



Universidad Católica del Maule  
Facultad de Ciencias de la Ingeniería  
Escuela de Ingeniería Civil Informática

# Práctico de Minería de Datos

---

Utilizando la metodología CRISP-DM

## **Integrantes**

*Matías Bustos, Ángel Ubilla, Chien Hao Chen, Diego Vergara, Felipe Jorquera*

## **Asignatura**

*Minería de datos*

## **Docente**

*Sergio Hernández*

## **Fecha**

*Jueves 05 de Noviembre de 2015*

# Primera etapa: Entendimiento del Negocio.

---

## Introducción:

Este trabajo tiene como objetivo promover la aplicación de técnicas avanzadas de investigación en Inteligencia Artificial a los problemas del mundo real. Se tendrá acceso a una gran cantidad de datos proporcionados por Tmall.com, la plataforma B2C más grande de China. El objetivo es aplicar técnicas avanzadas de aprendizaje automático y minería de datos para predecir qué compradores repetirán su acción después de una promoción de ventas en Tmall.com.

## Metodología:

Gran parte de este trabajo tiene relación con la minería de datos. En efecto, para lograr resultados convincentes, se utilizó una metodología de trabajo llamada CRISPDM 1.0. que ha sido validada por muchos *data scientists* del planeta. En términos estructurales, la metodología define un proceso que se compone de 4 subprocesos, en los que se establecen objetivos de trabajo, más o menos específicos, dependiendo del nivel de granularidad de las actividades, es decir que en los primeros niveles los objetivos son de carácter general y en los últimos de carácter procedural.

## Informe:

### Antecedentes:

Los comerciantes a veces hacen grandes promociones (por ejemplo, descuentos o cupones de efectivo) en fechas particulares (como “Boxing-day Sales”, “Black Friday” o “Double 11”) con el fin de atraer a un gran número de nuevos compradores. Desafortunadamente, muchos de los compradores atraídos son cazadores de solo una oferta, y estas promociones pueden tener poco impacto duradero en las ventas. Para aliviar este problema, es importante para los comerciantes lograr identificar quienes pueden convertirse en compradores reiterados. Al enfocarse en estos potenciales y leales clientes, los comerciantes pueden reducir en gran medida el costo de promoción y mejorar el costo de la inversión. Es bien sabido en el campo de la publicidad online, que encontrar el cliente “objetivo” es extremadamente difícil, especialmente para los compradores nuevos. Sin embargo, con el registro de comportamiento de los usuarios acumulado a largo plazo por Tmall.com, puede que ayude a resolver este problema.

En base a esta premisa, el grupo buscará dar solución a la problemática por medio de la aplicación de algoritmos de minería de datos. En primera instancia, el grupo plantea un modelo de clasificación tomando como entradas los campos que encontramos en el dataset. Como uno de los objetivos es conocer los compradores potenciales para el envío de ofertas, la

predicción de dichos compradores se puede obtener a través de éste modelo de minería de datos, ofreciendo una solución eficaz al cliente.

### **Objetivos de negocio y criterios de éxito:**

En este desafío, se dispone de un conjunto de datos de comerciantes y clientes, así como también de sus acciones registradas durante la promoción en el “Double 11 day”. La tarea consiste en predecir cuál de los nuevos compradores dado los comerciantes se convertirán en clientes habituales en el futuro. En otras palabras, es necesario predecir la probabilidad de que estos nuevos compradores compren artículos de los mismos comerciantes de nuevo dentro de 6 meses.

Como equipo suponemos que, dada la naturaleza de la organización, uno de los criterios de éxito del negocio es la participación de los clientes habituales dentro de las actividades de las tiendas o los vendedores ya que permite dar una primera impresión acerca del nivel de satisfacción que poseen en relación a los servicios a los que se ofrecen o disponen. Por ello, atender y prestar especial atención a las dinámicas del modelo de negocio a través de herramientas como la minería de datos es de suma importancia

### **Inventario de recursos:**

El personal para llevar a cabo el proyecto estará compuesto por 5 Ingenieros Civiles informáticos, entre ellos Matías Bustos, Ángel Ubilla, Chien Hao Chen, Diego Vergara, Felipe Jorquera, quienes en conjunto aplicarán las tecnologías necesarias para el análisis de Big Data y minería de datos. El conjunto de tecnologías que se utilizarán está compuesto de computadores personales con distintos sistemas operativos, entre ellos Linux, Windows y OSX, así también un súper computador armado con un coprocesador Intel Xeon Phi y Linux ubicado en las dependencias de la Universidad Católica del Maule, se utilizará la distribución HortonWorks montada sobre CentOS la cual contiene una serie de aplicaciones útiles para el tratamiento de Big Data y procesamiento distribuido, dentro de las más importantes Hadoop y su sistema de archivos distribuidos HDFS, Apache HIVE para procesamiento de consulta SQL, Apache Pig, Apache Spark, todas estas bajo la metodología MapReduce, entre otras. Cabe destacar que la distribución HortonWorks será utilizada en el supercomputador y un computador personal dotado con OSX.

La fuente de datos contiene registros anónimos de compras de usuarios en los últimos 6 meses antes y en el “Double 11 day”, además de la información que indica si son compradores repetidos. Estos datos se encuentran en distintos archivos en formato “CSV”.

### **Requerimientos, suposiciones y restricciones:**

Los resultados deberán estar en un formato de entrega determinado que está definido de la siguiente manera:

Atributo	Definición
user_id	Un id único para el comprador.
merchant_id	Un id único para el comerciante.
prob	Probabilidad predecida del usuario de convertirse en un comprador reiterado del comerciante dado. El valor debe estar entre 0 y 1.

### Riesgos y contingencias:

A continuación se presenta una lista de los riesgos que pudieran retrasar, o en el peor de los casos, detener el proyecto. También se anexan las acciones de contingencia pertinentes a modo de evitar un desastre mayor.

N	Descripción de riesgo	Plan de contingencia
1	Carencia de los recursos computacionales suficientes para ejecutar un ambiente diseñado para sistemas distribuidos.	Confeccionar un plan de mejora para el uso de los recursos computacionales.
2	Datos limpiados y transformados erróneamente debido a una mala comprensión de las reglas del negocio.	Escribir un registro de los inconvenientes y explorar el uso de los procedimientos del negocio.
3	Mal uso de la herramientas y déficit de aprendizaje debido a mala comprensión de la documentación.	Reproducir una serie de tutorías que permitan a los miembros del equipo de trabajo comprender en conjunto las herramientas del entorno y los datos de estudio.
4	Uso de un modelo inadecuado para estudiar y analizar los datos de entrada. Esto puede suponer conclusiones que se validan gracias a premisas erróneas.	Validar resultados con los datos estudiados y probar consistencias.

### Terminología:

- **Data Mining:** Es un proceso de análisis de grandes cantidades de datos para descubrir conocimiento e información oculta que proviene de diferentes perspectivas y así lograr un resumen de información útil para mejorar los ingresos o reducir costos del negocio... etc.
- **Big Data:** término asociado a los conjuntos de datos que son lo suficientemente grandes o complejos para no poder ser manipulados por aplicaciones convencionales.
- **Hadoop:** Es un framework que permite hacer el proceso distribuido en gran cantidad de datos a través de varios ordenadores usando los modelos simples de programación.
- **HortonWorks:** Es una compañía de software empresarial con sede en Santa Clara, California que se centra en el desarrollo y apoyo de Hadoop, un marco que permite el procesamiento distribuido de grandes conjuntos de datos a través de clusters de computadores.

- **YARN:** es el centro arquitectónico de Hadoop que permite a los motores de procesamiento de múltiples datos como SQL interactivo, streaming en tiempo real, la ciencia de datos y el procesamiento por lotes para manejar los datos almacenados en una única plataforma, abriendo todo un nuevo enfoque para el análisis de datos. Esta arquitectura funciona sobre MapReduce.
- **MapReduce:** Es un modelo de programación y una implementación asociada al procesamiento y la generación de grandes conjuntos de datos de forma paralela, algoritmo distribuido en clusters.
- **HDFS:** Hadoop Distributed File System, un sistema de archivo distribuido que proporciona alto rendimiento para acceder datos en una aplicación.
- **Apache Spark:** Spark proporciona el modelo de programación simple que soporta un rango amplio de aplicaciones, incluido ETL, aprendizaje de máquina, computación gráfica y procesamiento de flujo.
- **Apache Pig:** Un lenguaje de alto nivel para el flujo de datos y la ejecución de framework en computación paralela.
- **Apache Hive:** Una infraestructura de data warehouse que resume los datos y hace de la consulta en manera más apropiada y sencilla.
- **Ambari:** Una herramienta basada en web(Dashboard) para la provisión, administración, monitoreo Apache Hadoop cluster.

### **Costos y beneficios:**

En este caso puntual la minería de datos tiene un costo asociado directamente con la compra de equipo computacional suficientemente potente para realizar el análisis de los datos, debido a que se necesita una gran cantidad de memoria ram para procesar la información. Otro costo es el que se crea al adquirir un software o framework pagado. Si bien es cierto que hay alternativas libres, el hecho de adquirir una aplicación de pago, permite tener soporte en caso de tener algún problema con el entorno de trabajo.

A los comerciantes les interesa saber quienes serán compradores que vuelvan a consumir un producto en el futuro basándose en los datos que ya se tiene. Esto se puede conseguir a través de la minería de datos, por lo cual, es un beneficio directo para los comerciantes, debido a que el conocer esta información les puede hacer aumentar sus ventas y con esto sus ingresos.

### **Metas de minería de datos y criterios de éxito:**

La minería de datos busca conseguir modelos a partir de los datos para obtener predicciones sobre los comportamientos de clientes (en este caso). A los comerciantes les interesa saber qué clientes se convertirán en compradores repetidos, lo cual sería la meta de la minería. Un criterio adecuado para este caso, según el grupo de trabajo es tener una mayor probabilidad de que un cliente se convierta en comprador repetido para un determinado comerciante.

### Plan de proyecto:

Etapas	Duración	Recursos utilizados	Entradas	Salidas	Dependencias
<i>Entendimiento del negocio</i>	1 semana	Personal.	Información de negocio	Lista de objetivos	
<i>Entendimiento de los datos</i>	1 semana	Personal, computadores, oficinas.	Datos de clientes.	Histograma y tablas.	Entendimiento del negocio
<i>Preparación de los datos</i>	1 semana	Personal, computadores, oficinas.	Datos de clientes.	Datos Preparados	Entendimiento de los datos
<i>Modelado</i>	1 semana	Personal, computadores, oficinas.	Datos Preparados	Modelo de minería	Preparación de los datos
<i>Evaluación</i>	4 días	Personal, computadores, oficinas.	Modelo de minería	Decisión y Resultado de evaluación	Modelado
<i>Despliegue</i>	3 días	Personal, computadores, oficinas.	Decisión y Resultado de evaluación	Plan de mantención y monitoreo, Reporte final	Evaluación

### Evaluación inicial de herramientas y técnicas:

Los datos en el archivo csv que contiene el join de los datos de estudio, fueron cargados de manera exitosa en una tabla de datos creada en la herramienta Hive, cabe destacar que este procedimiento se realizó en un computador personal, se realizaron consultas simples sql a la tabla y los resultados se presentaron en un tiempo aproximado de 2.5 minutos, tiempo razonable para los 52 millones de datos que contenía la estructura, se probaron distintos tipos de consultas, entre ellas, consultas anidadas, agrupamientos, histogramas numéricos y los resultados demoraron aproximadamente el mismo tiempo. En algunas ocasiones cuando el equipo estaba muy sobrecargado de memoria ram, las consultas se colgaban, pero nuevamente realizadas eran completadas exitosamente.

La idea de ocupar Apache Hive, es obtener un estudio inicial de los datos, para el entendimientos de los datos que nos son proporcionados. Para el caso de la limpieza, y transformación en el caso de que lo sea necesario, se pretende utilizar igualmente Hive que no provee herramientas fáciles de utilizar. Para el análisis profundo de los datos y posteriormente la minería de datos, se intentará utilizar Apache Spark en conjunto de algún módulo de Python o R, en caso de que no se posible implementarlo en un corto tiempo, se tomará la determinación de utilizar el método del “muestreo”, en el cual se toman muestras de igual

tamaño y aleatoria de los datos y se analizan con técnicas de minería de datos a través de Python o R. Este análisis debe efectuarse en reiteradas iteraciones para obtener resultados que sean significativos con respecto a la proporción que fueron tomados y el universo de los datos.

## Segunda etapa: Entendimiento de los Datos.

### Introducción:

Con el proyecto se obtiene acceso a datos que se enlistan como recursos del proyecto. Esta “colección inicial de datos” incluye carga de datos, en caso de ser necesario para el entendimiento.

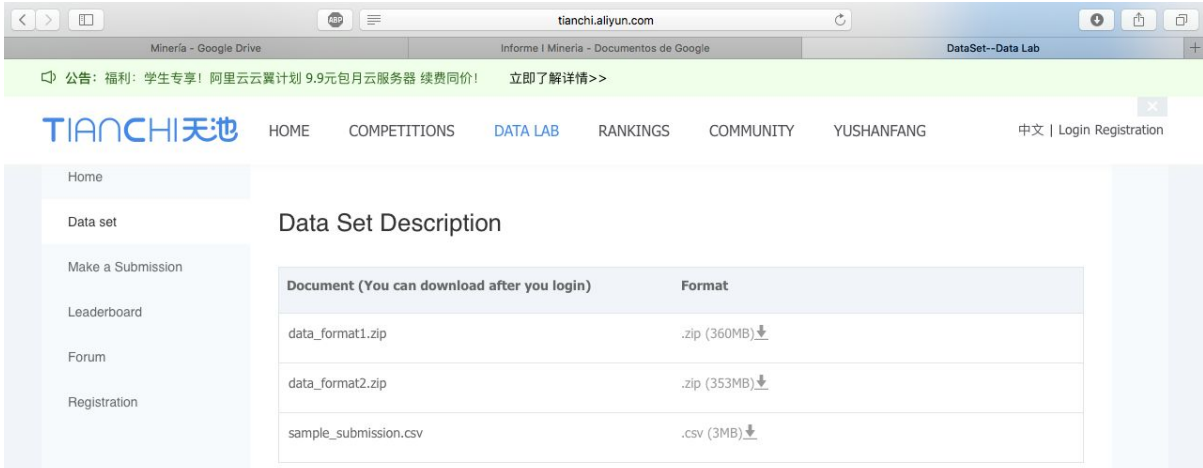
Se listan el o los datasets junto con su ubicación dentro del proyecto, los métodos usados para adquirirlo y los problemas ocurridos. Se guardan los problemas encontrados y se archivan las soluciones para ayudar en caso de futuros proyectos similares.

### Informe:

#### Informe inicial de recopilación de datos.

Los datos fueron descargados de un servidor web cuya dirección es la siguiente:

<http://tianchi.aliyun.com/datalab/dataSet.htm?spm=5176.100074.5678.2.wHMuHv&id=1>




The screenshot shows the Tianchi Data Lab website interface. At the top, there's a navigation bar with links: HOME, COMPETITIONS, DATA LAB, RANKINGS, COMMUNITY, YUSHANFANG, and 中文 | Login Registration. Below the navigation bar, there's a sidebar on the left with links: Home, Data set, Make a Submission, Leaderboard, Forum, and Registration. The main content area is titled "Data Set Description" and contains a table with the following data:


Document (You can download after you login)	Format
data_format1.zip	.zip (360MB) <a href="#">Download</a>
data_format2.zip	.zip (353MB) <a href="#">Download</a>
sample_submission.csv	.csv (3MB) <a href="#">Download</a>

El servidor contiene dos archivos en formato “zip” en los cuales están contenidos los dos formatos de los datos de análisis y almacenados en archivos de formato “csv”. La estructura, los volúmenes y resúmenes de datos se detallan a continuación.


## Formato 1:




**test\_format1.csv**




**train\_format1.csv**




**user\_info\_format1.csv**



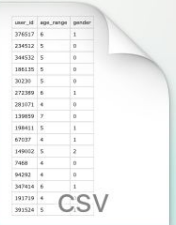
**user\_log\_format1.csv**



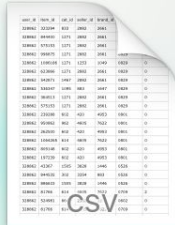
**test\_format1.csv**  
3,3 MB  
Última modificación 27-10-2015 4:36:31 p.m.



**train\_format1.csv**  
3,5 MB  
Última modificación 31-03-2015 5:32:00 a.m.




**user\_info\_format1.csv**  
4,5 MB  
Última modificación 31-03-2015 5:32:33 a.m.




**user\_log\_format1.csv**  
1,91 GB  
Última modificación 27-10-2015 4:04:12 p.m.


## Formato 2:




**test\_format2.csv**



**train\_format2.csv**



**test\_format2.csv**  
768,7 MB  
Última modificación 31-03-2015 6:00:12 a.m.



**train\_format2.csv**  
767,4 MB  
Última modificación 31-03-2015 6:01:44 a.m.

La dificultad de contar con tamaños de datos tan grandes es poder hacer realizar una buena visualización de los datos, algunas aplicaciones como SublimeText, permiten visualizar los datos cargandolos en memoria, pero el proceso es un poco extenso y los resultados son muy poco útiles, debido a la gran cantidad que hay que revisar, es por ello la necesidad de utilizar técnicas de resumen de datos.



## Informe de descripción de los datos.

Al aplicar una visualización de la cabecera de los archivos de datos, se logra observar los campos registrados y el formato para cada campo, los dos formatos contienen estructuras distintas pero con los mismos datos. A continuación se detalla dicha información en conjunto con el significado de cada uno de los campos antes mencionados.

Formato 1:

```
MacBook-Pro-de-Diego:data_format1 Diego$ head train_format1.csv
user_id,merchant_id,label
34176,3906,0
34176,121,0
34176,4356,1
34176,2217,0
```

```
MacBook-Pro-de-Diego:data_format1 Diego$ head user_info_format1.csv
user_id,age_range,gender
376517,6,1
234512,5,0
344532,5,0
186135,5,0
```

```
MacBook-Pro-de-Diego:data_format1 Diego$ head test_format1.csv
user_id,merchant_id,prob
163968,4605,
360576,1581,
98688,1964,
98688,3645,
```

```
MacBook-Pro-de-Diego:data_format1 Diego$ head user_log_format1.csv
user_id,item_id,cat_id,seller_id,brand_id,time_stamp,action_type
328862,323294,833,2882,2661,0829,0
328862,844400,1271,2882,2661,0829,0
328862,575153,1271,2882,2661,0829,0
328862,996875,1271,2882,2661,0829,0
```

- Formato 1.
  - Registro del Comportamiento del Usuario.

Atributo	Definición
user_id	Un id único para el comprador.
item_id	Un id único para cada ítem.
cat_id	Un id único para la categoría donde el ítem pertenece.
merchant_id	Un id único para el comerciante.
brand_id	Un id único para la marca del ítem.
time_stamp	Fecha en que la acción se realiza (formato: mmdd).
action_type	Tipo enumerado {0,1,2,3}, donde 0 es para click, 1 es para añadir al carro, 2 es por comprar y 3 añadir a favorito.

- Perfil de Usuario.

Atributo	Definición
user_id	Un id único para el comprador.
age_range	Rango de edad de los usuarios: 1 para <18; 2 para [18,24]; 3 para [25,29]; 4 para [30,34]; 5 para [35,39]; 6 para [40,49]; 7 y 8 para $\geq 50$ ; 0 y NULL para desconocido.
gender	Sexo del usuario: 0 para femenino, 1 para masculino, 2 y NULL para desconocido.

- Datos de Entrenamiento y de Prueba.

Atributo	Definición
user_id	Un id único para el comprador.
merchant_id	Un id único para el comerciante.
label	Es un tipo enumerado {0,1}, donde 1 es un comprador repetido, y 0 para el caso contrario. Este campo está vacío para los datos de prueba.

Formato 2:

```
MacBook-Pro-de-Diego:data_format2 Diego$ head test_format2.csv
user_id,age_range,gender,merchant_id,label,activity_log
163968,0,0,4378,-1,101206:812:6968:0614:0
163968,0,0,2300,-1,588758:844:3833:0618:0#71782:844:3833:1111:2
#71782:844:3833:1111:0#71782:844:3833:1102:0#702201:844:3833:11
02:0#1009809:844:3833:1102:0#71782:844:3833:1110:0#71782:844:38
833:0618:2

MacBook-Pro-de-Diego:data_format2 Diego$ head train_format2.csv
user_id,age_range,gender,merchant_id,label,activity_log
34176,6,0,944,-1,408895:1505:7370:1107:0
34176,6,0,412,-1,17235:1604:4396:0818:0#954723:1604:4396:0818:0#27
:0#548906:1577:4396:1031:0#368206:662:4396:0818:0#480007:1604:4396
:4396:0818:0#236488:1505:4396:1024:0
```

■ Formato 2.

Atributo	Definición
user_id	Un id único para el comprador.
age_range	Rango de edad de los usuarios: 1 para <18; 2 para [18,24]; 3 para [25,29]; 4 para [30,34]; 5 para [35,39]; 6 para [40,49]; 7 y 8 para ≥ 50; 0 y NULL para desconocido.
gender	Sexo del usuario: 0 para femenino, 1 para masculino, 2 y NULL para desconocido.
merchant_id	Un id único para el comerciante.
label	Valores entre {0,1,-1,NULL}, 1 denota que el user_id es un comprador reiterado para el merchant_id, mientras que 0 es lo contrario. -1 representa que el user_id no es un nuevo cliente del comerciante, por tanto, está fuera de la predicción. Sin embargo, estos registros pueden proporcionar información adicional. NULL solo se da en los datos de prueba, indicando que es un par a predecir.
activity_log	Conjunto de registros de interacción entre {user_id, merchant_id}, donde cada registro es una acción representada como {item_id:category_id:brand_id:time_stamp:action_type}. # Se utiliza para separar dos elementos vecinos. Los registros no están ordenados en un orden particular.

Como los dos formatos contienen los mismos datos estructurados de forma distinta, se procede a realizar una unión de los datos, a través de un join, separando cada una de las columnas, con el objetivo de mantener todos los datos disponibles y necesarios, bien estructurados.

Join:



```
MacBook-Pro-de-Diego:Datasets Tarea Final Diego$ head Join.csv
user_info_stage.user_id,user_info_stage.age_range,user_info_stage.gender,user_log_stage.user_id,user_log_stage.item_id,user_log_stage.cat_id,user_log_stage.seller_id,user_log_stage.brand_id,user_log_stage.time_stamp,user_log_stage.action_type
32,3,1,32,243113,184,3449,7394,1106,3
32,3,1,32,115484,1142,1487,6445,1106,3
32,3,1,32,365915,184,496,7168,0911,3
32,3,1,32,664468,602,1058,3369,0627,0
32,3,1,32,194491,662,1567,8364,0817,0
```

## **Informe de los datos.**

En el objetivo de la exploración de los datos es determinar la calidad de los datos, así como patrones que se puedan encontrar entre ellos. Primero se determinará las cantidades de datos con las cuales se cuenta para cada columna de la fuente de datos, también analizaremos los datos nulos que se encuentran y datos que no tengan relación con la especificación y límites especificados anteriormente, así como cualquier anomalía entre, también analizaremos los datos con respecto a varias columnas para observar si la información obtenida es importante para el descubrimiento de patrones o comportamientos que nos ayuden a formar una hipótesis inicial para concluir el trabajo de minería de datos.

Como parte de un estudio previo (estudio blando) de los datos se analizaron los datos correspondientes al perfil del usuario y su comportamiento. En primera instancia se estudia el comportamiento de las acciones de los compradores de acuerdo al género de cada uno de ellos, con esto determinar cuál es el género predominante en las compras, posteriormente, se estudia la relación de las acciones de compra con los rangos de edades, de manera de conocer las edades de las personas a las cuales los productos les llama la atención.

En el caso de que al realizar un estudio en profundidad de los datos nulos que se pueden encontrar en la base de datos fuente, es importante mencionar que se realizará una limpieza de estos al igual que los datos anómalos que no correspondan a las especificaciones dadas, de manera que no afecten a las conclusiones del estudio y análisis final. De igual forma, en el caso de que existan datos con formatos incorrectos, se procederá a la transformación de estos para poder utilizarlos de forma correcta y rápida. El objetivo fundamental de estas acciones es contar con datos de calidad que nos permitan realizar un trabajo y estudio de calidad.

En el conjunto de datos existe una columna en el cual se indica que un comprador tiene características de repetitivo hacia un vendedor, esta información es una de las más importantes debido a que concluimos que será parte de la variable dependiente del estudio.

## **Informe de calidad de datos<sup>1</sup>.**

Para determinar la calidad es necesario contar con elementos que permitan dilucidar -en base a supuestos- cuando un dato cumple o no con los requisitos del negocio (lo que no necesariamente están alineados con los técnicos). Por ello se utilizará una metodología que se enfoca en establecer un marco la calidad de los datos en proyectos con grandes volúmenes de datos.

Estructuralmente hablando, la calidad de los datos se concibe desde distintos niveles o facetas de especificidad, esto significa que existen consideraciones de alto nivel (las dimensiones) que consideran aspectos generales como: la frescura, la completitud y la exactitud de los datos y otras de bajo nivel (los factores) que corresponden a los aspectos

---

<sup>1</sup> Para más información acerca de la metodología:

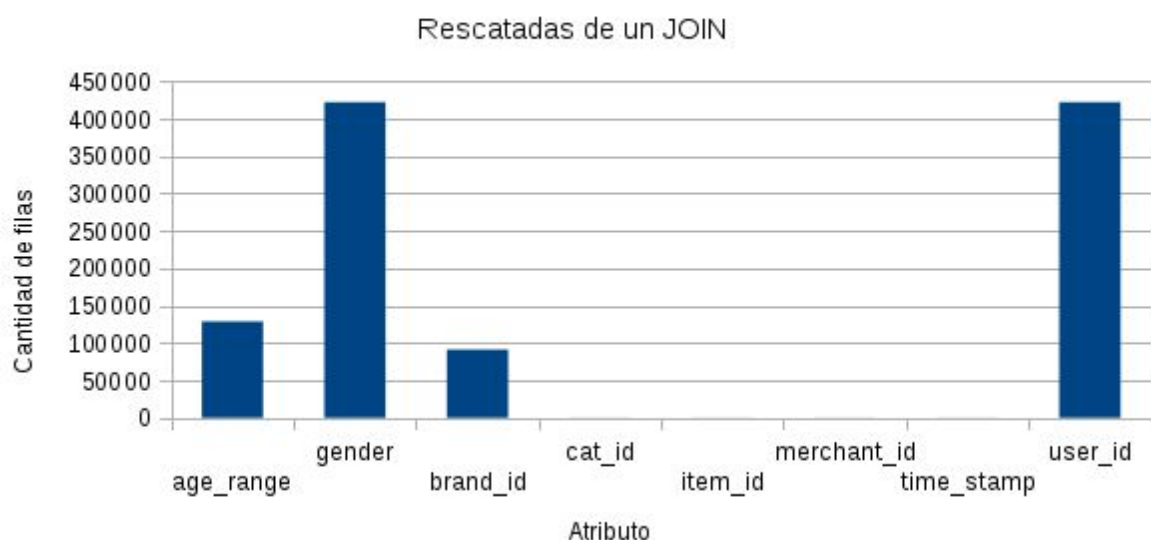
<https://www.fing.edu.uy/inco/cursos/caldatos/Transparencias/2-Dimensiones%20de%20calidad.pdf>

particulares de las dimensiones, por ejemplo la correctitud semántica y sintáctica o la precisión (exactitud).

A modo introductorio (o exploratorio), se analizarán los atributos que contienen un valor null en los registros de estudio.

Atributo	Cantidad de null (filas con presencia de celdas null)	Porcentaje del total (54505773 filas)
action_type	7	0.000012843
age_range	128606	0.23549319
gender	421938	0.774116166
brand_id	91028	0.167006163
cat_id	5	0.000009173
item_id	5	0.000009173
merchant_id	5	0.000009173
time_stamp	7	0.000012843
user_id	421938	0.774116166

Cantidad de filas con celdas null por atributo



Como se puede observar en la figura superior, la mayor cantidad de filas con celdas que poseen contenido null corresponden a aquellas que han sido analizadas por el atributo gender y user\_id. Esto advierte la presencia de una relación entre los atributos gender y user\_id que pudo darse por la falla de alguno de los sistemas de verificación de la organización

o un error en el desarrollo de la transformación de los datos. Por otro lado, la presencia de nulls en atributos como cat\_id, item\_id, merchant\_id o time\_stamp es insignificante, de hecho, como se puede observar en la tabla anterior, corresponde a un millonésima parte del total de las filas.

En relación a lo anterior podemos concluir que en términos de la completitud de los datos (presencia de nulls), el cruzamiento de filas (resultado de un JOIN) contiene un mínimo -pero no por ello menos importante- porcentaje de insatisfacción.

## Datos fuera de especificación.

Los datos que se presentan a continuación carecen de significado para el atributo con que se relacionan. Nuevamente se atribuye a este tipo de eventos los errores de transformación o -los comúnmente llamados- errores de tipeo.

Atributo	Dato	Especificación
gender	765657	Sexo del usuario: 0 para femenino, 1 para masculino, 2 y NULL para desconocido.
age_range	424134	Rango de edad de los usuarios: 1 para <18; 2 para [18,24]; 3 para [25,29]; 4 para [30,34]; 5 para [35,39]; 6 para [40,49]; 7 y 8 para ≥ 50; 0 y NULL para desconocido.

time_stamp	Fecha equivalente	Count
719	19-07	147564
727	27-07	126743
820	20-08	173092
829	29-08	191664
916	16-09	204480
928	28-09	194320
1024	24-10	321046
1031	31-10	540177
.	.	.
.	.	.
.	.	.

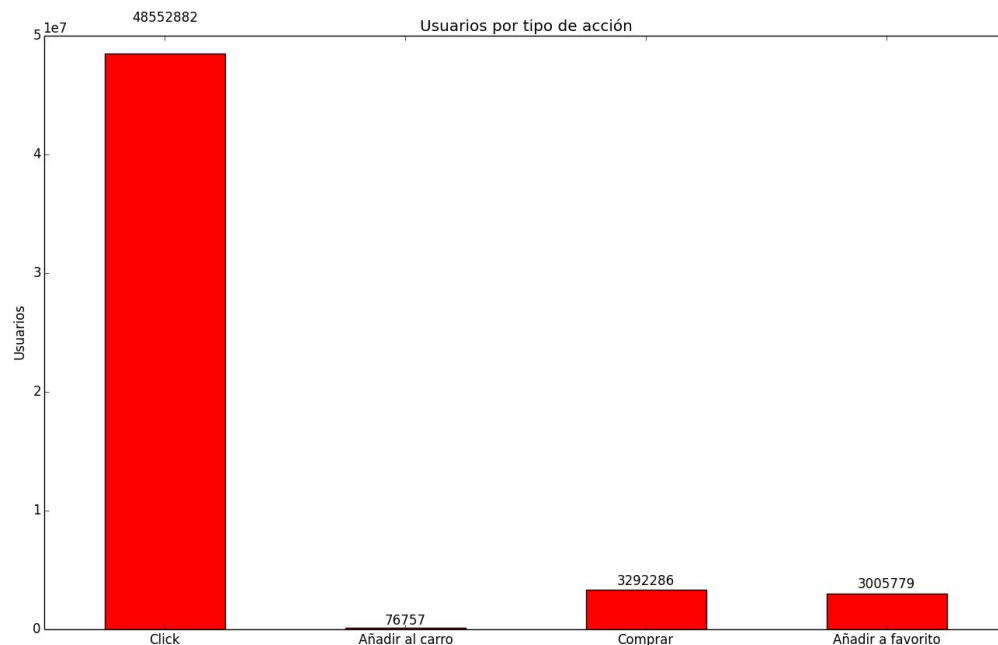


## Cantidad de datos distintos.

Otro punto importante a analizar es la diversidad de los datos, ya que nos permite evaluar desde un punto de vista muy general el contexto de las transacciones (por ejemplo, la relación entre cantidad de usuarios y total de compras, o ítems versus categorías, entre otros)

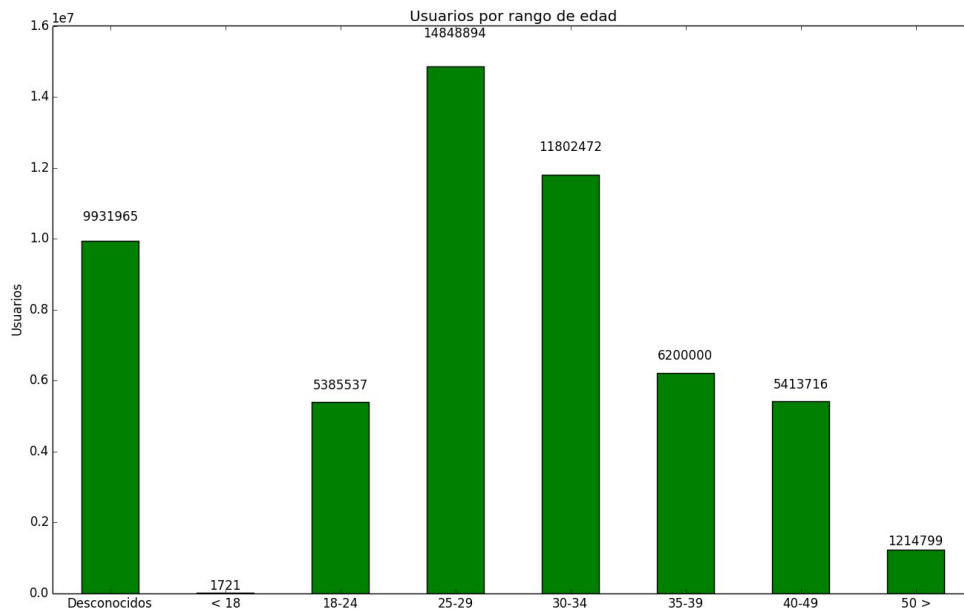
Usuarios	Items	Categorías	Comerciantes	Marcas
424171	1090390	1659	4996	8443

Gráfico 1



En este gráfico se muestra el comportamiento de los clientes en base a sus acciones realizadas al interactuar con el sistema. Es un gráfico de barras simple, que grafica la cantidad de clicks que se realizan, cuantas veces se añade al carro (posible compra), las compras efectuadas y la cantidad de veces que se añadió a favorito. Se aprecia que la cantidad de personas que añade a favorito y la cantidad que compra es bastante cercana al analizar mirando las barras, pero en cuanto a números, hay una diferencia considerable.

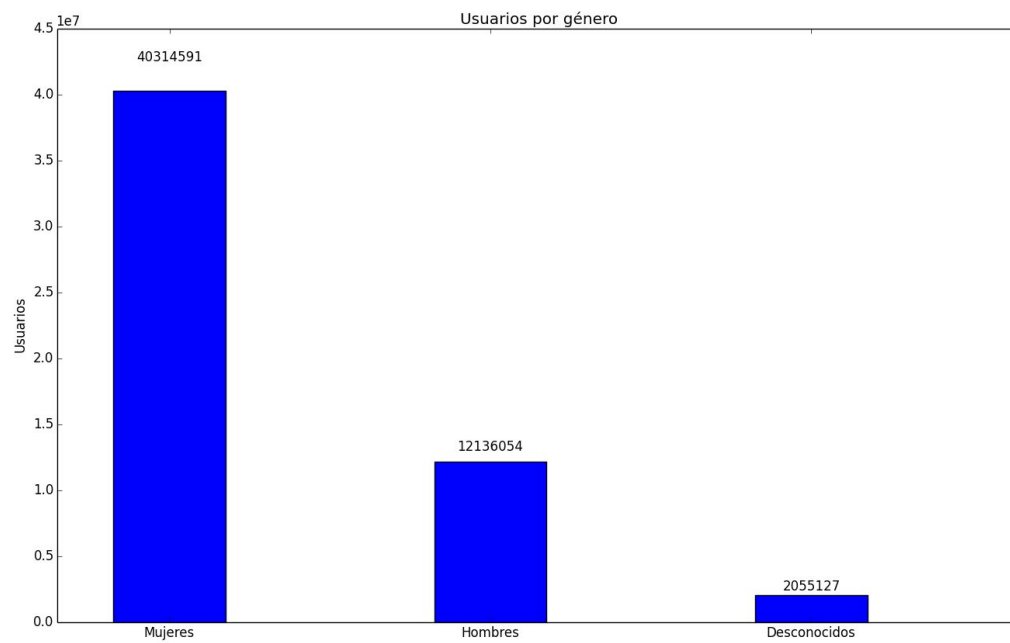
Gráfico 2



Respecto a la comprensión de datos relacionados con la edad, se aprecia que la mayor cantidad de usuarios está en el rango de 25 a 29 años, seguidos por el rango de 30 a 34 y desde allí comienza a descender a medida que aumenta la edad. Se puede ver claramente que hay una cantidad considerable de personas a las cuales no se les conoce la edad, lo cual es muestra clara de que los datos no son completos y por ende poseen falta de calidad.

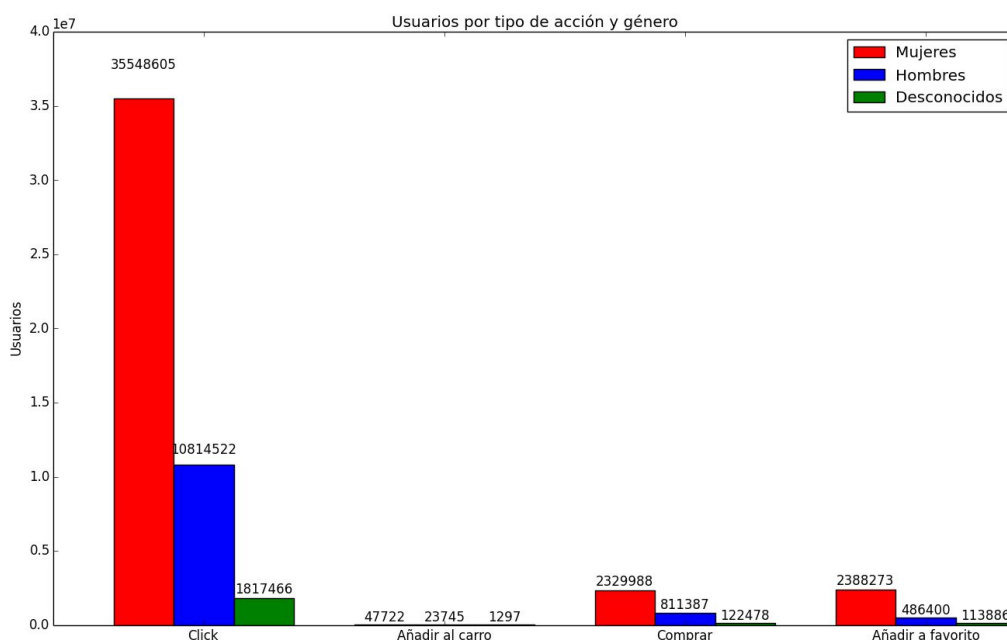


**Gráfico 3**



En esta figura se grafica la cantidad de mujeres y hombres en la base de datos, se aprecia que la cantidad de mujeres es más del triple de hombres. Por otra parte, se vuelve a apreciar que hay datos desconocidos.

Gráfico 4



En el análisis de de los datos se encontraron anomalías en cuanto a que existían datos nulos para cada columna de la fuente de datos, así como que para el caso del género, se encontró un dato que no pertenecía al rango especificado. Debido a lo anterior es que se debe realizar una limpieza de los datos en la cual se eliminen estos datos nulos y el dato anómalo encontrado.

Existe una columna en los datos fuentes, que representa la fecha de compra y se encuentra en formato mes-día, la irregularidad que se observó, fue que no poseía un formato fecha, si no que una concatenación numérica del mes y el día. Esta condición nos llevó a determinar que era necesario realizar una transformación de los datos correspondientes a esa columna de datos, de manera de especificarlos de acuerdo a una fecha correcta y de esta manera poder utilizar de forma adecuada ese campo que nos puede ser de mucha utilidad en un análisis de datos temporal.

Para concluir con el análisis de la calidad de los datos, se consideró importante mencionar el cumplimiento de la exactitud de los datos, en donde factores como el análisis de la semántica (cantidad de valores null) y la sintaxis (valores anómalos) de los registros almacenados fueron cruciales. Es más, una mínima parte del total de datos formó parte de estas deficiencias, sumando un total de menos de una milésima parte del conjunto completo.