Índice

[Índice 1](#_Toc449603502)

[1. Definición del problema 3](#_Toc449603503)

[1.1. Difusores de contenido audiovisual 3](#_Toc449603504)

[1.2. Kikvi 4](#_Toc449603505)

[1.3. El problema 6](#_Toc449603506)

[1.4. Objetivos 7](#_Toc449603507)

[2. Estado del arte 9](#_Toc449603510)

[2.1. Minería de datos 9](#_Toc449603511)

[2.2. Procesos de minería de datos 11](#_Toc449603512)

[Proceso de descubrimiento del conocimiento KDD 11](#_Toc449603513)

[SEMMA (Sample, Explore, Modify, Model and Assess) 14](#_Toc449603514)

[CRISP-DM (Cross-Industry Standard Process for Data Mining) 15](#_Toc449603515)

[2.3. Tareas de minería de datos 16](#_Toc449603516)

[Tareas descriptivas 16](#_Toc449603517)

[Tareas predictivas 20](#_Toc449603518)

[2.4. Herramientas de minería de datos 22](#_Toc449603519)

[3. Diseño de la solución 25](#_Toc449603520)

[3.1. Entendimiento del negocio 25](#_Toc449603521)

[3.2. Entendimiento de los datos 27](#_Toc449603522)

[Visualización de datos usando Tableau (OLAP) 27](#_Toc449603523)

[4. Desarrollo de la solución 37](#_Toc449603524)

[4.1. Preparación de los datos 37](#_Toc449603525)

[4.2. Modelo 37](#_Toc449603526)

[4.3. Evaluación 37](#_Toc449603527)

[4.4. Despliegue 37](#_Toc449603528)

[5. Conclusiones 38](#_Toc449603529)

[6. Bibliografía 39](#_Toc449603530)

[7. Referencias 40](#_Toc449603531)

[8. Anexos (color: 6b59a6) 41](#_Toc449603532)

[Anexo 1: Modelo de datos de Kikvi 41](#_Toc449603533)

[Anexo 2: Scripts para APIs 41](#_Toc449603534)

[Facebook (PHP) 41](#_Toc449603535)

[Youtube (RoR) 42](#_Toc449603536)

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. 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]](#footnote-1)), y aún más invasiva en otros, como es el caso de portales de contenido viral con *banners* y *popups[[2]](#footnote-2)*. Con el paso del tiempo, los usuarios de internet se han vuelto reacios y menos susceptibles al segundo tipo de publicidad, ocurriendo fenómenos como la llamada **ceguera del banner[[3]](#footnote-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. 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: los **publicadores**, y los **creadores**. El grupo de **creadores** 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 consistía en una serie de personas 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 produce 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:

* Losq 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. 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. Se hace necesario entender el negocio, entender los procesos y entender a los usuarios 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 relaciona a un usuario como en relación 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. 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, mejorar la difusión de campañas audiovisuales a través de redes sociales, es la correcta.

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 delos usuarios de la plataforma ante ella, entonces, ¿cómo puede **Kikvi** apoyar esta campaña?

Es necesario 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. Objetivos

Objetivo principal

Mejorar la percepción y experiencia usuaria 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 e 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.

1. 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.

* 1. Minería de datos

De forma general, la minería de datos consiste en el proceso de analizar datos[[4]](#footnote-4) de múltiples fuentes, desde diferentes perspectivas, con el fin de resumirla en información[[5]](#footnote-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 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 almacenamiento y software estadísticos, ha logrado incrementar dramáticamente la precisión de los análisis, mientras disminuyen los costos de realizarlos.

De una forma muy simplificada, la minería de datos consiste en la identificación de patrones en sets, generalmente de grandes dimensiones, de datos, con el fin de adquirir algún conocimiento[[6]](#footnote-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 datos se usa hoy vastamente por todo tipo de organizaciones. 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 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.

La minería está compuesta de 5 etapas principales:

* Extraer, transformar, y cargar datos en el *data warehouse*.
* Almacenar y administrar los datos en un sistema de bases de datos multidimensional.
* 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:

* Redes neuronales: modelos predictivos no lineales que se 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.
* 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. Llamada también *k-nearest* *neightbours*.
* Reglas de inducción: Conjunto de reglas *if-then* útiles, basadas en significancia estadística.
* Visualización de datos: interpretación visual de relaciones complejas en datos multidimensionales.
  1. 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[[7]](#footnote-7)

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 *Azevedo y Santos* (Referencia 4 - KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW - Azevedo & Santos), *KDD* es el proceso de usar métodos de minería de datos para extraer lo que es considerado conocimiento de acuerdo a una seria de medidas y umbrales, usando bases de datos en conjunto con cualquier pre-procesamiento necesario, extracto de muestras, o transformación. El proceso cuenta con 5 fases fundamentales: Selección, Pre-proceso, Transformación, Minería de datos, Interpretación/Evaluación.

El *KDD* es precedido por el desarrollo de un entendimiento del área de aplicación, cualquier conocimiento previo relevante y los objetivos del usuario final. Es un proceso iterativo e interactivo, e involucra numerosos pasos con muchas decisiones tomadas en el camino por el usuario.

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.

Ilustración : Proceso del descubrimiento del conocimiento (KDD) (agregar referencia)



1. **Selección**: esta etapa consiste en definir un conjunto de datos, o enfocar los esfuerzos en una serie de variables de los mismos. Durante esta etapa, es fundamental contar con un conocimiento previo del negocio, que ayudará a definir cuáles variables son relevantes para el estudio y cuales 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-proceso**: 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.

1. **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. Algunos tipos de transformación son:

* **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) contra 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**: consiste en, como dice su nombre, 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).

1. **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 (más detalles ya vistos en la sección 2)
2. **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.

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

*SEMMA* es una serie de pasos 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* (**Muestra**): consiste en la selección de un set 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* (**Explorar**): durante esta fase se visualizan de 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* (**Modificar**): 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* (**Modelar**): 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* (**Evaluar**): 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[[8]](#footnote-8)*, que cuentan con fases[[9]](#footnote-9) enfocadas en estos aspectos.

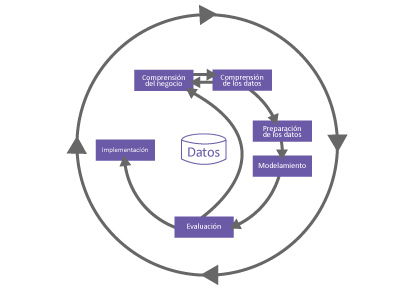
CRISP-DM (Cross-Industry Standard Process for Data Mining)

*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 set de datos inicial, y consiste en actividades con la finalidad de familiarizarse con los datos, para identificar problemas de calidad de los datos, para descubrir una primera mirada sobre los datos o bien descubrir sub conjuntos 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 : Ciclo de vida de CRISP-DM (agregar referencia)



* 1. 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.

Tareas descriptivas

En este tipo de tareas el objetivo es, como dice su nombre, describir los datos existentes. Busca de proporcionar información entre las relaciones existentes entre 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[[10]](#footnote-10)*.

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[[11]](#footnote-11), 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:

* 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.

Tabla : Clasificación de reglas de asociación

|  |  |  |
| --- | --- | --- |
| Clasificación | Primer caso | Segundo caso |
| Tipos de valores que maneja la regla: booleana o cuantitativa | Asociación indica ausencia o presencia de elementos. | Asociación describe relaciones entre atributos cuantitativos. |
| Dimensiones de los datos involucrados: unidimensional o multidimensional | Referencia a una única dimensión. | Referencia a dos o más dimensiones. |
| Instantáneas o secuenciales | Indica relaciones inmediatas, relacionadas con el tiempo, contemporáneas | Establece un orden temporal en la relación |
| Positivas o negativas | Indica la ocurrencia o presencia de los elementos relacionados | Indica la ausencia de al menos uno de los elementos relacionados |

Además, la asociación se puede clasificar de acuerdo a los tipos de valores que maneja la regla, dimensiones de los datos involucrados, instantáneas o secuenciales y positivas o negativas. En la Tabla 1**,** se ilustran las diferentes clasificaciones de acuerdo a los criterios mencionados.

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[[12]](#footnote-12) 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[[13]](#footnote-13) 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 sub conjuntos, 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 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 Minkowksi:

* Distancia de Manhattan (r=1)

* Distancia Euclideana (r=2)

* Distancia de Chebyshev (

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

* Particionamiento: estos métodos construyen *k* particiones de un set de datos, representando cada una de estas un grupo. *K* debe ser menor o igual al número de elementos del set de datos mencionado. Ejemplos: *K-Means, K-Medoids.*
* Jerárquicos: como su nombre lo dice, genera una descomposición jerárquica del conjunto de grupos, en otras palabras, crea una serie de sub conjuntos de datos, en los que algunos engloban a otros. Ejemplos: *BIRCH, CURE*.
* Basados en densidad: se va creciendo el grupo (agregando elementos) hasta que la densidad (número de objetos) del mismo alcance algún umbral definido. 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: utilización de modelos hipotéticos para cada uno de los grupos, ajustando los datos a los 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 *Douglas M. Hawkins et al, 1980*.

“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'“*[[14]](#footnote-14)*

Tareas predictivas

Las tareas predictivas, son problemas en los que se hace necesario predecir un (o varios) valores para un set 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 predictivas de clasificación buscan definir un modelo que, como dice su nombre, sea capaz de predecir la clase de un objeto que no la tiene definida. A pesar de que sus objetivos son predecir la clase de un elemento en particular, pueden ser utilizadas para predecir 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.
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 decisión (o clasificación)

Los árboles de decisión son un método de clasificación en el cual se somete un dato a una seria de condiciones, que lo can clasificando 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 *divide & conquer[[15]](#footnote-15)*, 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 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 *Towell et al, 1990*. 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. Pueden ser del tipo *naive* (ingenuos), que 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 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.

* 1. 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[[16]](#footnote-16).(agregar referencia)
* *Angoss*: enfocada principalmente para organizaciones involucradas con ventas, marketing y análisis de riesgo, esta herramienta cuenta con una interfaz gráficas 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. Un plus de esta herramienta es que cuenta con una amplia gama de representaciones gráficas de datos. Si bien cuenta con una gama de implementaciones de los algoritmos más conocidos, no cuenta con suficientes herramientas para personalizar los procesos, por lo que no es la opción para quienes prefieran un ambiente fácilmente extensible. (agregar referencia)
* *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[[17]](#footnote-17)* 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. (agregar referencia)
* *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[[18]](#footnote-18), 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 en sí) para sacar el máximo provecha de esta herramienta. (agregar referencia)

1. 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.*

* 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 de interés mencionada (cantidad de usuarios activos) se encuentre en sus valores más bajos, para así poder evitarlos.

La finalidad detrás de buscar el aumento de usuarios activos en la plataforma recae en que, para los clientes del servicio, este 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 estos no son los que tienen valor para el cliente final, ya que estos 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.

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 de usuario**. Finalmente, los indicadores que serán de interés para este estudio se pueden apreciar de manera resumida en la Tabla 2.

Tabla : Indicadores de interés

|  |  |
| --- | --- |
| Indicador | Descripción |
| Usuarios activos | Hace referencia a la cantidad de usuarios activos (que han tenido actividad con la plataforma en la última semana) en un instante de tiempo. |
| Penetración | Hace referencia a la proporción de usuarios activos que comparten un video en particular. |
| Calidad de usuario | 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 3 y 4 las posibles variables de entrada, cuya variación podría afectar o no a los indicadores de interés.

Tabla : 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 | Género 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 : 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. |

* 1. 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 se puede apreciar en el **Anexo 1**. Cabe destacar que se incluyeron sólo las tablas y variables más significativas en el diagrama.

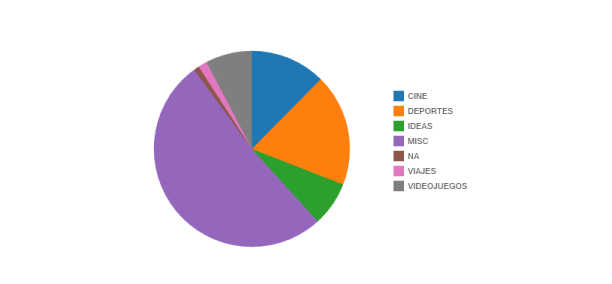
Una primera mirada a los datos deja en evidencia los siguientes problemas:

* Para el caso de datos de usuario, la variable **uni** (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%[[19]](#footnote-19) 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[[20]](#footnote-20) 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.

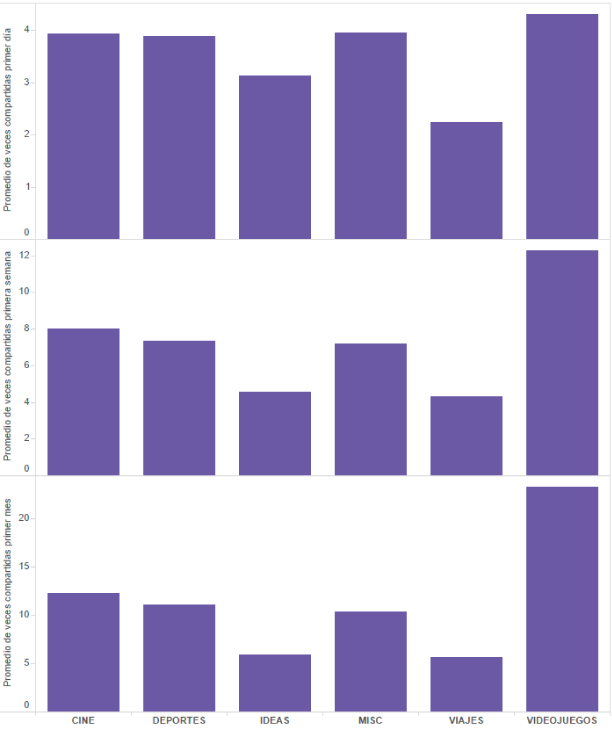
Ilustración : Distribución categorías de videos



En la Ilustración 3 se muestra la distribución de videos de acuerdo a su categoría. Aproximadamente un 50% de ellos pertenecen a la categoría *MISC*, que es una categoría relativamente general para categorizar a todos los videos que no pertenecen al resto de las categorías, 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.

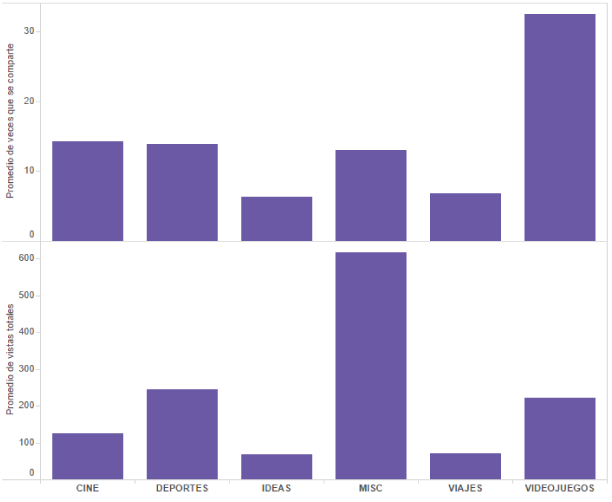
En la Ilustración 4 se muestra cómo se relaciona la categoría de un video con la cantidad de veces que este se comparte el primer día, la primera semana, y el primer mes (luego de su publicación). Para el primer caso, no se aprecia una diferencia muy significativa con respecto a las categorías. Dejando de lado la categoría de *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 y mes respectivamente, se aprecia que la categoría de *videojuegos* tiende a mantener su popularidad, mientras que el resto de las categorías parecen dejar de ser interesantes para el usuario.

Ilustración : Categoría de video vs promedio de veces que se comparte en intervalos



En base a esta potencial relación, se decide estudiar cómo se comportan al largo plazo cada una de las categorías 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 aprecia 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 hace ver que hay *outliers[[21]](#footnote-21)* 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 : Categoría de video vs promedio de vistas y veces que se comparte



En base al comportamiento no previsto de la Ilustración 4, se decide revisar si existe alguna relación entre la cantidad de vistas que consigue un 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.

Ilustración : Veces que se comparte vs veces que se ve

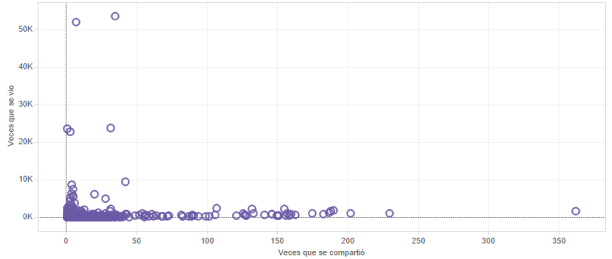
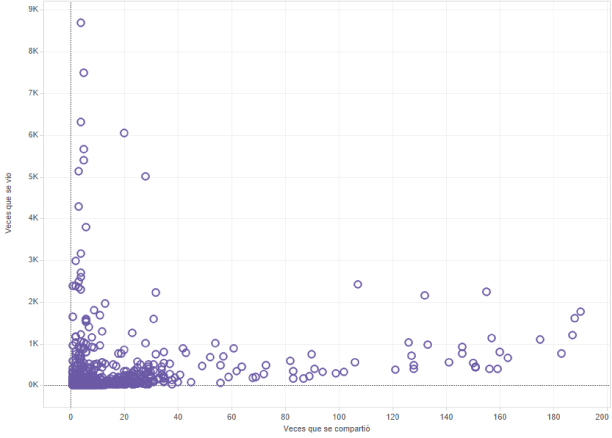


Ilustración : Veces que se comparte vs veces que se ve (filtrado)



Para el caso de usuarios, no se encuentra ninguna relación o distribución interesante de mencionar. Se hace necesario poblar variables nulas o vacías, por lo que se procede a la etapa de preparación de los datos.

1. Desarrollo de la solución
   1. Preparación de los datos

Durante esta etapa se realizaron 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 esto se utilizan las herramientas (*API*) de *Facebook* y de *YouTube* para buscar información faltante de los usuarios y videos respectivamente. Se automatiza el proceso utilizando los scripts presentes en el **Anexo 2**.

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 7 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 5 se muestran los criterios para asignación de clases de esta variable.

**Tabla 5: Calidad de usuario**

|  |  |  |  |
| --- | --- | --- | --- |
| Calidad | Descripción |  |  |
| No interesado,  No entendió | Usuario se registró en la plataforma, pero nunca tuvo interacción[[22]](#footnote-22) 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 constante | Usuario se registró en la plataforma e interactuó con ella diariamente, por un período mayor a 30 días. | | |
| Semanal mensual | Usuario se registró en la plataforma e interactuó con ella semanalmente (al menos), por un período de entre 1 y 4 semanas. | | |
| Semanal constante | Usuario se registró en la plataforma e interactuó con ella semanalmente (al menos), por un período mayor a 4 semanas. | | |

Habiéndose definido el primero de los indicadores clave para el negocio mencionados en la Tabla 2, en la Tabla 6 se explica cómo fueron calculados los dos restantes.

Tabla : Cálculo de indicadores clave

|  |  |
| --- | --- |
| Variable | Descripción y cálculo |
| active\_users | 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 (*Active Users: Measure the Success of Your Business,* *Claudiu Murariu, 2014)*. Se define en este caso en particular una interacción 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. |

Para poder tener un mejor entendimiento del grupo usuario, se definió y calculó una serie de variables que podrían ser significativas para explicar el comportamiento y calidad usuaria, listadas en la Tabla 7.

Tabla : Nuevas variables para usuarios

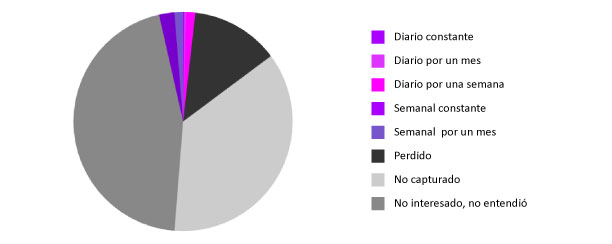
|  |  |
| --- | --- |
| Variable | Descripción |
| edad | Edad (en años) del usuario  Posibles valores: Número entero |
| sistema\_registro | Hace referencia al sistema a través del cual se registró el usuario  Posibles valores:  *Facebook campaign*: hace referencia a los usuarios que se registraron 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[[23]](#footnote-23).  *Through video display*: Registros a través de links dispuestos en la “vitrina”[[24]](#footnote-24). 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 |
| 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. Luego, se procede a una segunda etapa de visualización de datos, ahora con variables limpiadas, transformadas, y nuevas.

Visualización de datos tratados, usando Tableau (OLAP)

A continuación, 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 2.

Ilustración : Distribución calidad usuaria



En la Ilustración 8 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 contra la de calidad negativa es preocupante[[25]](#footnote-25). Se hace necesario entonces realizar una sobre-muestra (o sub-muestra) en caso de que se quiera utilizar técnicas de clasificación. 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).

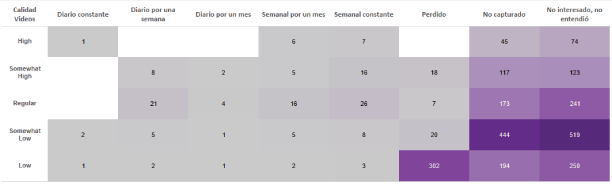
Para poder utilizar técnicas de clasificación sobre este indicador se hace necesario hacer *undersampling[[26]](#footnote-26)* o en su defecto *oversampling[[27]](#footnote-27)* de los datos. Se decide proceder con *oversampling*. Como criterio de creación de datos, se utilizan todos los valores existentes 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.

En la Ilustración 4 se comparan la calidad usuaria con la calidad de los videos (recientes) al momento que dichos usuarios se registraron. En la tabla, un color más saturado 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 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 : Calidad usuaria vs calidad de videos



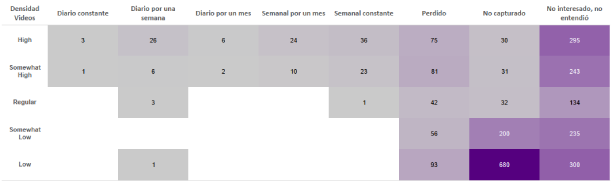
En el caso de los usuarios perdidos (sexta columna), para la gran mayoría de los casos la calidad de los videos al momento de registro es mala (*Low*). Además, para las clases de calidad usuaria “No capturado” y “No interesado, no entendió” (últimas 2 columnas) se ve que los valores más comunes corresponden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

Ilustración : Calidad usuaria vs densidad de concursos



En la Ilustración 5: Calidad usuaria vs densidad de concursos, siguiendo los niveles y propiedades 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.

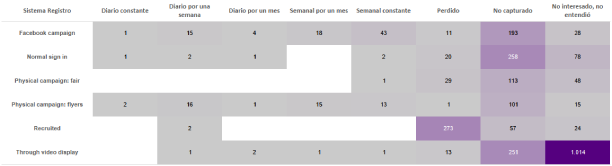
Ilustración : Calidad usuaria vs densidad de videos



En la Ilustración 6: Calidad usuaria vs densidad de videos 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.

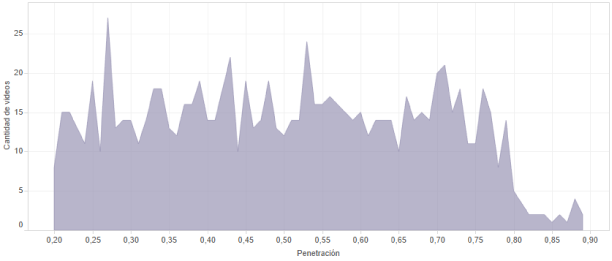
En la Ilustración 7 se aprecia la relación entre calidad usuaria y sistema de registro. Cabe destacar que, para esta variable, los valores de las clases no tienen ningún tipo de escala. Salta a la vista que para la clase “No interesado, no entendió” el sistema de registro más común es a través de la “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 : Calidad usuaria vs sistema de registro



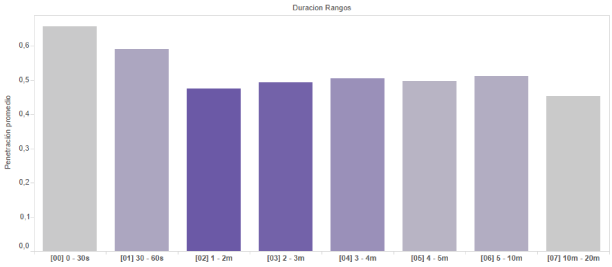
Hasta el momento se han comparado diferentes variables con la variable de interés **calidad usuaria**. A continuación, se realizarán comparaciones en búsqueda de patrones que pudiesen influir en las variables **penetración** y **usuarios activos**, en este mismo orden.

Ilustración : Distribución de videos de acuerdo a su penetración



En la Ilustración 8 se aprecia la distribución de valores presentes para penetración. En el eje horizontal se presentan los valores de penetración aproximada a 2 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).

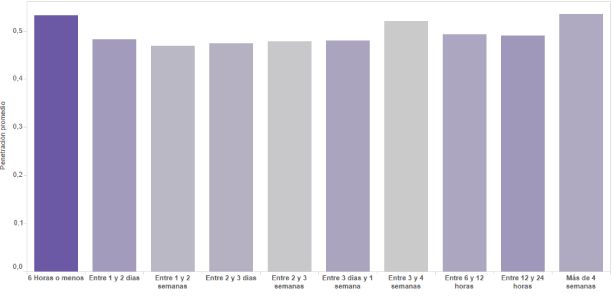
Ilustración : Penetración vs duración



En la Ilustración 9 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.

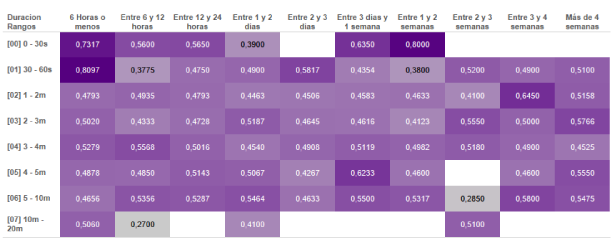
Se teoriza que es importante ser de las primeras plataformas en difundir un contenido en particular para conseguir un buen recibimiento del grupo usuario, traducido en una penetración alta. En la Ilustración 15, se busca entender una posible relación entre la diferencia de lanzamiento[[28]](#footnote-28) 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 : Penetración contra diferencia de lanzamiento



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

Ilustración : Relación penetración vs diferencia de lanzamiento y duración



En la Ilustración 16 se aprecia la exploración propuesta. En una primera interpretación pareciera que los videos en el primer rango de diferencia de lanzamiento y primeros rangos de duración alcanzan una mejor penetración. Llama la atención la penetración considerablemente marcada del primer rango de duración 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*. ­­

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[[29]](#footnote-29)). De esta base, se ve en las Ilustraciones 17 y 18, la relación entre la cantidad de usuarios activos promedio para estos instantes de tiempo en comparación a la cantidad de concursos[[30]](#footnote-30) y canjes[[31]](#footnote-31) activos respectivamente. En ambos casos se puede apreciar 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 : Usuarios activos vs concursos activos

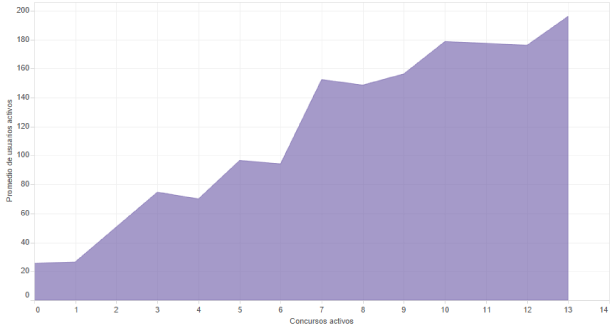
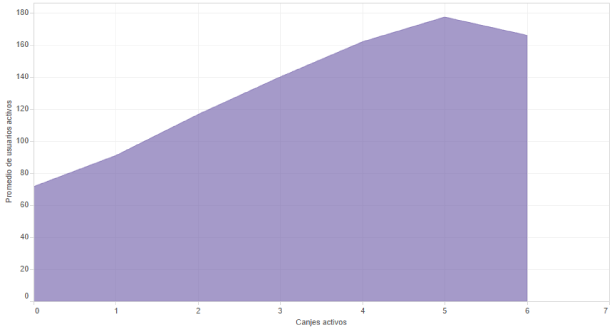


Ilustración : Usuarios activos vs canjes activos



En el siguiente capítulo 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 cono 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 8 se presentan las variables categorizadas, con sus valores posibles.

Tabla : Variables categorizadas para usuarios

|  |  |
| --- | --- |
| Variable | Cambio aplicado |
| puntos\_historicos | Se categoriza en diferentes rangos.  Posibles valores:  *“[ 0, 4651)”*  *“[ 4651, 18212)”*  *“[ 18212, 46406)”*  *“[ 46406,118093)”*  *“[118093,299581]”* |
| shares\_totales | Se categoriza en diferentes rangos.  Posibles valores:  *“[ 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.  Posibles valores:  *“1”*: El usuario reclutó 1 o más usuarios.  *“0”*: El usuario no reclutó a otros usuarios. |
| concursos\_participados | Se categoriza en diferentes rangos.  Posibles valores:  *“[ 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.  Posibles valores:  *“1”*: El usuario canjeó 1 o más premios.  *“0”*: El usuario no canjeó ningún premio. |
| tickets\_canjeados | Se categoriza en diferentes rangos.  Posibles valores:  *“[ 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.  Posibles valores:  *“[14.0,21.0)”*  *“[21.0,23.6)”*  *“[23.6,27.8)”*  *“[27.8,36.3)”*  *“[36.3,64.0]”* |

Tabla : Variables categorizadas para videos

|  |  |
| --- | --- |
| Variable | Cambio aplicado |
| active\_raffles | Se categoriza en diferentes rangos.  Posibles valores:  *“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.  Posibles valores:  *“[0, 40]”*  *“(40, 80]”*  *“(80, 120]”*  *“(120, 160]”*  *“Más de 160”* |
| duracion | Se categoriza en diferentes rangos.  Posibles valores (segundos):  *“30 o menos”*  *“(30, 60]”*  *“(60, 120]”*  *“(120, 180]”*  *“(180, 240]”*  *“(240, 300]”*  *“(300, 600]”*  *“600 o más”* |
| release\_difference | Se categoriza en diferentes rangos.  Posibles valores:  *“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.  Posibles valores:  *“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.  Posibles valores:  *“[ 1, 386)”*  *“[ 386, 1484)”*  *“[ 1484, 4417)”*  *“[ 4417,20842)”*  *“[20842,53557]”* |
| shares\_first\_day | Se categoriza en diferentes rangos.  Posibles valores:  *“[ 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.  Posibles valores:  *“[ 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.  Posibles valores:  *“[ 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.  Posibles valores:  *“[ 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.  Posibles valores:  *“0.000”*  *“[0.693,2.193)”*  *“[2.193,3.612)”*  *“[3.612,5.112)”*  *“[5.112,6.000]”* |

* 1. Modelo

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.

Reglas de asociación

A modo de primer acercamiento, se decide utilizar *A priori*, 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).

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 este. Entrando en detalle en estos registros, la cantidad total de datos de cada clase positiva de calidad usuaria se aprecia en la Tabla 10. De acuerdo a esta información, ninguna regla de asociación conseguida podría tener un soporte mayor al 2.25%. Se decide finalmente descartar el uso de *apriori* para encontrar reglas de asociación para estas clases.

Tabla : 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, se sabe que la cantidad que cuentan con calidades usuarias negativas es del 94.68%. En la Tabla 11 se presenta el total de registros de cada clase negativa.

Tabla : 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 a la clase *“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 12 se presentan las reglas encontradas para estas clases. Cabe destacar que estas 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 : Reglas de asociación para clases *Not captured* y *Lost*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Clase | Mano izquierda | Mano derecha | Soporte | Confianza |
| Not captured | densidad\_videos=Low | quality=Not captured | 25.48% | 63.31% |
| densidad\_concursos=Low | quality=Not captured | 26.26% | 63.44% |
| Lost | sistema\_registro=Recruited | quality=Lost | 10.29% | 76.69% |
| premios\_canjeados=0, sistema\_registro=Recruited | quality=Lost | 9.98% | 76.08% |
| recruitments=0,  sistema\_registro=Recruited | quality=Lost | 10.04% | 76.35% |
| recruitments=0,  premios\_canjeados=0,  sistema\_registro=Recruited | quality=Lost | 9.7% | 75.73% |

Estas reglas encontradas concuerdan con la información visualizada en la primera etapa de *CRISP-DM*.

Luego se procede a aplicar este algoritmo sobre la base de videos. De la misma forma que para la base de usuarios, se busca las reglas cuya mano derecha contenga valores relacionadas con los indicadores de interés. Lamentablemente, las reglas encontradas no aportan información extra para el estudio.

Técnicas de clasificación

En segunda instancia, se decide utilizar técnicas de clasificación, para así poder definir un modelo que se adecue a la mayor cantidad de calidades usuarias. Primero se desea probar con árboles de decisión, y se utilizan implementaciones de algoritmos de generación de árboles presentes en los paquetes *party* y *rpart*.

* 1. Evaluación

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.

* 1. 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.

1. Conclusiones

1. Referencias bibliográficas (completar con fechas de visita, ediciones de libros, años de publicación)

[1] ***UCLAAnderson*** *- school of management, http://www.anderson.ucla.edu*

[2] ***KDD, SEMMA AND CRISP-DM: A PARALLEL*** *OVERVIEW - Azevedo & Santos*

[3] ***Performance Dashboards: Measuring, Monitoring, and Managing your Business***

[4]***Advanced Tech Computing Group UTPL*** *- http://advancedtech.wordpress.com/*

[5] ***OLAP TOOLS*** *- http://www.informationbuilders.com/olap-online-analytical-processing-tools*

[6]***Identification of Outliers*** *-- Douglas M. Hawkins, 1980*

[7] ***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*

[9] ***Refinement of approximate domain theories by knowledge-based neural networks****, G. Towell, J. Shavlik, and M. Noordewier, 1990.*

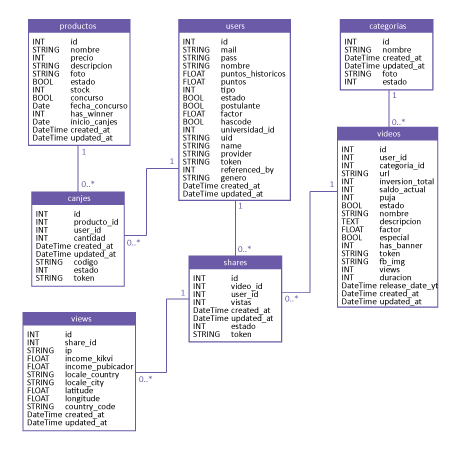
[10]***Rule extraction from artificial neural networks****, H. MAYER, Huber, Rohde, and Tamme, Universitsat Salzburg, 12th October 2006 2006.*

[11] ***Active Users: Measure the Success of Your Business****,* *Claudiu Murariu, 2014 - https://blog.innertrends.com/active-users-2/748 (visitado 3 Mayo 2016)*

1. Anexos (color: 6b59a6)

Anexo 1: Modelo de datos relacional

Se muestran a continuación las tablas relevantes al estudio, se omiten tablas intermedias en relaciones de muchos a muchos, tablas obsoletas, y extensiones de tablas.



Anexo 2: Scripts para APIs

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);

?>

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

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] [↑](#footnote-ref-1)
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] [↑](#footnote-ref-2)
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% (Información de red de *Display de Google*). Estudios de seguimiento de ojos (*eye-tracking*) demuestran a través de mapas de calor, que los usuarios de internet no fijan su atención en los banners, reconociéndolos inconscientemente como basura. [Fuentes: Gorka Garmendia (<http://www.gorkagarmendia.com/la-ceguera-del-banner-el-truco-de-mortadelo>), Red de *Display* de Google] [↑](#footnote-ref-3)
4. Hecho, número, o texto que puede ser procesado por un computador. [↑](#footnote-ref-4)
5. Los patrones, asociaciones o relaciones entre datos pueden generar información. A diferencia de los datos, la información tiene uso, utilizad. [↑](#footnote-ref-5)
6. La información puede ser transformada en conocimiento sobre patrones históricos o modas futuras. [↑](#footnote-ref-6)
7. *Knowledge discovery in databases* [↑](#footnote-ref-7)
8. ***Cr****oss-****I****ndustry* ***S****tandard* ***P****rocess for* ***D****ata* ***M****ining*, o en español, Proceso estándar multi-industria para minería de datos. [↑](#footnote-ref-8)
9. Fase *Business Understanding phase* (Fase de entendimiento del negocio) de *CRISP-DM*. [↑](#footnote-ref-9)
10. *http://www.tableausoftware.com* [↑](#footnote-ref-10)
11. Para más información sobre tareas supervisadas y no supervisadas refiérase al Anexo B: Técnicas supervisadas y no supervisadas [↑](#footnote-ref-11)
12. Medida para cuantificar los casos en los que e antecedente se hace verdadero. Puede ser número de casos o porcentaje. [↑](#footnote-ref-12)
13. Número de casos en que, habiéndose cumplido el antecedente de la regla, se cumple el consecuente. [↑](#footnote-ref-13)
14. Traducción: "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'" [↑](#footnote-ref-14)
15. Dividir y conquistar. [↑](#footnote-ref-15)
16. Para más información sobre la licencia de tipo AGPL, referirse a http://www.gnu.org/licenses/agpl-3.0.html. [↑](#footnote-ref-16)
17. *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. [↑](#footnote-ref-17)
18. Más información sobre licencia de R: https://www.r-project.org/COPYING [↑](#footnote-ref-18)
19. 2277 Registros, de un total de 2669, tiene el campo ***uni*** con valor faltante, nulo, o vacío. [↑](#footnote-ref-19)
20. Otras variables: *difference\_last\_and\_first\_share, difference\_last\_raffle\_first\_share* [↑](#footnote-ref-20)
21. En este caso, un video con más de 50.000 vistas, y tres videos con más de 20.000. [↑](#footnote-ref-21)
22. 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. [↑](#footnote-ref-22)
23. 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. [↑](#footnote-ref-23)
24. Vitrina es el nombre utilizado para referirse a la página en la que se muestra un video. [↑](#footnote-ref-24)
25. De forma más exacta, hay 2527 (94.68%) de registros de usuario negativos contra 142 (5.32%) positivos. [↑](#footnote-ref-25)
26. 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. [↑](#footnote-ref-26)
27. 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 similiar de ocurrencias. [↑](#footnote-ref-27)
28. Diferencia de tiempo entre que el contenido es publicado en YouTube en comparación a su publicación en la plataforma. [↑](#footnote-ref-28)
29. Se publican videos repetidas veces a la semana, sin dejar pasar múltiples días sin nuevos videos (excepto en excepciones pequeñas) [↑](#footnote-ref-29)
30. 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. [↑](#footnote-ref-30)
31. 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 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. [↑](#footnote-ref-31)