1. Introducción
   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 han 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 Kikvi 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 doble, consistente en 2 grupos principales: los **publicadores**, y los **creadores**.

La idea era la siguiente: el grupo de **creadores** consistía en 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 adeptas al 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. El modelo de negocios sería el siguiente: 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).

En a estas alturas que 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 2 aristas con las que se estaba teniendo problemas:

* Lo 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.

Hoy se está a puertas de un nuevo lanzamiento, una nueva iteración en la historia del producto, y se espera tomar las decisiones correctas apoyándose en lo que se espera extraer de los datos que se han recolectado hasta la fecha. Este nuevo lanzamiento será complementado con el desarrollo de aplicaciones móviles y una nueva imagen de producto (rediseño de logos, paletas de colores y plataforma web), estos puntos serán abordados más a fondo más adelante en este estudio.

* 1. Objetivos
     1. 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).

* + 1. 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 medir su variación al tomar acciones.
* Mejorar la percepción de usuarios enfocando esfuerzos a áreas de interés, al desarrollar un entendimiento de qué es lo que más usan y menos usan de la plataforma.
* 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 .

1. Estado del arte
   1. Minería de datos

De forma general, la minería de datos consiste en el proceso de analizar datos[[4]](#footnote-4) de diferentes fuentes, desde diferentes perspectivas, con el fin de resumirla en información[[5]](#footnote-5) útil, o sea, información que pueda ser utilizada para aumentar ganancias, disminuir costos, 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 de 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 determinar 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 elementos principales:

1. Extraer, transformar, y cargar datos en el *data warehouse*.
2. Almacenar y administrar los datos en un sistema de bases de datos multidimensional.
3. Dar acceso a los datos a analistas del negocio y profesionales de TI.
4. Analizar los datos con aplicaciones especializadas.
5. Presentar la información en formatos útiles, como gráficos o tablas.

En relación a los niveles de análisis del paso número 4. , 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 pensamiento a través de un set de decisiones. Estas decisiones generan reglas de clasificación para un set 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**: Extracción de reglas *if-then* útiles, basadas en significancia estadísticas.
* **Visualización de datos**: Interpretación visual de relaciones complejas en datos multidimensionales. Se usan herramientas gráficas para ilustrar relaciones en los datos.

En la sección **2.1.2.2.1** - **Clasificación**, se ahondará más en estos métodos.

* + 1. Procesos de minería de datos
       1. 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. (**Referencia 1** - Curso Bussiness Inteligence -- UTFSM)

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.

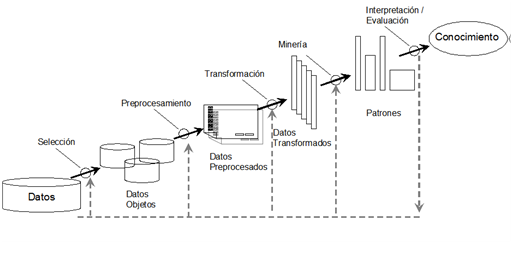
El proceso cuenta con 5 etapas definidas:

1. **Selección**: Esta etapa consiste en crear un set de datos, o enfocar en un sub conjunto de variables o muestras de datos, sobre la cual se realizará el descubrimiento
2. **Pre-proceso**: Proceso de limpieza del set de datos seleccionado con el fin de obtener datos consistentes.
3. **Transformación**: En esta etapa, como dice su nombre, se transforman los datos utilizando *dimensionality reduction[[8]](#footnote-8)* o métodos de transformación.
4. **Minería de datos**: Esta etapa consiste en la búsqueda de patrones de interés en alguna forma particular de representación, dependiendo del objetivo de la minería. Comúnmente se busca predicción.
5. **Interpretación/Evaluación**: En esta etapa final, se interpretan y evalúan los patrones encontrados.

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.

De forma gráfica, en la **Ilustración 1** se aprecian los 5 pasos del KDD:

Ilustración 1: Proceso del descubrimiento del conocimiento (KDD)



* + - 1. 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: La primera fase del proceso, 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: Esta fase consiste en la visualización 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 se enfoca únicamente al modelo en proyectos de minería de datos, dejando los aspectos propios del negocio afuera, a diferencia de otros procesos como *CRIPS-DM[[9]](#footnote-9)*, que cuentan con fases[[10]](#footnote-10) enfocadas e estos aspectos.

* + - 1. 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**: Aplicación de varias técnicas de modelo, calibrando sus parámetros a valores óptimos.
5. **Evaluación**: En esta etapa 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 el ciclo de vida de *CRISP-DM*.

Ilustración 2: Ciclo de vida de CRISP-DM



* + 1. Tareas de minería de datos
       1. 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.

Para los objetivos de este estudio, este tipo de tareas es de especial interés ya que ayudará a establecer patrones y características que definan a un estudiante de excelencia.

* + - * 1. 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[[11]](#footnote-11)*.

* + - * 1. 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[[12]](#footnote-12), 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 solo pueden aplicarse sobre variables nominales (todas las involucradas). La asociación puede presentarse de dos maneras:

1. 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.
2. Dependencias: A diferencia del caso anterior, este tipo de asociaciones son direccionales, o sea que el cumplimiento de una serie de condiciones implican que se cumplirán otras, y no al revés.

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: Clasificación de reglas de asociación,** se ilustran las diferentes clasificaciones de acuerdo a los criterios mencionados.

Tabla 1: 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 |

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 set de ítems frecuentes con un soporte[[13]](#footnote-13) mayor o igual al deseado , o sea, que se buscan conjuntos de elementos que cuenten con cierto criterio establecido, sin separarlos aún. Luego, en la segunda fase, se hacen particiones de los set de ítems, calculando la confianza[[14]](#footnote-14) de cada una, y reteniendo las reglas que tengan confianza mayor o igual a la deseada.

Segmentación (Agrupamiento)

Esta tarea consiste 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 tipo. A continuación se listan algunos de ellos:

* 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)
  + SOM, mapas auto organizados (redes neuronales)
    - * 1. Detección de anomalias

La detección de valores e instancias anómalas es una tarea necesaria al momento de realizar minería de datos. En todo set 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* (**Referencia 2** - Identification of Outliers -- Douglas M. Hawkins, 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'"[[15]](#footnote-15)*

* + - 1. Tareas predictivas

Las tareas predictivas, como dice su nombre, 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.

Las tareas predictivas de pueden clasificar en:

* Tareas de clasificación.
* Regresión estadística.
* Pronósticos.
  + - * 1. 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 sub conjunto de entrenamiento para la construcción del modelo.
3. Utilización del sub conjunto 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 predictiva.

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[[16]](#footnote-16)*, empezando con todos los elementos del grupo de entrenamiento en la raíz, y continuando dividiéndolos en el atributo que se elija para ramificarlo.

Inducción de reglas (reglas de clasificación)

Los métodos de inducción de reglas tiene 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* (1990) (**Referencia 5** - G. Towell, J. Shavlik, and M. Noordewier, Refinement of approximate domain theories by knowledge-based neural networks", 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

Los métodos bayesianos son herramientas estadísticas 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 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. Estos métodos se basan en el teorema de Bayes[[17]](#footnote-17)

Métodos basados en casos y vecindad

Estos métodos 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 elobjetivo de conseguir una mejor precisión final en el clasificador

* + - * 1. 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. Dashboards

Los *Dashboards* o *pizarras* son despliegues de información pertinente para quien la consulta. En particular, hay un tipo de *dashboard* que es de importancia para este estudio, los *performance dashboards*. Este tipo de despliegue de información tiene la característica de que muestra información crítica relacionada con el desempeño de la actividad de una empresa, organización o entidad, haciéndola ideal para hacer seguimientos.

*Eckerson*, autor del libro *“Performance Dashboards: Measuring, Monitoring, and Managing your Business”*, pone énfasis en que es de vital importancia para una empresa el dirigir el foco del esfuerzo, para sacarle el mejor provecho posible, y es por esto que los *performance dashboards* son útiles para este fin, ya que es de vital importancia el enfocar bien los esfuerzos para poder lograr una educación de excelencia, y por lo tanto, estudiantes y luego profesionales de excelencia.

A través de un *dashboard*, se pueden transformar estrategias, en objetivos concretos, métricas valiosas, iniciativas y tareas. Parte de los beneficios que se pueden conseguir de un buen despliegue de información de desempeño son:

* Comunicación de estrategia.
* Refinación de estrategia.
* Visibilidad (de situaciones que no se aprecian de otras maneras).
* Coordinación.
* Motivación.
* Focalizar esfuerzos

Para poder construir un *performance dashboard* efectivo, se hace necesaria la identificación de los indicadores clave de desempeño (KPI), definidos como:

*"Los indicadores clave de desempeño son métricas financieras o no financieras, utilizadas para cuantificar objetivos que reflejan el rendimiento de una organización, y que generalmente se recogen en su plan estratégico. Estos indicadores son utilizados en inteligencia de negocio para asistir o ayudar al estado actual de un negocio a prescribir una línea de acción futura"*

-- (**Referencia 3** - 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)

1. Desarrollo

La primera pregunta que se hace necesario responder antes de poder identificar a un estudiante de excelencia es, justamente:

**¿Qué define a un estudiante de excelencia?**

Hay discusión en torno a esta pregunta, en particular relacionada con la disputa de si la excelencia viene dada solo por un rendimiento académico resaltado, o si bien se ve complementada con la integridad como estudiante, o sea, su participación en diferentes áreas de disciplina, además de su enfoque principal, que en este caso particular es su carrera de estudios superiores.

Lamentablemente, la información con la que se cuenta es casi absolutamente académica, por lo que no será posible ahondar sobre esta discusión. En base a esto, se deciden tomar en consideración la siguientes variables académicas, que se detallarán a continuación:

* Promedio ponderado
* Retención
* Titulación
* Reprobación (tasa de reprobación)

Estas variables no se encuentran especificadas explícitamente en los datos facilitados, por lo que serán calculadas en este estudio.

* 1. Preparación de los datos

Previo a realizar minería de datos, es necesario preparar la información facilitada, con el fin de definir y calcular variables de interés, dejar a un lado los datos que no tienen valor para el estudio.

* + 1. Variables de interés
       1. Promedio ponderado

El promedio ponderado de un estudiante refleja qué tan buenas fueron las clasificaciones del mismo en un período de tiempo. Para objetivos de este estudio, se considerará 1 año como período, por lo que el promedio ponderado estará dado por:

Donde es el promedio ponderado del estudiante, es la nota obtenida por el estudiante en el semestre ***i***del año en cuestión, y son la cantidad de créditos que el estudiante cursó en el semestre ***i*** del año en cuestión.

* + - 1. Retención

La **retención** será una variable booleana[[18]](#footnote-18) que reflejará si un estudiante se encuentra '*retenido'* o no en un semestre o año en particular. Un estudiante *'retenido'* será aquel que cuente con ramos inscritos el semestre siguiente al cual se está estudiando si retención, o sea:

Donde es la variable booleana en cuestión, relacionada con el semestre ***i***, y la cantidad de ramos inscritos por el estudiante en el semestre ***i+1***.

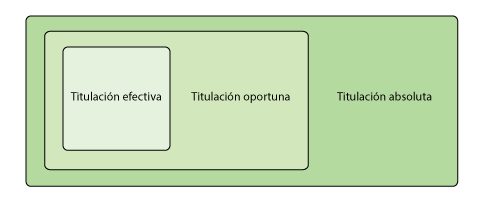
* + - 1. Titulación

La **titulación**  será una variable booleana que reflejará si un estudiante está titulado o no. En base a esto, se consideran 3 tipos de titulación:

* **Titulación efectiva**: Un estudiante entrará en esta categoría si se titula en el tiempo de duración de la carrera que estudia o menos.
* **Titulación oportuna**: Un estudiante entrará en esta categoría si se titula en el tiempo de duración de su carrera **más un año** o menos. Es decir, si un estudiante que estudia una carrera de 6 años se titula en 7 o menos (incluyendo 6 años), entrará en esta categoría.
* **Titulación absoluta**: Un estudiante entrara en esta categoría si cursa la totalidad de su carrera y logra titularse, independiente del tiempo que le tome.

De las descripciones anteriores se desprende que las categorías no son independientes entre sí, la **Ilustración 3: Categorías de titulación,** expresa gráficamente los alcances de las categorías mencionadas.

Ilustración 3: Categorías de titulación



* + - 1. Reprobación (tasa de reprobación)

La **reprobación**, o **tasa de reprobación** es una variable que refleja la proporción entre los ramos cursados por un estudiante y los ramos reprobados por el mismo. De forma matemática:

Donde será la reprobación, o tasa de reprobación, de un estudiante dado en el semestre o año ***i***, será la cantidad de ramos reprobados por el estudiante en el semestre o año ***i***, y finalmente será la cantidad total de ramos cursados por el estudiante en el semestre o año ***i***.

* + 1. Limpieza de datos sin valor para el estudio

Debido a la naturaleza genérica de la información facilitada, hay una gran cantidad de esta que no cuentan con valor para este estudio. El criterio principal para filtrar los datos que no serán considerados en las variables de interés se aprecia en la **columna 9 del rendimiento para todos los cohortes: estado actual del estudiante**, en la que se detalla a través de una serie de estados definidos la situación actual del alumno (para más detalles, refiérase a **Anexo C: Diccionario de datos**). En base al valor de esta columna, se consideraron filtros de acuerdo a **Tabla 2: Filtros aplicados a variables de interés**, para las variables de interés, donde un **√** en una casilla indica que un alumno con ese estado será considerado para el cálculo de la variable de esa columna.

Tabla 2: Filtros aplicados a variables de interés

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Estado / Variable | Retención | Titulación | Reprobación | PP |
| abandono por cambio de carrera |  |  |  |  |
| cambio de carrera |  |  |  |  |
| eliminado causal academica | **√** | **√** | **√** | **√** |
| eliminado no titulacion | **√** | **√** | **√** | **√** |
| eliminado por no matricula | **√** | **√** |  |  |
| fallecimiento |  |  |  |  |
| memorista | **√** | **√** | **√** | **√** |
| no matriculado | **√** | **√** |  |  |
| postergacion c/ expr. De causa | **√** | **√** |  |  |
| postergacion de estudios | **√** | **√** |  |  |
| Recons. Causal acedemica | **√** | **√** | **√** | **√** |
| reincorporacion decreto 2609 | **√** | **√** | **√** | **√** |
| reincorporacion postergacion | **√** | **√** | **√** | **√** |
| reincorporacion R.T. C/Expr de causa | **√** | **√** | **√** | **√** |
| reincorporacion renuncia | **√** | **√** | **√** | **√** |
| reincorporacion Post C/Expr de causa | **√** | **√** | **√** | **√** |
| reincorporacion por eliminacion art eliminado no titulacion | **√** | **√** | **√** | **√** |
| reincorporado | **√** | **√** | **√** | **√** |
| renuncia voluntario nueva conc. Sem. | **√** | **√** | **√** | **√** |
| renuncia voluntaria nuevo ex. Arancel | **√** | **√** | **√** | **√** |
| renuncia carrera (antiguos) | **√** | **√** | **√** | **√** |
| renuncia USACH (nuevos) | **√** | **√** | **√** | **√** |
| retiro temporal c/expr de causa | **√** | **√** |  |  |
| retiro temporal | **√** | **√** |  |  |
| sin variacion | **√** | **√** | **√** | **√** |
| titulado | **√** | **√** | **√** | **√** |

Además se representan en la **Tabla 3: Filtros extra aplicados a variables de interés**, las consideraciones en base a qué datos e intervalos de tiempo considerar relevantes para el estudio, en base a recomendaciones sugeridas por vice rectoría académica de la Universidad de Santiago.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Estado / Variable | Retención | Titulación | Reprobación | PP |
| PP menores (o iguales) a 1,2 | **√** | **√** | **√** |  |
| Alumnos que, por tiempo, no se han titulado (no se ha cumplido el tiempo de su carrera) | **√** |  | **√** | **√** |
| Contar si no tienen ramos inscritos en primer semestre del año |  |  |  |  |
| Alumnos de bachillerato |  |  | **√** | **√** |

Tabla 3: Filtros extra aplicados a variables de interés

Dentro de las razones no evidentes para la aplicación de estos filtros destacan (en orden relativo a las tablas **Tabla 2** y **Tabla 3**):

Para **Tabla 2: Filtros aplicados a variables de interés**:

* No se consideran para ninguna variable los alumnos que se han **cambiado de carrera**, debido a que la gran mayoría de estos casos es reflejo de estudiantes que entraron a una carrera por imposibilidad de entrar a la que realmente estaban buscando. Los estados de cambio de carrera se asignan a alumnos que se han cambiado dentro de la misma universidad.
* No se considerará a los estudiantes **eliminados por no matrícula** para los cálculos de las variables de reprobación y promedio ponderado ya que este estado refleja un abandono voluntario por parte del alumno, que en repetidas ocasiones ocurre a mitad de año o de semestre, lo que podría perjudicar la significancia de las variables, ya que al abandonar los ramos tendrá una reprobación del 100%, o un promedio ponderado de 0 (o muy bajo si es que el estudiante rindió algún examen antes de abandonar sus estudios).
* No se considerarán a los estudiantes **no matriculados** por la misma razón que el punto anterior (para este estudio, es una redundancia en los datos facilitados).
* Los alumnos con **postergación de estudios** no serán considerados en la variables de reprobación ni de promedio ponderado ya que debido al abandono temporal, su información se encuentra incompleta.
* Los alumnos con **retiro temporal** no serán considerados en el cálculo de las variables de reprobación ni de promedio ponderado por la misma razón descrita en el punto anterior.

Para **Tabla 3: Filtros extra aplicados a variables de interés**:

* No se considerará para la interpretación de los resultados de la variable promedio ponderado a aquellos estudiantes que cuenten con un valor menor o igual a 1,2, ya que en la gran mayoría de los casos, estos promedios se dan por abandono no documentado de la carrera, y no por desempeño académico real, que es el enfoque buscado en este estudio.
* No se considerarán los alumnos que **no se han titulado por tiempo** para el cálculo de la variable 'titulación', ya que esto quiere decir que, en el momento en que fueron documentados los datos, no se ha cumplido el tiempo total de duración de la carrera. Por ejemplo, no se considerará como 'no titulado' a un estudiante de la carrera de ingeniería civil que haya comenzado sus estudios el año 2008[[19]](#footnote-19).
* No se contarán para ningún cálculo ni interpretación de este estudio aquellos estudiantes que **no cuenten con ramos inscritos el primer semestre del primer año** de estudios, ya que esto significa que dichos alumnos nunca comenzaron sus estudios, por razones desconocidas pero ejemplificadles, como cambio de preferencia de estudios, causas monetarias, fallecimientos no documentados, etc.
* No se contarán los **alumnos de bachillerato** para el cálculo de retención ni de titulación, debido a la naturaleza transitiva de la carrera.
  + 1. Cálculo de variables de interés
       1. Plataforma

Para el cálculo de variables de interés se utilizó una plataforma creada para el estudio, desarrollada en ***ruby on rails***[[20]](#footnote-20) por las siguientes razones:

* Al tratarse de un *framework* que favorece convención sobre configuración (*convention over configuration*), el desarrollo de aplicaciones pequeñas con manejo de base de datos es expedito. La rapidez de desarrollo va de la mano con la generación completa de *scaffolds*, que son un modelo, vista y controlador relacionados, creados solo con un comando, incluyendo vistas y funciones de creación/eliminación/lista, migraciones para bases de datos y elementos de *test* entre otros.
* Alta disponibilidad de *pluguins* y *scripts*, llamados *gems*, o gemas en español. Aportadas por la gran comunidad usuaria, las gemas pueden cumplir funciones desde el manejo de imágenes a adaptación para dispositivos móviles. En el caso particular de este estudio, se utilizarán gemas para el (NOMBRAR GEMA) manejo y importación de archivos de Excel (extensiones .csv, .xlsx), ya que la información facilitada se encuentra en estos formatos.
* Es una herramienta muy útil y completa para desarrollar aplicaciones web dinámicas, basadas en bases de datos, lo que resulta muy ventajoso para el desarrollo de la plataforma de análisis al final de este estudio.
* Hay una comunidad creciente muy numerosa en torno al lenguaje, lo que da como resultado una documentación muy detallada y extensa.
* Quien desarrolla este estudio cuenta con años de experiencia en el uso del lenguaje, lo que tendrá como resultado una plataforma estable, bien programada y rápidamente desarrollada.
* *Ruby on Rails* es un lenguaje de alto nivel, y madurado, actualmente en su versión 4, ya se trata de un lenguaje "probado" como efectivo.
* Debido a la base completa del *framework*, el uso de *Ruby on Rails* es una perfecta ilustración del caso de "no reinventar la rueda".
  + - 1. Cálculo y limpieza

La necesidad de cálculo de variables, junto con la limpieza de los datos, se realizó a través de los siguientes pasos:

* **Puesta en marcha de la plataforma**: Se genero una plataforma en blanco, a través de los comandos de RoR[[21]](#footnote-21), para luego generar los siguientes modelos:

|  |  |
| --- | --- |
| Modelo: Estudiante | |
| variable | **tipo** |
| id[[22]](#footnote-22) | integer |
| comuna\_id | integer |
| dependencia\_colegio\_id | integer |
| ingreso\_bruto\_id | integer |
| estado\_id | integer |
| ano\_ingreso | integer |
| facultad | string |
| codigo\_carrera | integer |
| nem | float |
| psu\_lenguaje | float |
| psu\_matematicas | float |
| ranking | float |
| duracion\_carrera | integer |
| nombre\_carrera | string |
| ano\_titulacion | integer |
| titulado\_efectivo | boolean |
| titulado\_oportuno | boolean |
| titulado\_absoluto | boolean |

|  |  |
| --- | --- |
| Modelo: Comuna | |
| variable | **tipo** |
| id | integer |
| nombre | string |

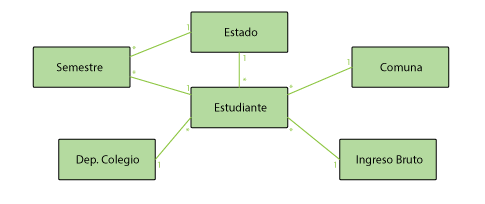
|  |  |
| --- | --- |
| Modelo: Semestre | |
| variable | **tipo** |
| id | integer |
| estado\_id | integer |
| estudiante\_id | integer |
| numero\_semestre | integer |
| asignaturas\_inscritas | integer |
| asignaturas\_aprobadas | integer |
| asignaturas\_reprobadas | integer |
| promedio\_semestre | float |
| promedio\_aprobadas | float |

|  |  |
| --- | --- |
| Modelo: Dependencia Colegio | |
| variable | **tipo** |
| id | integer |
| dependencia | string |

|  |  |
| --- | --- |
| Modelo: Ingreso Bruto | |
| variable | **tipo** |
| id | integer |
| rango | string |

|  |  |
| --- | --- |
| Modelo: Estado | |
| variable | **tipo** |
| id | integer |
| estado | string |

Que se relacionan de la siguiente forma:



* **Importación de datos para poblar tablas (ETL)[[23]](#footnote-23)**: El siguiente paso es la importación de los datos facilitados en formato Excel para poblar las tablas. Se hace necesario estandarizar los datos antes de ser importados, debido a la gran cantidad de incongruencias que presentan, entre las que destacan:
  + Poca estandarización de columna 'dependencia colegio' (Part. Pagado = P PAGADO = Par pagado, etc)
  + Falta de datos esencialmente numéricos (como número de ramos inscritos) es solucionado en los datos con instancias como "N/A" o "S/I", lo que resultaría en error al momento de importar los datos como numéricos.

Una vez estandarizados los datos, se prosigue con la importación (**extract**), la que se realiza utilizando la gema (GEMA), el proceso tardó (TIEMPO). Con los datos en memoria, se procede al proceso de transformación (**transform**), el que se realiza utilizando los algoritmos dispuestos en el (ANEXO).

Finalmente, con los datos transformados, se finaliza el proceso cargando (**load**) la información a la base de datos descrita en el punto anterior.

* + 1. Observaciones
  1. Minería de datos
     1. Descripción de herramientas utilizadas
        1. Plataforma
        2. Algoritmos
     2. Detalles del proceso de minería de datos
     3. Resultados de minería de datos
     4. Observaciones

1. Bibliografía

[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/

1. Referencias

Referencia 1 - Curso Bussiness Inteligence -- UTFSM

Referencia 2 - Identification of Outliers -- Douglas M. Hawkins, 1980

Referencia 3 - 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

Referencia 4 - KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW - Azevedo & Santos

Referencia 5 - G. Towell, J. Shavlik, and M. Noordewier, Refinement of approximate domain theories by knowledge-based neural networks", 1990.

Referencia 6 - H. MAYER, Huber, Rohde, and Tamme, Rule extraction from artificial neural networks", Universitsat Salzburg, 12th October 2006 2006.

1. Anexos

Anexo A: Cálculo del RANKING USACH

El RANKING utilizado en este estudio hace referencia a Ranking USACH, y no el Ranking MINEDUC. El Ranking USACH para un estudiante perteneciente a un colegio en particular se calcula de la siguiente manera:



Donde:

|  |  |
| --- | --- |
|  | Nota del estudiante. |
|  | Nota mínima del colegio. |
|  | Nota máxima del colegio |
|  | Nota promedio del colegio |

Anexo B: Técnicas supervisadas y no supervisadas

Dentro del aprendizaje automático, existen técnicas de clasificación, las que se pueden dividir en técnicas supervisadas y no supervisadas. Su principal diferencia radica en que las técnicas supervisadas cuentan con un conocimiento a priori, o sea, que para clasificar un objeto dentro de una clase se cuenta con modelos ya clasificados, a diferencia de las técnicas no supervisadas, en las que no se cuenta con conocimientos a priori, por lo que se cuenta con un área de entrenamiento disponible para la tarea de clasificación. En las técnicas no supervisadas no se cuenta con grupos de objetos clasificados en clases por sus características similares (como en el caso de las técnicas supervisadas), sino que se deben encontrar las clases que describan al grupo de objetos, agrupando a aquellos con características similares en la misma clase.

Anexo C: Diccionario de datos

El marco muestral de estudio corresponde a los estudiantes de pregrado de la Universidad de Santiago de Chile, de la totalidad de las carreras y los cohortes 2002, 2003, 2004, 2005, 2006, 2007 y 2008. Se cuenta con aproximadamente 23000 registros de alumnos, con datos correspondientes a **antecedentes** y **OTRO NOMBRE**

Antecedentes para las cohortes de ingreso 2002, 2003

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | VARIABLE | ETIQUETA | TIPO | DETALLE |
| 1 | RUT | RUT | VARCHAR(8) | Rol Único Nacional |
| 2 | Dígito Verificador | DV | VARCHAR(1) | Dígito verificador RUT. |
| 3 | Nombre | NOMBRE | VARCHAR(26) | Primer y Segundo nombre (según corresponda) |
| 4 | Apellido Paterno | APELLIDO PATERNO | VARCHAR(16) |  |
| 5 | Apellido Materno | APELLIDO MATERNO | VARCHAR(16) |  |
| 6 | Género | GENERO | VARCHAR(9) | Femenino/Masculino |
| 7 | Código de colegio | CODIGO COLEGIO | VARCHAR(4) |  |
| 8 | Nombre de colegio | COLEGIO | VARCHAR(45) | Nombre del establecimiento |
| 9 | Dependencia de colegio | DEPENDENCIA COLEGIO | VARCHAR(15) | Relativo al financiamiento: Municipal / P Pagado/ P Subvencionado |
| 10 | Rama educacional | RAMA EDUCACIONAL COLEGIO | VARCHAR(2) | Relativo a modalidad educacional:  H1: Humanista Científico Diurno  H2: Humanista Científico Nocturno  T1: Técnico Profesional Comercial  T2: Técnico Profesional Industrial  T3: Técnico Profesional Servicios y Técnica  T4: Técnico Profesional Agrícola  T5: Técnico Profesional Marítima |
| 11 | IVE del Colegio | IVE COLEGIO 2011 | FLOAT | Índice de Vulnerabilidad por colegio |
| 12 | Comuna | COMUNA | VARCHAR(25) | Comuna de origen |
| 13 | Nivel educacional de la madre | NIVEL EDUCACIONAL DE LA MADRE | INTEGER | Definido según:  0: S/I  1:Sin estudios  2: Básica incompleta  3: Básica completa  4: Media incompleta  5: Media completa  6: Centro de Formación Técnica incompleta  7: Centro de formación Técnica completa  8: Universitaria incompleta  9: Universitaria completa  10: Otros estudios |
| 14 | Nivel educacional del padre | NIVEL EDUCACIONAL DEL PADRE | INTEGER | Definido según:  0: S/I  1:Sin estudios  2: Básica incompleta  3: Básica completa  4: Media incompleta  5: Media completa  6: Centro de Formación Técnica incompleta  7: Centro de formación Técnica completa  8: Universitaria incompleta  9: Universitaria completa  10: Otros estudios |
| 15 | Nacionalidad | NACIONALIDAD | VARCHAR(10) |  |
| 16 | Integrantes del grupo familiar | GRUPO FAMILIAR | INTEGER | Cantidad de personas del grupo familiar. |
| 17 | Ingreso Bruto Grupo Familiar | INGRESO BRUTO | INTEGER | Definido según:  1: 0 - 278  2: 278,001 - 834  3: 834,001 - 1.400.000  4: 1.400.001 - 1.950.000  5: 1.950.001 - 2.500.000  6: 2.500.001 - O más |
| 18 | Vía de ingreso | VIA DE INGRESO | VARCHAR() | C.Dep; C. Ind.; C. Tel.; C. USACH; Cupo Ind; DepDest; Hi jo Funcionario USACH; Minusválido; Of. U Chile; PAA; BEA; Indígena; LAE; List. Adic.; Oficios Varios; Otros; Repostulación; Supernumerario; Teletón; Unesco;PSU |
| 19 | Año de Egreso Enseñanza Media | AÑO EGRESO ENSENANZA MEDIA |  | Año de egreso de enseñanza media |
| 20 | Promedio de Notas de la Enseñanza Media | PROMEDIO NOTAS ENSEÑANZA MEDIA | FLOAT | Promedio de notas de los 4 años de enseñanza media. |
| 21 | Notas Enseñanza Media | NEM | INTEGER | Escala de puntaje que corresponde al promedio de todos los promedios anuales de cada asignatura cursada durante los cuatro años de enseñanza media |
| 22 | PAA Matemática | MAT | INTEGER | Prueba de Aptitud Matemática |
| 23 | PAA Lenguaje | LENG | INTEGER | Prueba de Aptitud Lenguaje |
| 24 | PAA Historia | PCE HISTORIA | INTEGER | Prueba de conocimiento específico Historia |
| 25 | PAA Ciencias | PCE CS | INTEGER | Prueba de conocimiento específico Ciencias |
| 26 | PAA Ciencias Sociales | PCE C. SOC | INTEGER | Prueba de conocimiento específico Ciencias Sociales |
| 27 | PAA Física | PCE FIS | INTEGER | Prueba de conocimiento específico Física |
| 28 | PAA Matemática | PCE MAT | INTEGER | Prueba de conocimiento específico Matemática |
| 29 | PAA Química | PCE QCA | INTEGER | Prueba de conocimiento específico Química |
| 30 | Preferencia de Postulación | PREFERENCIA DE POSTULACIÓN EN LA QUE QUEDO SELECCIONADO | INTEGER |  |
| 31 | Bonificados | INGRESO CON BONIFICACIÓN | VARCHAR() | SI, NO, si corresponde. |
| 32 | Año de ingreso | AÑO DE INGRESO | INTEGER |  |
| 33 | Código de carrera | CARRERA | INTEGER | Código que corresponde a alguna de las carreras de la USACH. |

Desde el 2003 en adelante, cambia el sistema de selección de los postulantes a las Universidades del Consejo de Rectores de Chile, reemplazando la Prueba de Aptitud Académica (PAA) por la Prueba de Selección Universitaria (PSU). Por lo mismo, para las cohortes 2004, 2005, 2006, 2007, 2008, la base de datos se modifica, reemplazando las variables 22, 23, 24, 25, 26, 27, 28 y 29, para el resto de las variables, la base se mantiene.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 34 | PSU Matemáticas | PSU MAT | INTEGER | Prueba de Selección Universitaria Matemática |
| 35 | PSU Lenguaje | PSU LENJ | INTEGER | Prueba de Selección Universitaria Lenguaje |
| 36 | PSU Historia | PSU HISTORIA | INTEGER | Prueba de Selección Universitaria Historia |
| 37 | PSU Ciencias | PSU CIENCIA | INTEGER | Prueba de Selección Universitaria Ciencias |

Rendimiento para todos los cohortes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | VARIABLE | ETIQUETA | TIPO | DETALLE |
| 1 | RUT | RUT | VARCHAR(8) | Rol Único Nacional |
| 2 | Año de Ingreso | AÑO INGRESO | INTEGER |  |
| 3 | Código Carrera Ingreso | CARRERA INGRESO | INTEGER | Código que corresponde a alguna de las carreras de la USACH. |
| 4 | Plan de Estudio | PLAN DE ESTUDIO INGRESO | VARCHAR() | Nombre formal plan de estudios de ingreso. |
| 5 | Año de Titulación | AÑO DE TITULACIÓN | INTEGER |  |
| 6 | Código Carrera Egreso | CARRERA EGRESO | INTEGER | Código que corresponde a la carrera de egreso de un estudiante en la USACH. |
| 7 | Plan de Estudios Egreso | PLAN DE ESTUDIOS EGRESO | VARCHAR() | Nombre formal plan de estudios de egreso. |
| 8 | Duración de la carrera | DURACIÓN CARRERA | INTEGER | Número de semestres de duración formal de una carrera USACH. |
| 9 | Estado Actual del estudiante | ESTADO ACTUAL | VARCHAR() | Estado actual de un estudiante, según:  Abandono por cambio de carrera  Cambio de Carrera  Eliminado por causal académica  Eliminado no titulación  Eliminado por no matrícula  Fallecimiento  Memorista  No matriculado  Postergación c/ expr. De causa  Postergación de estudios  Recons. Causal Académica  Reincorporación Decreto 2609  Reincorporación Postergación  Reincorporación R.T. C/Expr de causa  Reincorporación Renuncia  Reincorporación Post C/Expr de causa  Reincorporación por eliminación Art. Eliminado no titulación  Reincorporado  Renuncia Voluntario Nuevo Canc. Sem.  Renuncia Voluntaria nuevo ex. Arancel  Renuncia Carrera (antiguos)  Renuncia USACH ( Nuevos)  Retiro Temporal C/Expr de causa  Retiro Temporal  Sin variación  Titulado |
| 10 | Actividades extraprogramáticas | ACTIVIDADES EXTRAPROGRAMÁTICAS APROBADAS; CURSO 1 | VARCHAR() | Nombre de actividad extraprogramática cultural o deportiva aprobada por un estudiante. |
| 11 | Actividades extraprogramáticas | ACTIVIDADES EXTRAPROGRAMÁTICAS APROBADAS; CURSO 2 | VARCHAR() | Nombre de actividad extraprogramática cultural o deportiva aprobada por un estudiante. |
| 12 | Actividades extraprogramáticas | ACTIVIDADES EXTRAPROGRAMÁTICAS APROBADAS; CURSO 3 | VARCHAR() | Nombre de actividad extraprogramática cultural o deportiva aprobada por un estudiante. |
| 13 | Actividades extraprogramáticas | ACTIVIDADES EXTRAPROGRAMÁTICAS APROBADAS; CURSO 4 | VARCHAR() | Nombre de actividad extraprogramática cultural o deportiva aprobada por un estudiante. |
| 14 | Actividades extraprogramáticas | ACTIVIDADES EXTRAPROGRAMÁTICAS APROBADAS; CURSO 5 | VARCHAR() | Nombre de actividad extraprogramática cultural o deportiva aprobada por un estudiante. |
| 15 | Cantidad de asignaturas del Plan de Ingreso | CANTIDAD DE ASIGNATURA PLAN DE INGRESO | INTEGER | Número de asignaturas totales por carrera. |

Para un total de 23, la información por semestres es la siguiente:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 16 | Semestre X; Estado actual | ESTADO | VARCHAR() | Estado de un estudiante al semestre X, según:  Abandono por cambio de carrera  Cambio de Carrera  Eliminado por causal académica  Eliminado no titulación  Eliminado por no matrícula  Fallecimiento  Memorista  No matriculado  Postergación c/ expr. De causa  Postergación de estudios  Recons. Causal Académica  Reincorporación Decreto 2609  Reincorporación Postergación  Reincorporación R.T. C/Expr de causa  Reincorporación Renuncia  Reincorporación Post C/Expr de causa  Reincorporación por eliminación Art. Eliminado no titulación  Reincorporado  Renuncia Voluntario Nuevo Canc. Sem.  Renuncia Voluntaria nuevo ex. Arancel  Renuncia Carrera (antiguos)  Renuncia USACH ( Nuevos)  Retiro Temporal C/Expr de causa  Retiro Temporal  Sin variación  Titulado |
| 17 | Semestre X; Asignaturas Inscritas | CANTIDAD DE ASIGNATURAS INSCRITAS | INTEGER | Número total de asignaturas inscritas en el semestre X. |
| 18 | Semestre X; Asignaturas Aprobadas | CANTIDAD DE ASIGNATURAS APROBADAS | INTEGER | Número total de asignaturas aprobadas en el semestre X. |
| 19 | Semestre X; Asignaturas Reprobadas | CANTIDAD DE ASIGNATURAS REPROBADAS | INTEGER | Número total de asignaturas reprobadas en el semestre X. |
| 20 | Semestre X; Promedio Ponderado | PROMEDIO PONDERADO TOTAL | FLOAT | Promedio Ponderado del total de asignaturas, en el semestre X. |
| 21 | Semestre X; Promedio Ponderado Sólo Asignaturas Aprobadas | PROMEDIO PONDERADO APROBADOS | FLOAT | Promedio Ponderado de las asignaturas aprobadas, en el semestres X. |

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. Proceso durante el cual se reduce la cantidad de variables aleatorias bajo consideración. [↑](#footnote-ref-8)
9. ***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-9)
10. Fase *Business Understanding phase* (Fase de entendimiento del negocio) de *CRISP-DM*. [↑](#footnote-ref-10)
11. *http://www.tableausoftware.com* [↑](#footnote-ref-11)
12. Para más información sobre tareas supervisadas y no supervisadas refiérase al **Anexo B: Técnicas supervisadas y no supervisadas** [↑](#footnote-ref-12)
13. Medida para cuantificar los casos en los que e antecedente se hace verdadero. Puede ser número de casos o porcentaje. [↑](#footnote-ref-13)
14. Número de casos en que, habiéndose cumplido el antecedente de la regla, se cumple el consecuente. [↑](#footnote-ref-14)
15. 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-15)
16. Dividir y conquistar. [↑](#footnote-ref-16)
17. , donde son las probabilidades a priori, la probabilidad de de *B* en la hipótesis *A* y la probabilidad a posteriori. [↑](#footnote-ref-17)
18. Puede tomar valor *'verdadero' (1)*  o *'falso' (0)*. [↑](#footnote-ref-18)
19. Ya que la duración de la carrera es de 6 años, y los datos facilitados no documentan lo ocurrido durante el año 2013, que sería teóricamente el año de egreso del estudiante ejemplificado. [↑](#footnote-ref-19)
20. *Framework* para aplicaciones web en base al lenguaje de programación *ruby.* [↑](#footnote-ref-20)
21. Ruby on Rails [↑](#footnote-ref-21)
22. Se usará el rut sin código verificador para fines de identificación [↑](#footnote-ref-22)
23. Extraxt, transform and load. [↑](#footnote-ref-23)