UNIVERSIDAD AUSTRAL DE ARGENTINA

Materia
Introducción a Data Mining

Profesor
Mg. Gaston Pezzuchi, MSc.

Programa
MAESTRÍA EN EXPLOTACIÓN DE DATOS Y GESTIÓN DEL CONOCIMIENTO

Alumnos
Hernán Ifrán
Damián Joglar
José Eduardo Valdés Castro
Marcela Distefano

Noviembre 2022

TABLA DE CONTENIDO

LISTADO DE TABLAS

LISTADO DE ILUSTRACIONES

CUADRO DE CONTROL DE CAMBIOS

No. VERSION	FECHA	MODIFICACIÓN
Versión 1.1	Noviembre 27 de 2022	Primera versión del caso No. 1 de la materia Introducción a Data Mining

INTRODUCCIÓN

El presente trabajo se realiza en el desarrollo de la materia Introducción a Data Mining con el objeto de entender el proceso KDD a través de un caso de estudio que se aborda de manera conceptual, este caso corresponde al primer trabajado en la materia en mención y tiene que ver con una cadena de hoteles que pretende implementar un proyecto en minería de datos.

Caso IDM-1

Considere el siguiente caso:

Una cadena de hoteles que gestiona numerosos establecimientos en varios países registra información acerca de sus huéspedes en los diferentes hoteles. La gerencia desea implementar un **proyecto de Data Mining** a efectos de extraer el máximo provecho de esos datos. Más precisamente, la cadena desea conocer mejor a sus clientes de manera de poder informarlos sobre eventos especiales, promociones especiales, etc., durante su estancia como así también después de esta.

A la llegada de un huésped, se registran sus datos demográficos y se le pide completar un cuestionario cuando se le hace entrega de una credencial que le permite ingresar a distintos lugares recreativos tales como piscinas, bares, etc. Esos lugares son gratuitos, pero vía la credencial es posible registrar cuando un huésped hace uso de alguno de esos servicios. Además, la credencial sirve como tarjeta de crédito para pagar ciertas bebidas como así también para pagar productos en las pequeñas tiendas que la cadena posee. Todas esas transacciones son registradas en una base de datos.

Por medio de su canal de TV privado, el hotel informa a sus huéspedes sobre los próximos eventos, actividades y promociones especiales, etc. Sin embargo, puesto que los huéspedes pasan la mayor parte de su tiempo afuera y no mirando TV, el hotel necesita un sistema para enviar anuncios altamente personalizados de manera de garantizar que sólo se envíen anuncios en los que el huésped esté posiblemente interesado. Además, el sistema de data mining también tiene que dar una sugerencia razonable sobre qué anuncios enviar a un huésped particular ya durante los primeros días de su estadía, cuando el sistema tiene pocas posibilidades de aprender los hábitos de ese huésped particular. Después de la estadía, se le solicita al huésped que complete un formulario de evaluación. Este formulario contiene una lista de preguntas en las que se debe consignar un puntaje de 1 a 5. Para cada respuesta se pueden explicar los motivos o agregar comentarios adicionales breves.

Durante el invierno, la cadena envía publicaciones de una selección de sus hoteles a antiguos huéspedes, para obtener nuevas reservas en uno de sus hoteles. La selección para este mailing personalizado tiene que ser hecha de modo tal que sólo las publicaciones de los hoteles más interesantes para un huésped particular son enviadas a ese huésped. Por lo expresado, es importante considerar que en la mayoría de los casos un huésped de hotel no elige muchas veces exactamente el mismo hotel.

En conclusión, el sistema de data mining debe proporcionar información para:

1. Decidir cuáles anuncios de TV van a ser enviados a los huéspedes durante su estadía; un subproblema de este primer problema es decidir cuales 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.

Recorrer los diferentes pasos del proceso KDD, explicando cómo se aplican en este caso. Indicar para cada paso cuáles técnicas usaría y justificar su elección.

SOLUCIÓN DEL CASO

Para el desarrollo del caso se tiene en cuenta las siguientes consideraciones:

- El almacenamiento de datos utilizado por la cadera de hoteles es en la nube. Este está compuesto por los database de cada una de las sucursales.
- Debido a la información inicial que la compañía hotelera está localizada en distintos países, se asume que el canal de TV privado referenciado es un canal difundido por la web a través de streaming y el contenido difundido es generalizado de los aspectos de la cadena, esto debido a que aún no se implementa el proyecto de data mining. Este mismo contenido particularizado al país donde opera cada hotel, es presentado en todas las zonas sociales del complejo en cada una de sus localidades.

Teniendo en cuenta el proceso de knowledge discovery in databases (KDD), referenciado en el enunciado del caso para solventar la problemática referenciada, se cuenta con la siguiente estructura:

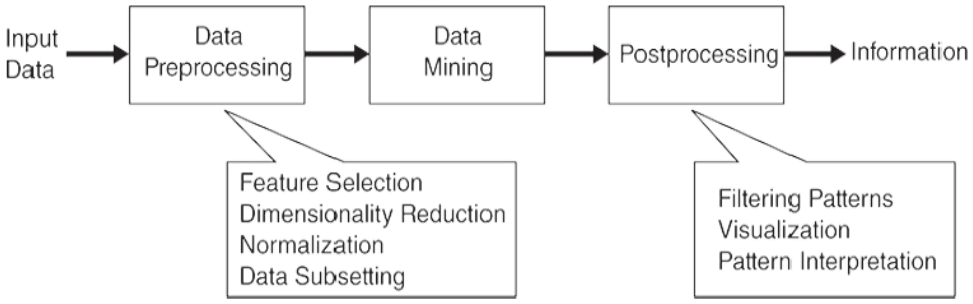

Figure 1.1.
The process of knowledge discovery in databases (KDD).

Ilustración 1: El proceso de descubrimiento de conocimiento en bases de datos. (KUMAR)

Como se observa en la imagen de arriba, las cinco fases del proceso KDD son los momentos en donde se produce el tratamiento de datos para poder implementar la solución en data mining. A continuación, se describe cada una de estas etapas:

1. INPUT DATA

Como se informó inicialmente se asume que la base de datos de la compañía está localizada en la nube. Además, la cadena de hoteles cuenta con un sistema de información con el cual realiza su actividad económica, este permite la adquisición de información a nivel general del proceso desempeñado.

De los datos recolectados en el proceso de la actividad comercial se obtiene información esencial que permite identificar a los clientes, estos pueden ser catalogados como datos comunes y acorde a la implementación del proyecto de data mining se adquirirán otros datos. En tal sentido, se podría clasificar los datos recolectados de la siguiente forma:

- Datos comunes: La recolección de información en este apartado se realiza a través de formularios con datos básicos como el nombre del cliente, número de identificación, fecha de nacimiento, lugar de residencia, tipo de habitación a facturar, tiempo de estadía, servicios prestados, motivos del hospedaje, época del año, entre otros.

  Se aclara que, al contar con la base histórica de cada uno de los hoteles y está al contener una gran cantidad de información se puede llegar a perfilar a los usuarios que vuelvan a utilizar los servicios de la cadena hotelera. Este perfilamiento serviría de datos iniciales para ofertar servicios de los huéspedes conocidos y en cada una de las visitas actualizar el perfil con los nuevos datos recolectados.

- Otros datos:
  - Credencial: A través de este medio entregado a los huéspedes para facilitar la estancia, acceso y uso de los servicios ofrecidos por la compañía a los clientes en la cadena de hoteles, se recolectará información de todos los eventos cursados por los huéspedes, entre estos se podrá conocer gustos y/o preferencias al momento de usar los productos ofertados, identificando servicio utilizados o producto adquiridos, horario de la solicitud, cantidad de dinero gastado, entre otros datos generados del uso del instrumento entregado.
  - Formulario de evaluación: Este instrumento permitirá recolectar datos de la percepción de los clientes con motivo de los servicios recibidos durante su estancia en el hotel. En este se registrará un puntaje de satisfacción, preguntas tales como ¿si volvería a hospedarse?, opinión de servicios, si recomendará el hotel y unas observaciones y así lo decidirá. Aunque este formulario no es obligatorio, asumimos que la mayoría de los clientes lo diligencia.

De la información recolectada por la cadena de hoteles que estaría en línea y con una disponibilidad alta, debido al manejar servicios en la nube, se realizaría la conexión y/o consulta de los datos generados para inicial el proyecto de data mining referenciado en el caso de estudio. En tal sentido, hasta este punto es claro que se tendría el acceso a toda la información generada por la cadena de hoteles y con ella poder realizar las siguientes etapas del KDD para lograr implementar un proyecto de ciencias de datos exitoso.

2. DATA PREPROCESSING

En esta etapa posterior a la adquisición de datos, se buscaría utilizar técnicas tales como la limpieza de datos, selección de variables relevantes a estudiar, integración de distintas bases, detección y tratamiento de outliers.

Con respecto a la reducción de las variables relevantes que se van a tomar será dependiendo a donde se busque enfocar el análisis. En nuestro caso buscaremos sugerir servicios al cliente en base a su historial de consumos y pronosticando en base a las variables elegidas que serán el país, cantidad de estrellas de hotel elegido, la temporada (debido a que no es lo mismo alojaste en temporada baja que una alta), que tipo de servicios consumió, en que frecuencia, que tipo de actividades realizó, horarios de salida, tiempo de estadía, tipo de habitación, consumos recurrentes y la cantidad de personas (dependiendo el grupo se pueden ofrecer servicios familiares o no).

Una aplicación importante será la reducción de la dimensión, esto debido a que, si tomamos una base histórica de los clientes, nos vamos a encontrar con cambios de preferencias o gustos. Por ejemplo, si una persona de 20 años, hace 10 años optó por una estadía con servicios tales como gimnasio, bares o fiestas, siendo el caso de que hoy vuelva con familiares, es muy probable que utilice otro tipo de servicios. Otro motivo de recortar la mirada hacia atrás es el avance de la tecnología ya que hoy un cliente hace 10 años no le interesaba tener wifi gratuito, hoy es casi excluyente que lo tenga, o puede que la televisión ya no sea preferencia cuando antes era esencial. En tal sentido, toda la información histórica que contenga la compañía permitirá realizar un primer perfil del cliente, que será actualizado acorde a los nuevos datos que se generen en las nuevas visitas realizadas por clientes antiguos y la reducción de dimensionalidad permitirá tener las principales variables explicativas del usuario para alimentar el modelo desarrollado para el proyecto en mención.

3. DATA MINING

El acceso a los datos históricos en la base de los clientes va a permitir efectuar asociaciones de estos para confeccionar perfiles de los huéspedes. Esto es importante porque va a permitir pronosticar comportamientos que permitirán a la cadena ofrecer servicios dirigidos con alta probabilidad de aceptación por parte de los usuarios, estos perfiles serán actualizados con la nueva información recolectada en las nuevas visitas de los huéspedes. A continuación, se presenta las variables que serían objeto de tratamiento para el desarrollo del proyecto:

Cliente	Habitación	Actividades	Excursiones
Id	N°	Gym	Aire libre
Edad	Tipo	Masajes	Museos

Cliente	Habitación	Actividades	Excursiones
Tipo	Consumo	Pileta	Hop on -Hop off
Consumo Total		Restaurant	Night club
		Bar	

Tabla 1: Variables a tratar.

La clasificación de los clientes se desarrollará de la siguiente forma:

Tipo de cliente	Nivel de consumo
Convencionista	
Alto	> 200000
Medio	150000-200000
Bajo	0 - 100000
Vacacionista	
Alto	> 200000
Medio	150000-200000
Bajo	0 - 100000

Tabla 2: Tipos de clientes.

De esta forma se pueden proponer como técnicas la asociación en cuanto al comportamiento de los clientes para luego clusterizar en segmentos de acuerdo con el perfil observado y por último pronosticar cual podrá ser el comportamiento de cualquier cliente en una determinada situación. Con esta información ya se está en condiciones de proponer estrategias de marketing dirigidas a esas situaciones.

Existe la posibilidad de que no contemos con información del cliente, y será necesario ir sumando información a medida que vaya consumiendo.

En este caso, para lograr una mejor información en el clúster, se implementará un árbol de decisión que se desplegará en primera medida si el cliente ya fue a ese hotel, luego si vino con hijos o no. Esto permitirá recomendar lugares turísticos típicos o no tan típicos en el caso de que ya conozca la zona, se adaptará la recomendación a si es apta para familia, ya que no se podrá recomendar un recorrido de trekking de 10 km a una familia con niños pequeños.

En el caso de que el cliente ya haya estado en el hotel, el árbol se desplegará nuevamente para saber si está en familia, y se le recomendará excursiones diferentes y relacionadas a las veces anteriores. Además, en base a consumos anteriores, se clusterizará.

4. POSTPROCESSING

En esta etapa se evaluarán los resultados obtenidos y se determinará realizar re-clusterización en función de variables de consumo total en el hotel, esto siempre y cuando si los patrones que resultaron no lograban descubrir lo que inicialmente se estaba buscando, como tipos de actividades realizadas, ofertas aceptadas, entre otras.

En la ilustración 2 se puede observar una posibilidad de cómo se verían los clústers una vez obtenido el resultado del modelo aplicado.

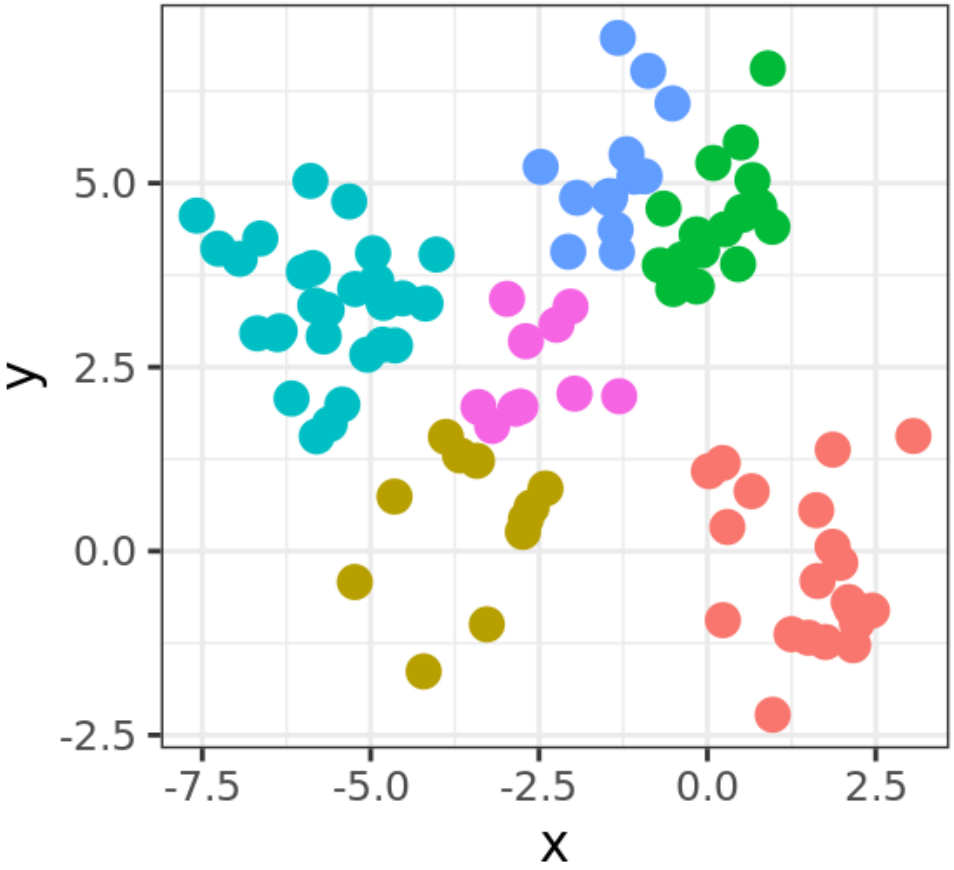

Ilustración 2: Grafico KNN, k=6.

En base a los consumos que realicen vamos a aplicar reglas de asociación, se le va a ofrecer servicios relacionados, por ejemplo, si hace salidas nocturnas, se le ofrecerá el bar del hotel.

La salida de este proceso de asociación en base a los comportamientos permitirá segmentar los clientes en función de sus preferencias. Por ejemplo: no se malgastará dinero en publicitar el bar del hotel a una familia con niños pequeños.

Esto también permitirá distinguir diferentes patrones de consumo que pueden ser bastantes obvias pero que abrirán la puerta a una mejor calidad en las recomendaciones. Un ejemplo de esto es que si el cliente utilizó el Gym y seguido de esto tomo una sesión de spa o masajes. Se podrá ofrecer servicios relacionados a estos consumos.

Un patrón es interesante si es (1) fácilmente comprensible por humanos, (2) es válido en datos nuevos o de prueba con cierto grado de certeza, (3) es potencialmente útil y (4) novedoso. Un patrón también es interesante si valida una hipótesis que el usuario pretendía confirmar. Un patrón interesante representa conocimiento. (Han, 2012)

Se determinó que las ofertas para cada tipo de cliente se estructuren de la siguiente manera y con los códigos generados por la cadena de hoteles se buscará asociar las soluciones del proyecto a los grupos de clientes que se determinen. A continuación, se presenta las clasificaciones asociadas:

Cod.	Propuestas
1	Gym
2	Masajes
3	Pileta
4	Restaurant
5	Bar
6	Parques temáticos
7	Museos
8	Hop on off
9	Night club
10	Recorridos turísticos
11	Salón de juegos
12	Paseo de compras
13	Recorridos de día completo
14	Alquiler de auto
15	Alquiler de bicicletas

Tabla 3: Propuestas de la cadena de hoteles.

		Ofertas	
Tipo de cliente	Nivel de consumo	Actividades	Excursiones
Convencionista	Alto	1, 2, 3, 4, 5, 14	12
	Medio	1, 3, 4	
	Bajo	3, 4	
Vacacionista	Alto	14	12

		Ofertas	
Tipo de cliente	Nivel de consumo	Actividades	Excursiones
	Medio	3,11	6,7
	Bajo	3,15	8,10

Tabla 4: Ofertas de la cadena de hoteles.

Se mostrará la información recabada con sus conclusiones de manera amigable y entendible para él área de negocio, en gráficos que no contengan términos demasiado técnicos para un mejor entendimiento.

Utilizando Looker Studio se realizar un tablero que permitirá la toma de decisiones del área de ventas. A continuación, se presenta un ejemplo del tablero que se observaría:

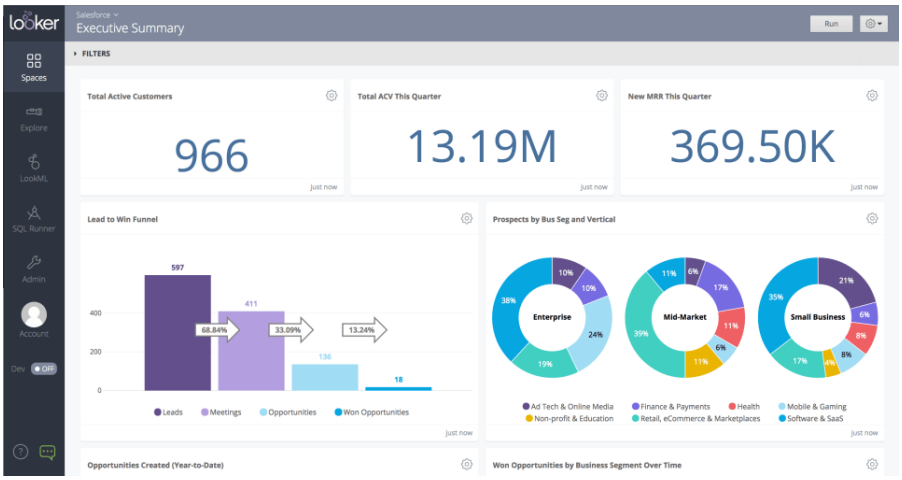

Ilustración 3: Ejemplo de interfaz Looker Studio.

Con este tipo de soluciones se pueden observar fácilmente alertas que permitan determinar propuestas a grupos de clientes o resultados de clientes insatisfechos para mejorar las ofertas que se estén presentando.

En función del conocimiento generado se determinan las publicidades a visualizar considerando a los diferentes tipos de clientes.

5. INFORMATION

En esta etapa se le presenta al área de negocios las conclusiones del proceso. Es decir, las sugerencias de marketing para cada tipo de cliente en función de los hallazgos luego de la aplicación de las técnicas mencionadas. Además, se presentarán las visualizaciones para una mejor comprensión de las conclusiones.

Preguntas para solventar por parte del sistema:

1. *Decidir cuáles anuncios de TV van a ser enviados a los huéspedes durante su estadía; un subproblema de este primer problema es decidir cuales anuncios van a ser enviados durante los primeros días de la estadía de un huésped.*

Respuesta: Dentro de los anuncios de tv que disponga el hotel se clasificarán entre publicidad general y publicidad particular, esto teniendo en cuenta la ubicación de cada hotel.

Se aclara como se anunció al inicio del proceso que este canal de tv es transmitido vía streaming lo que permitiría presentar dicha publicidad a través de las redes sociales de los clientes. Partiendo del hecho que los huéspedes durante su estancia estarán poco tiempo dentro de las habitaciones y/o contarán con poco tiempo para atener los anuncios tv que disponga los hoteles en las zonas comunes en los tv físicos. Esta forma de presentar la información permitirá que inicialmente se presente el contenido a los clientes de forma generalizadas, la misma publicidad que se transmite en los tv físicos de los hoteles será presentada en las redes sociales de los usuarios. Esta publicidad generalizada será presentada también a los usuarios del hotel que adquieran los servicios por primera vez, esto prioritariamente porque no se tendrían datos para categorizar al usuario hasta que se vayan adquiriendo información que permita asociarlo a una de las categorías establecidas.

Posterior a conocer datos particulares del huésped, se ira presentando información de acuerdo al tipo de cliente que hace usos del servicio, utilizando toda la información que se levante en cada momento que se encuentre ubicado como huésped.

Para el caso de los clientes cuando terminan su estancia en el hotel, la publicidad de tv que se enviará a sus redes sociales está encaminada a la determinación del tipo de cliente en la que fue catalogado, de esta forma se realizará una actividad post comercial para buscar que en algún momento vuelva a hacer uso de los servicios del hotel, presentando tanto publicidad del segmento de usuario que fue catalogado, como promociones que se puedan desprender del grupo correspondiente.

2. Decidir sobre la selección de hoteles para el mailing privado durante el verano.

Respuesta: Siguiendo con el análisis del caso, en este punto se van a tomar decisiones concernientes a los correos a enviar promocionando los distintos hoteles.

Se le enviará a cada cliente de manera personalizada en base a la regla de consumos que tuvo en su histórico y como fue clusterizado dentro de sus preferencias y condiciones, ofreciéndole ofertas que se ajusten a su interés. Dentro de esas preferencias, estarán los consumos realizados por temporada, que tipo de consumo presenta, tipo de actividad y servicio tomado, junto a los patrones específicos de cada cliente.

Un paso posterior que no entrará en análisis será la medición de cantidad de clics que se dan en la apertura a la publicidad enviada.

BIBLIOGRAFÍA

Han. (2012). *Data Mining: Concepts and Techniques.*

KUMAR, V. (s.f.). *Introduction to Data Mining.* Pearson.