.sql”;
  file_put_contents($file, $updateQuery);
?>
```

Anexo 3: Script API Youtube (RoR)

```
 def get_duracion_videos_from_yt
  if not params[:videos].nil?
     concatenated_yt_ids = Video.where(:id =>
params[:videos]).map(&:url).join(“,”)
        url = <GOOGLE API URL WITH PRIVATE TOKEN>
        begin
           video_info = open(url)
           video_info = JSON.parse video_info.read
           video_info[“items”].each do |v|
              duration = v[“contentDetails”][“duration”]
              seconds = 0
              duration = duration.sub! “PT”, ““
              if duration.include? “H” and duration.include? “M” and
duration.include? “S”
                 duration_hours = duration.split(“H”)[0].to_i
                 duration_minutes = duration.split(“M”)[0].split(“H”)[1].to_i
                 duration_seconds = duration.split(“M”)[1].gsub! “S”,””
                 duration_seconds = duration_seconds.to_i
                 seconds = duration_hours*3600 + duration_minutes*60 +
duration_seconds
              elsif duration.include? “M” and duration.include? “S”
                 duration_minutes = duration.split(“M”)[0].to_i
                 duration_seconds = duration.split(“M”)[1].gsub! “S”,””
                 duration_seconds = duration_seconds.to_i
                 seconds = duration_minutes*60 + duration_seconds
              elsif duration.include? “H” and duration.include? “M”
                 duration_hours = duration.split(“H”)[0].to_i
                 duration_minutes = duration.split(“H”)[1].gsub! “M”,””
                 duration_minutes = duration_minutes.to_i
                 seconds = duration_hours*3600 + duration_minutes*60
              elsif duration.include? “H” and duration.include? “S”
                 duration_hours = duration.split(“H”)[0].to_i
                 duration_seconds = duration.split(“H”)[1].gsub! “S”,””
                 duration_seconds = duration_seconds.to_i
                 seconds = duration_hours*3600 + duration_seconds
              elsif duration.include? “H”
                 duration_hours = duration.split(“H”)[0].to_i
                 seconds = duration_hours*3600
              elsif duration.include? “M”
                 duration_minutes = duration.split(“M”)[0].to_i
```

```
            seconds = duration_minutes*60
          elsif duration.include? “S”
            duration_seconds = duration.split(“S”)[0].to_i
            seconds = duration_seconds
          end
          Video.where(:url => v[“id”]).update_all(:duracion => seconds)
        end
      rescue StandardError=>e
        puts “WOOOPS”
      end
    end
    render :nothing => true
end
```