UNIVERSIDAD AUSTRAL

INTRODUCCION A DATA MINING

CASO IDM-1

DOCENTES:

Gastón Pezzuchi
Eduardo Poggi

INTEGRANTES:

Fernando Damián Coz
Tamara S. Gásquez
Anibal Martín Glaniver
Johanna Noval

Noviembre 2022

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:

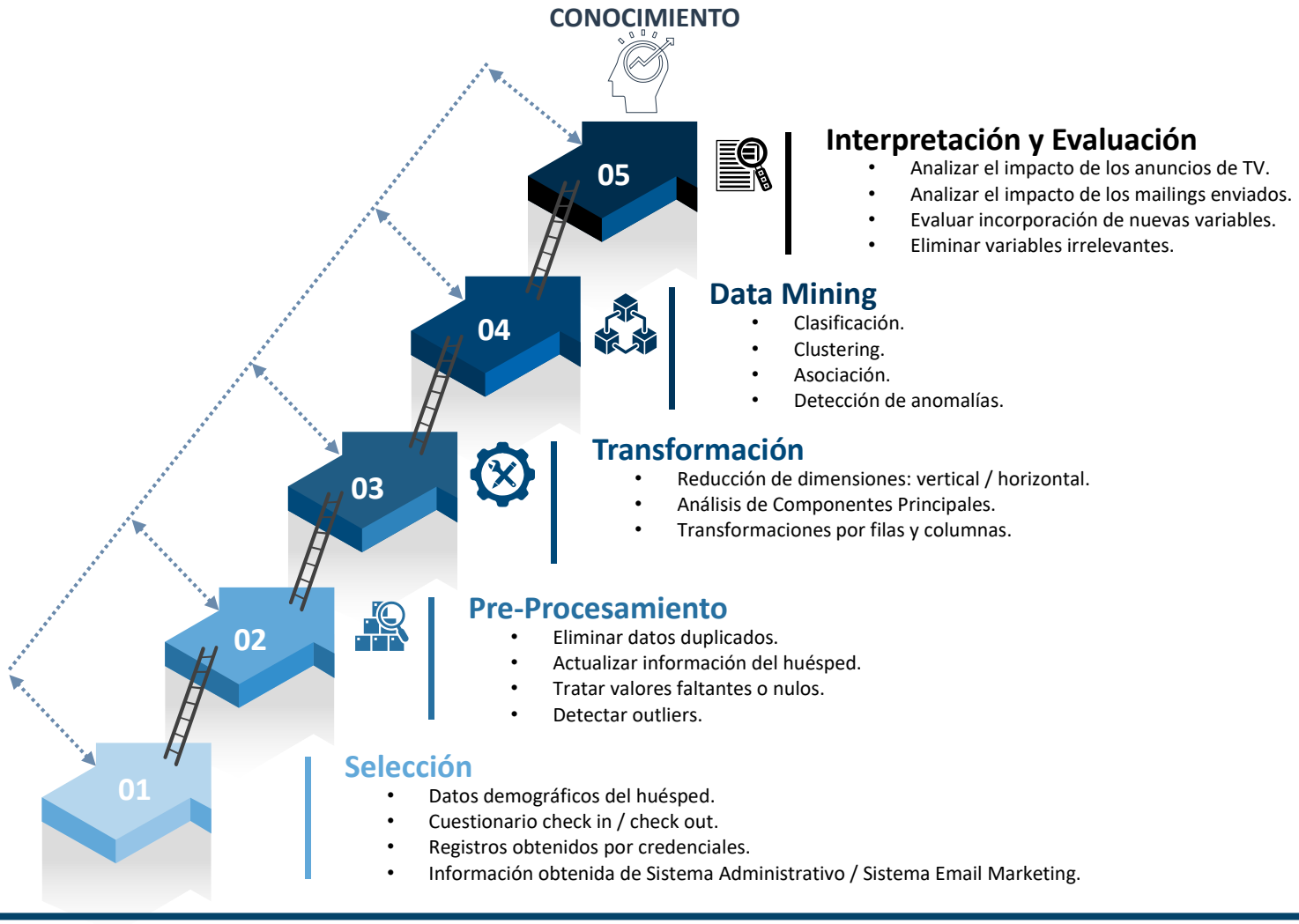
1. *Decidir cuáles anuncios de TV van a ser enviados a los huéspedes durante su estadía; un subproblema de este primer problema es decidir cuales anuncios van a ser enviados durante los primeros días de la estadía de un huésped.*
2. *Decidir sobre la selección de hoteles para el mailing privado durante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

El proceso Knowledge Discovery in Databases (KDD), es utilizado para llevar a cabo la extracción automatizada de conocimiento partiendo de grandes volúmenes de datos. Es de naturaleza iterativa, por lo tanto, es aplicable tantas veces como sea necesario hasta obtener la información requerida. El proceso KDD tiene como motivación la detección de información no trivial y potencialmente útil que permita resolver los problemas o necesidades que surgen de las empresas.

PROCESO KDD

Aplicado a una cadena hotelera:



Como puede apreciarse en la figura anterior, visto como un proceso de descubrimiento de conocimiento, los pasos del proceso KDD para el caso propuesto son:

SELECCIÓN

Esta etapa consiste en seleccionar los datos, sobre los cuales se llevará a cabo la tarea de DM. Los datos con los que cuenta la cadena de hotel son:

a. Los datos demográficos del huésped. Asumimos que los datos con los que cuenta son: nombre, apellido, domicilio, DNI, correo electrónico, teléfono, estado civil, profesión, género, nacionalidad, fecha de nacimiento.

b. El cuestionario que completa al hacer el check in. Asumimos que los datos con los que cuenta son: el motivo del viaje, con quién viaja, dieta alimentaria, preferencia de horario de limpieza, motivo por el cual eligió el hotel.

c. Los registros de la base de datos relacionados con el uso de la credencial. Por un lado, el detalle de los servicios del hotel que usó (piscina, spa, gimnasio, restaurante, etc.) y horario en que usa los servicios. Por otro lado, las compras y transacciones realizadas con la credencial (tiendas en qué realizo compras, ítems comprados, importe gastado).

d. El formulario de evaluación que completa al finalizar la estadía. Suponemos que la encuesta realiza preguntas sobre los siguientes temas: nivel de satisfacción con los servicios del hotel, equipamiento de la habitación, la ubicación, la limpieza, mantenimiento del hotel, los servicios brindados, el personal, el restaurante, si volvería a alojarse en el hotel, concordancia con las fotos del hotel mostradas al momento de la reserva.

Asimismo, si bien no está detallado en el enunciado, consideramos que también se podría registrar información obtenida del sistema administrativo de la cadena de hoteles:

e. La duración y temporada de la estadía. Es decir, cuánto tiempo estuvo alojado en el hotel y en qué fechas. Esto nos permitirá detectar aquellos huéspedes que siempre reservan en la temporada de verano o de invierno, y brindar ofertas y publicidades más segmentadas.

f. En cuáles hoteles de la cadena suele alojarse habitualmente.

Si el huésped ya se registró previamente en algunos de los hoteles, también podríamos tener información del sistema de email marketing, tales como:

g. Qué mails abrió y cuales no, si hizo clic en alguno de los links que le enviamos, si reservó a través del correo enviado o no y si se dio de baja en la suscripción.

PRE-PROCESAMIENTO

En esta etapa realizamos la limpieza y el preprocesamiento de las observaciones, para obtener datos consistentes. Las tareas por realizar son:

a. Eliminar datos duplicados. Por ejemplo, podríamos detectar que un huésped que se haya alojado en más de una oportunidad en algún hotel de la cadena, en lugar de actualizar el registro de dicho cliente, se hubiera cargado un nuevo registro, duplicando esta observación.

b. Actualizar la información del huésped. Por ejemplo, si un huésped se vuelve a hospedar en algunos de los hoteles, confirmar o actualizar la información, tales como el domicilio, información de contacto, profesión, estado civil, hijos.

c. Tratar valores faltantes o nulos. Puede suceder que existan huéspedes que sean reacios a brindar cierta información y dejan campos de la encuesta o del formulario sin completar, o bien que por algún tipo de error no se hayan relevados estos datos, o los mismos no hayan sido cargados al sistema.
   Cuando existen datos faltantes en el conjunto de datos, se debe decidir cómo actuar. De esta forma, siguiendo a *Tan,P-N, et. al.(2018)*, si faltasen algunos datos categóricos, tales como sexo, nacionalidad, estado civil, domicilio (provincia), se podrían reemplazar con la moda. En cambio, si los valores nulos se correspondiesen con una variable cuantitativa, tales como como la edad, cantidad de hijos, cantidad de mascotas, calificación de la encuesta final, se podría reemplazar con la media o la mediana.

d. Detectar si existen outliers. Los outliers pueden generarse por un error del data entry, en cuyo caso, se eliminarían. O bien, los valores atípicos también pueden deberse a anomalías las cuales no deben ser eliminadas, pero sí integrarse con el resto de los datos (por ejemplo, usando técnicas robustas).
   Los outliers pueden corresponderse con errores de carga (por ejemplo, valores fuera de escala en la encuesta de satisfacción), en cuyo caso deberíamos eliminar estos valores (no la observación completa, sino sólo el dato en particular), o bien, reemplazarlos por la moda o la media. También pueden deberse a anomalías (por ejemplo, una familia con una gran cantidad de hijos), en cuyo caso, podría ponderarse para quitarle peso en el cálculo de los parámetros estadísticos.

Dependiendo de la forma en la cual se recolectan los datos, el preprocesamiento puede ser más simple o más complejo. Por ejemplo, si el cuestionario que llena el huésped es manual y luego un empleado del hotel tiene que cargar dicha información en la base de datos, pueden existir errores de carga del empleado, o bien podría perderse información por no entender la letra del cliente. Mientras que, si el formulario es completado por el cliente de manera digital, y tiene algún tipo de validación (por ejemplo, un máximo de edad permitida), podría ayudar a que existan menos errores y a evitar datos nulos.

TRANSFORMACIÓN

En esta etapa, buscamos características útiles para representar los datos dependiendo de la meta del proceso. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos. Los métodos de reducción de dimensiones pueden simplificar una tabla de una base de datos horizontal o verticalmente.

La reducción horizontal implica la eliminación de tuplas idénticas como producto de la sustitución del valor de un atributo por otro de alto nivel, en una jerarquía definida de valores categóricos o por la discretización de valores continuos (por ejemplo, edad por un rango de edades). La reducción vertical implica la eliminación de atributos que son insignificantes o redundantes con respecto al problema, como la eliminación de llaves, la eliminación de columnas que dependen funcionalmente (por ejemplo, edad y fecha de nacimiento).

A partir de las variables recolectadas previamente, podríamos llegar a encontrar cuáles se encuentran correlacionadas y cuáles no. Por ejemplo, en base a la información recolectada podemos detectar que las personas que viajan con niños pequeños suelen valorar que el hotel tenga restaurante, pileta, juegos y actividades para ellos.

Podríamos detectar también variables que no aportan información nueva o relevante, por ejemplo, el DNI, el teléfono o el correo electrónico.

En caso de detectarse que existen variables con un alto nivel de correlación lineal, podríamos usar el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad y facilitar el análisis de los datos.

Se podría realizar también transformaciones por filas, las cuales tienen como objeto hacer comparables los valores asignados a los distintos individuos u objetos de análisis. Estas transformaciones podrían aplicarse en el cuestionario final que completaron los huéspedes, respecto de las calificaciones recibidas. Por ejemplo, podría suceder que un huésped se haya peleado con su pareja o haya sufrido algún robo o accidente durante el viaje y, por ende, va a tener más tendencia a dar una baja calificación, o que un huésped haya ganado en el casino, y nos dé una calificación más alta.

Las transformaciones en columnas se utilizan también para hacer comparables las magnitudes. Por ejemplo, si antes los formularios se completaban con una calificación del 1 al 10, y ahora la calificación que se pide es del 1 al 5, a los fines de poder comparar y trabajar con estas diferencias de escala, se podría realizar una estandarización de la variable.

DATA MINING

En esta etapa, la cual es la más representativa del proceso de KDD, buscamos los patrones de interés a través de la utilización de algoritmos.

"La minería de datos es el proceso de descubrir patrones y conocimientos interesantes a partir de grandes cantidades de datos." (Han, J., Kamber, M., 2005).

Clasificación

En base a la información recolectada del huésped, podríamos clasificarlos y asociarlos a una etiqueta en concreto: viaje familiar, viaje de negocio, tiempo de estadía, tipo de uso de las instalaciones del hotel, dieta alimentaria, movilidad (si tiene auto o no), cliente nuevo o frecuente.

Anuncios de publicidad en TV: De esta forma, podemos orientar los anuncios que se muestran en la TV según las preferencias de cada etiqueta. Por ejemplo, durante los primeros días de la estadía, cuando aún no tenemos mucha información sobre el huésped, podríamos considerar las variables como movilidad, con quien viaja y el motivo de viaje (datos con obtenidos en los formularios de ingreso) para segmentar los anuncios. Por ejemplo, si viaja con familia podemos mostrar publicidad sobre actividades para chicos, y si viaja por negocios, podemos ofrecerle los servicios del bar del hotel.

Durante los últimos días de la estadía, a medida que tenemos información sobre los servicios del hotel que usa y las transacciones y compras que realizó, podemos brindar publicidades más específicas. Por ejemplo, si el cliente suele ir habitualmente al bar del hotel, podemos ofrecerle alguna promoción o descuento en bebidas.

Envíos de mails: De esta forma, si se detecta que un huésped suele viajar en familia, y se tiene información previa de que en esta etiqueta (viaje familiar) se valoran las instalaciones del hotel de piscinas, juegos para chicos y restaurante, podríamos enviar mailings con información de los hoteles que brinden estos servicios.

Clustering

La agrupación, en general como técnica de minería de datos, nos permitirá segmentar audiencias o poblaciones.

Anuncios de publicidad en TV: En base a información previa obtenida de un modelo aplicado, podríamos tener conocimiento de que es muy probable que las mujeres mayores a 50 años y divorciadas pasen más de 4 horas por día en el casino. Por ese motivo, aun cuando un huésped haya ingresado recientemente al hotel, si detectamos que, en función a los datos demográficos recolectados al momento de hacer el check in, encuadra en este cluster, enviarles anuncios de TV con publicidad del casino.

Envíos de mails: En el caso de estudio, suponiendo que el cliente ya se hospedó alguna vez en algún hotel de la cadena, podríamos agruparlos en función de este dato y enviarle mailings con descuentos en ese hotel para el verano. En cambio, si es un cliente nuevo, el cluster que podría realizarse es en función a los datos demográficos y al cuestionario que completó al realizar el check in.

En nuestra opinión, esta técnica de DM tiene más significancia cuando el huésped es recurrente, ya que, en base a toda la información recolectada, obtendríamos mejores agrupamientos.

Asociación

Este tipo de técnica, al igual que sucede con los sistemas de recomendación, permiten predecir la preferencia de los usuarios.

Anuncios de publicidad en TV: En base a información previa, podríamos haber detectado en qué horario se suelen usar los servicios del hotel: el gimnasio se suele usar a la mañana, el restaurante al mediodía, la pileta a la tarde y el casino a la noche. Considerando estas asociaciones, los anuncios de TV que se muestran, podrían estar vinculados con el horario y las actividades a realizarse en el mismo.

Envíos de mails: Asimismo, para enviar mails de promoción para alojarse en los hoteles durante el verano, se podría asociar con las instalaciones que más se usan en dicha estación, tales como la piscina, servicio de playa. Por otro lado, también tendríamos en cuenta si los servicios del hotel se ajustan a la estación del verano. Por ejemplo, si el hotel se encuentra en un centro de esquí, no lo tendría en cuenta para incluirlo en las publicidades a enviarse para la época estival.

Detección de anomalías

Nos permitiría identificar comportamientos inusuales de los usuarios.

Anuncios de publicidad en TV: En base a la información obtenida de los huéspedes, por ejemplo, si es hotel en el que suelen hospedar personas que viajan por negocios, pero se alojaron muchas familias con niños, podríamos ofrecer algún servicio o actividades para niños y promocionar las mismas por TV a estas personas que viajan en familia.

Envíos de mails: A modo de ejemplo, si detectamos que tenemos muchas reservas para el verano, en un hotel que se encuentra ubicado en un centro de esquí, podría deberse a un evento que se va a realizar en dicha zona, o por descuentos muy altos en los pasajes aéreos hacia dicha localidad. Esto podría tenerse en cuenta para enviar promociones sobre dicho hotel, aunque los servicios que ofrece el mismo no se ajustan a la época estival.

INTERPRETACIÓN Y EVALUACIÓN

Esta etapa nos permitirá tomar decisiones luego de interpretar y evaluar los resultados obtenidos.

Anuncios de publicidad en TV: Respecto de las recomendaciones que se pasaron por TV a los huéspedes, detectar si las mismas tuvieron impactos. Por ejemplo, si luego de mostrar una publicidad del restaurante del hotel, el huésped cenó allí.

Envíos de mails: De igual forma, respecto de la campaña de envío de mails con promociones para el verano, podríamos evaluar cuáles clientes efectuaron una reserva.

En ambos casos, si la campaña no tuvo el impacto esperado, deberíamos evaluar si necesitamos incorporar nuevas variables en el análisis (agregando nuevas preguntas al cuestionario, recolectar más información del uso de los servicios), identificar cuáles son los atributos que no agregan valor, y eliminar dichas variables del modelo, del formulario y de las encuestas utilizados para obtener los datos.

BIBLIOGRAFÍA

- Han, J., Kamber, M. (2005). *Data mining: concepts and techniques.* San Francisco. Kaufmann. ISBN: 1558604898 9781558604896. Pag. 8.

- Tan, P-N.; Steinbach, M.; Karpatne, A.; Kumar, V. (2018): *Introduction to Data Mining.* 2nd Edition, Addison Wesley, Boston, MA, ISBN 978-0133128901. Pag. 163.