Materia: Introducción al Data Mining

Tema: Caso 1

Octubre de 2022

Contenido

Autores

- Gabriel Donadia
- Gastón Larregui
- Agustin Lopez Fredes
- Aureliano Chavarria

Inicialmente se describen las etapas del proceso KDD y se hace énfasis en la etapa de minería de datos, detallando algunas técnicas utilizadas, como son la clasificación, la asociación, clustering en modo conceptual con el fin de dar un enfoque de los mismo al problema presentado en el CASO 1.

1. Proceso de KDD (Knowledge Discovery in Databases)

El proceso KDD es básicamente un proceso automático en el que se combinan el descubrimiento y análisis de los datos, con el fin de dar valor a la información que contiene una organización. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. Estas tareas implican generalmente pre-procesar y almacenar los datos, luego realizar minería de datos (data mining) examinando los diferentes patrones para realizar la presentación de los resultados

Las etapas involucradas en el proceso KDD que se ilustra en la figura 1 se resumen a continuación:

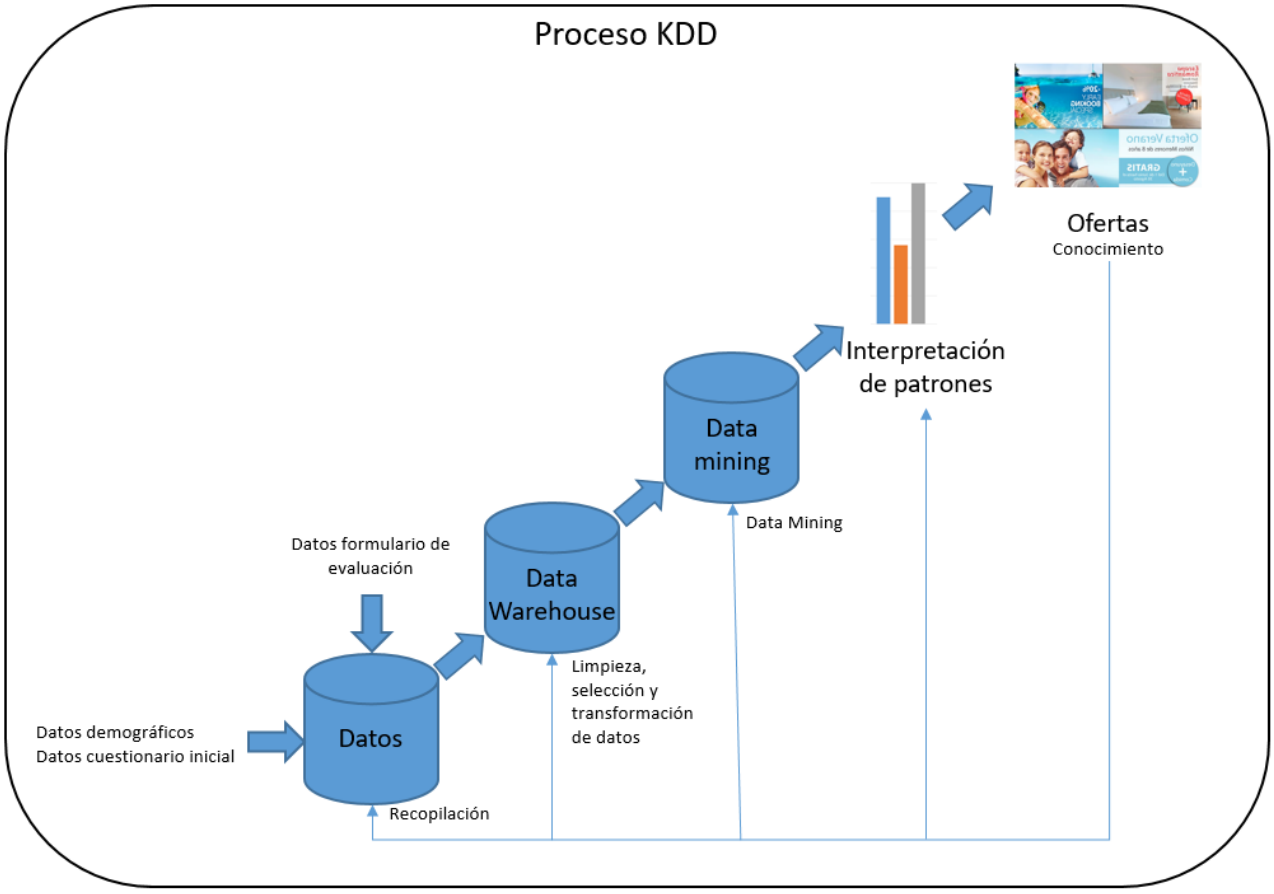

Figura 1

1. **Recopilación de datos:** Una vez identificado el conocimiento relevante, prioritario y definidas las metas del proceso KDD, desde el punto de vista del usuario final, se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos o una muestra representativa de este, sobre el cual se realiza el proceso de descubrimiento.

2. **Pre-procesamiento y Limpieza de datos:** Esta etapa es la de mayor duración de tiempo durante la ejecución del proyecto, dado que se analiza la calidad de los datos, se aplican operaciones para la remoción de datos ruidosos, se seleccionan estrategias para el manejo de datos desconocidos, datos nulos, datos duplicados y técnicas estadísticas para su reemplazo. En esta etapa, es de suma importancia la interacción con el usuario.

   Los datos ruidosos son valores que están significativamente fuera del rango de valores esperados; se deben principalmente a errores humanos, a cambios en el sistema, a información no disponible a tiempo entre otros.

3. **Selección y Transformación de los datos:** En esta instancia se buscan características útiles para representar los datos dependiendo de la meta del proceso de negocio. Se utilizan métodos de reducción de dimensiones, agregaciones, compresión de datos, segmentación, discretización basada en entropía, muestreo y/ o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos.

4. **Datamining y Análisis de Patrones:** El objetivo de la etapa es la búsqueda y descubrimiento de patrones insospechados y de interés, aplicando tareas de descubrimiento como clasificación, clustering, asociaciones, regresiones entre otras.

   Las técnicas de minería de datos crean modelos que son predictivos o descriptivos.
   Los modelos predictivos, utilizan las variables para predecir valores desconocidos o futuros de otras variables.
   Los modelos descriptivos identifican patrones que explican los datos.

   En la figura 2 se ilustran los modelos para el análisis de patrones.

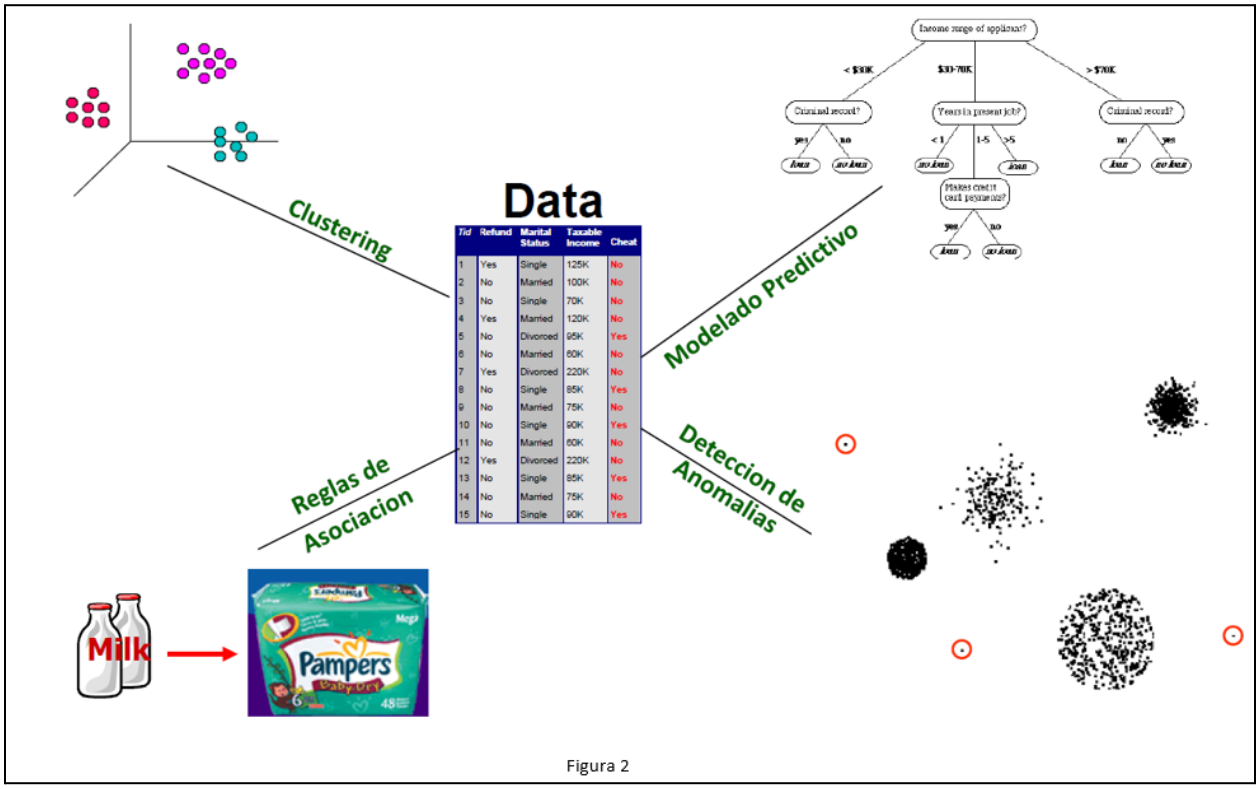

Figura 2

5. **Presentación:** Se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones. Esta etapa puede incluir la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de los patrones útiles en términos que sean entendibles para el usuario. Por otra parte, se consolida el conocimiento descubierto para incorporarlo en otro sistema para posteriores acciones o, simplemente, para documentarlo y reportarlo a las partes interesadas; también para verificar y resolver conflictos potenciales con el conocimiento previamente descubierto.

2. Propuesta Modelo Data Mining Hotel

De acuerdo a las diferentes etapas de KDD detalladas previamente, se realiza en forma conceptual el desarrollo correspondiente al Caso 1.

Como objetivos específicos se define lo siguiente:

1. Decidir cuáles anuncios de TV van a ser enviados a los huéspedes durante su estadía; un subproblema de este primer problema es decidir cuáles anuncios van a ser enviados durante los primeros días de la estadía de un huésped.
2. Decidir sobre la selección de hoteles para el mailing privado durante el verano.

Etapa 1 – Recopilación de datos iniciales

En primer lugar, se definen los datos correspondientes a la información que será solicitada a cada huésped al ingreso del hotel, dicha información cuenta con datos demográficos y otros de interés con el fin de crear un conjunto de datos objetivo por usuario (huésped).

Se presentan algunos de los datos solicitados al huésped.

- Edad
- Sexo
- Lugar de residencia
- Nacionalidad
- Estado civil
- Ocupación
- Cantidad de hijos
- Intereses recreativos
- Es la primera vez en el hotel

Etapa 2 - Pre-procesamiento y Limpieza de datos

Los datos registrados en la etapa 1 ingresan al DataWarehouse, donde se realizará la limpieza de los mismos dado que existe la posibilidad de que se presenten datos ruidosos (campos vacíos, sin sentido, errores humanos, etc), estos datos serán codificados de tal forma que no alteren los análisis posteriores.

Etapa 3 – Selección y Transformación de los datos

Se definirán grupos de datos de similares categorías, los cuales se denominan “perfiles”.
Los mencionados perfiles contendrán diferentes promociones de eventos para los que luego se realizará el envío de los anuncios.
Se listan algunos perfiles a modo de ejemplo:

- Perfil 1 – Eventos Musicales
- Perfil 2 – Eventos Gastronómicos
- Perfil 3 – Eventos Infantiles
- Perfil 4 – Actividad físicas

- Perfil 5 – Actividad recreativas familiares
- Perfil 6 – Promociones Verano / Invierno / Fines de Semana
- Perfil 7 – Ofertas productos femeninos en tienda
- Perfil 8 – Ofertas productos masculino en tienda
- Perfil 9 – Varios.

Etapa 4 – Data Mining y Análisis de Patrones

Una vez recolectados los datos de interés y del tipo de conocimiento que se desea extraer se selecciona los modelos de machine learning para predecir los anuncios a ser enviados para cada huésped:

1. Durante los primeros días de estadía del cliente, únicamente se cuenta con los datos demográficos y las respuestas del cuestionario, por lo que se utilizará un modelo de clustering (agrupamiento), donde se asigna qué clusters corresponde a cada huésped en función de los datos ingresados. Una vez identificado el cluster del cliente, se realizarán anuncios personalizados de acuerdo a los patrones de consumo identificados en huéspedes anteriores del mismo cluster.
2. Transcurrido el período inicial, se cuenta con registros de consumos y lugares visitados por el cliente actual. En esta instancia se puede realizar un análisis de asociación en base a los patrones de consumo de huéspedes anteriores, analizando por ejemplo qué lugares visitaron o qué consumos realizaron juntos. De esta forma, podemos aplicar los descubrimientos realizados en patrones de consumo de huéspedes anteriores, y en base a los patrones de consumo del cliente actual realizar recomendaciones.
3. Por último, para la selección de hoteles para el mailing privado durante el verano, podemos aplicar técnicas de clasificación para la selección de ofertas. Para esto, utilizamos como datos de entrenamiento los datos de clientes anteriores (tanto la información demográfica y cuestionario ingresados al inicio de la estadía, así como los patrones de consumo registrados durante la misma). Sobre estos huéspedes anteriores contamos con la elección de hotel que realizaron para el verano, por lo que podemos realizar un entrenamiento supervisado de un modelo de clasificación, y aplicarlo a los clientes actuales para seleccionar las ofertas relevantes a enviar. A modo ilustrativo en la figura 3 se describe el modelo a utilizar, teniendo en cuenta de realizar una división de los datos del dataset para entrenamiento y testeo

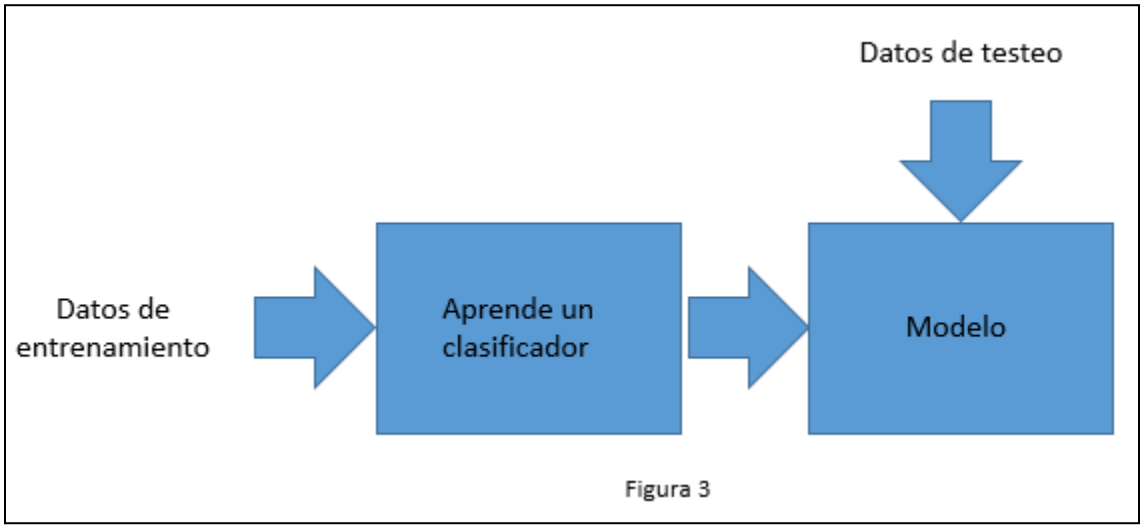

Figura 3

Etapa 5 – Presentación de ofertas

En función de las técnicas indicadas en la etapa 4 (puntos 1, 2 y 3), se presentan las ofertas relevantes surgidas de las mismas:

1. Para los primeros días se emiten ofertas de acuerdo al cluster que el modelo asoció al nuevo cliente y los patrones de consumo de clientes anteriores que pertenezcan a ese cluster.
2. Cuando ya contamos con datos de transacciones, utilizamos el análisis de asociación de clientes anteriores para emitir ofertas relevantes al cliente actual en base a sus consumos registrados.
3. Para la selección de ofertas para el verano, utilizamos el modelo de clasificación detallado en la etapa 4 para clasificar al cliente y enviarle ofertas relevantes.