UNIVERSIDAD AUSTRAL

# Introducción a Data Mining

## Caso IDM-1


**Integrantes:**

1. Falcones Johanna
2. Matsusaka Daniel
3. Rodríguez Diego
4. Rodríguez José Antonio

## Introducción

La importancia de la adaptación al cambiante entorno empresarial hotelero es una etapa de transición crucial no solo para seguir siendo competitivos ante la creciente oferta, sino también para poder sobrevivir en momentos de crisis como por ejemplo la pandemia. Como resultado de la presión ejercida por los constantes cambios, la tecnología se ha convertido en una herramienta de creciente importancia. En la industria hotelera, la acumulación y generación de datos es gigantesca siendo esta una fuente de información valiosa para todo el sector. A partir de la información obtenida se producen nuevos fragmentos de datos que se analizan por separado, y que al asociarlos se genera un nuevo conocimiento que en manos de un experto (data scientist) y bajo las herramientas analiticas adecuadas obtendra un poderoso conocimiento para las futuras operaciones del hotel.

Con el fin de desarrollar este conocimiento se ha desarrollado el *Knowledge Discovery in Database* (KDD) que tiene la función de extraer patrones de comportamiento válidos, novedosos y útiles para posteriormente ser analizados por el usuario y aportar de manera significativa en la toma de decisiones empresariales. Por ejemplo, crear campañas de forma personalizada vía mail, planificar promociones de temporada a los clientes, crear anuncios personalizados, nicho de mercado donde se quiere establecer o ingresar la empresa hotelera, saber qué segmentos de mercados están creciendo más rápidamente, determinar la logística necesaria en cada temporada con base en los clientes potenciales, retener e incrementar a los clientes " leales" .

El objetivo del presente trabajo es realizar la propuesta de un sistema de data mining que proporcione información para poder decidir sobre los anuncios de tv enviados a los huéspedes durante su estadía y mails con propuestas de hoteles perteneciente a la cadena que serán enviados a antiguos huéspedes.

**Consideración previas:**

a. Se otorga una credencial a cada miembro del grupo familiar en el caso de que no sean huéspedes individuales.
b. El cuestionario del huésped es cargado por él mismo en un sistema interno mediante una tablet del hotel.

A continuación recorreremos los diferentes pasos del proceso KDD, explicando cómo se aplica cada uno en este caso:

### 1. Data cleaning

Antes de iniciar el proceso de Data Cleaning es importante tener un entendimiento claro de la información contenida dentro de nuestro dataset, para así poder definir qué información es relevante en nuestro caso de estudio. Como primer paso se debería analizar si la fuente de información tuvo origen a partir de una migración anterior al sistema, en ese caso se deberían limpiar errores de inconsistencia que hayan surgido del proceso de migración; estandarizar los datos, y así editar aquellos valores no esperados que surgieron por un error de tipeo o data *input*. Seguidamente se deberá identificar y depurar todos aquellos errores de carga por parte del huésped en su cuestionario. Otro de los escenarios que deben formar parte del data cleaning, son aquellos movimientos

que pudieron haberse originado por transacciones no asociadas a huéspedes, como ser el uso interno del staff, todo esto con el fin de no tener en cuenta esos datos al momento del análisis.

## 2. Data integration

La cadena hotelera tendrá dos tipos de data sources; el primero que será la base de datos del hotel, la que contendrá los datos de registros (demográficos), cuestionarios cargados en el sistema interno del hotel y los gastos mediante la credencial. Mientras que los datos del uso de los servicios gratuitos con la credencial, se los extraerá de un archivo plano contenido en el dispositivo que da acceso al servicio, el cual posteriormente será migrado a la base de datos del hotel.

## 3. Data selection

A través de la información recopilada mediante los datos demográficos registrados, el cuestionario y la información transmitida por la credencial entregada, se tiene un importante banco de datos con los registros de diversos clientes; lo cual nos permitirá conocer cuáles son los principales intereses, gustos, preferencias y las necesidades particulares de cada cliente, como así también el del grupo familiar, con el fin de determinar a qué tipos de segmento podemos orientar nuestra publicidad según preferencias y estilo de vida, generando valor agregado al cliente y una ventaja competitiva para el hotel.

Dentro de la base de datos del hotel existen atributos que no son relevantes para el estudio, por ejemplo el número de teléfono fijo, que actualmente es irrelevante para el caso en estudio. Por otro lado, reducir variables que puedan surgir de otras, como ser el caso de la fecha de nacimiento y la edad. El objetivo de este paso es realizar una selección de atributos que permitan representar a los datos según los objetivos trazados.

Como método de distinción de los datos más relevantes para el análisis, se puede emplear la selección de características, definiendo en primera instancia los atributos para la toma de decisiones de nuestro análisis. En este caso el diseño conceptual de un árbol de decisión con un split binario, permite elegir los atributos para llevar a cabo las divisiones entre los nodos internos.

## 4. Data transformation

En esta parte del proceso KDD se busca representar a los datos mediante características que favorezcan los objetivos trazados para el caso:

Una de las metodologías a ser aplicadas al proceso de transformación de los datos, corresponde a la discretización de ciertas variables relevantes para el caso de estudio, como ser el consumo por clase, el uso de los servicios, y la edad del usuario. Este proceso implicará convertir datos continuos en un conjunto de intervalos de datos. Los valores de atributos continuos se sustituyen por etiquetas de intervalos pequeños. Esto hace que los datos sean más fáciles de estudiar y analizar. Si una tarea de minería de datos maneja un atributo continuo, entonces sus valores discretos pueden ser reemplazados por atributos de calidad constante. Esto mejorará la eficiencia de la tarea.

Por ejemplo, los valores para el atributo de edad pueden ser reemplazados por etiquetas de intervalo como (0-10, 11-20…) o (niño, joven, adulto, senior) para asignar los algoritmos que permitan clusterizar al data set.

Otro de los métodos que se utilizará en el caso es Data Aggregation para poder disponer de un panel central entre todos los hoteles de la cadena que contenga una información útil, resumida, eficiente y así tener una ventaja estratégica frente a la competencia.

## 5. Data mining

En esta etapa procedemos a trabajar con las tareas de Data Mining basadas en el tipo de modelo que se desea emplear para el análisis hotelero.

Dado que el objetivo del análisis, es decidir, los anuncios que serán enviados a los huéspedes del hotel durante su estadía y, que la mayoría de información dentro de la base de datos corresponde a datos categóricos, es necesario partir de un modelo descriptivo que permita identificar patrones de comportamiento e intereses de los huéspedes. En este sentido, se pueden definir las siguientes tareas de Data Mining:

- *Reglas de asociación:* Determinar reglas de dependencia para medir la probabilidad de ocurrencia de un evento basándose en la ocurrencia de otros. Por ejemplo:
    - La combinación de si el huésped se encuentra solo o acompañado en el hotel, con la frecuencia de uso de espacios recreativos.
    - El incremento en los usos de servicios en los días con mal clima.
    - Uso bajo de los servicios cuando el motivo del viaje es por trabajo.
    - Permanencia en habitación por rango horario.
    - Motivo particular del viaje
  - Esta regla de asociación nos permitirá saber en qué momentos pasar publicidad en el televisor a determinados huéspedes y en determinados lugares.

- *Clustering:* Subdividir la muestra de huéspedes en subconjuntos objetivos para enviar las publicidades a través de la unificación de atributos y características similares.
    - Geográficos: País y ciudad procedente; es viajero frecuente.
    - Demográficos: Edades o intervalo de edades, género, con o sin familia, número de integrantes en la familia, viaja sólo o acompañado.
    - Sociales: el hospedaje es por trabajo, estudio o turismo.
    - Relación con el hotel: cliente frecuente (“leal”) o nuevo cliente.
    - Cómo llegó a conocer a la empresa: instagram, mail, publicidad, twitter etc. Red social más usada por el cliente.
  - Este método será utilizado para poder enviar mails con las distintas ofertas de hoteles durante la temporada baja a los distintos huéspedes registrados en la base y que forman los distintos subconjuntos.

## 6. Pattern evaluation

Para este caso, podemos realizar pruebas de accuracy para evaluar su precisión predictiva. Por ejemplo, un modelo diseñado para predecir quién responderá a una promoción debe basarse en una oferta anterior en la que se sepa quién respondió o no. Después de que se construye el modelo, se puede analizar un grupo "rechazado" de una

promoción anterior para verificar la confiabilidad. Si las predicciones de retención no replican los resultados de la promoción anterior, es posible que el modelo no sea significativamente predictivo. Para mejorar aún más la precisión, se puede asignar una puntuación al modelo en función del nivel de acuerdo entre el grupo reservado y todo el grupo. Los modelos refinados posteriores se pueden probar y puntuar.
Otra prueba que podría realizarse para evaluar su precisión predictiva es la del TestSet. Donde se comparará las predicciones de los tipos de publicidad y mails enviados a determinados clientes respecto de los servicios adquiridos por los mismos huéspedes en oportunidades anteriores.

## 7. Knowledge presentation

Para el knowledge presentation, se utilizarán diagramas de clústeres, donde se asociaría un grupo de potenciales huéspedes con cada instancia que podría representarse con distintos tipos de publicidades y sus respectivos canales de distribución. En nuestro caso, el agrupamiento permitirá que un huésped pueda pertenecer a más de un clúster, lo cual permitiría saber que a ese huésped se le podría ofrecer más de una oferta existente. Esta agrupación además se podría combinar con un árbol de decisión o con algún conjunto de reglas.

Es importante resaltar, que implementar recomendaciones obtenidas de un trabajo predictivo a nivel corporativo puede ser una tarea difícil, razón por la cual se debe brindar capacitación adecuada a todos los usuarios potenciales de los resultados de la minería de datos expuestos en los outputs arriba descritos.

Esto incluye a gerentes generales, gerentes comerciales de ventas y marketing, así como el staff de ventas. Se debe instruir a los usuarios sobre el tipo de output disponible y cómo interpretar correctamente la información. para considerar al proyecto de data mining exitoso.

## Conclusiones

Las herramientas de data mining permiten a las empresas identificar patrones de comportamiento en la información recabada ya sea de sus clientes o procesos, generando una oportunidad para la optimización de procesos, servicios o relación con los clientes. El *Knowledge Discovery in Database* (KDD) es un proceso que al ser interactivo y centrado en el usuario, posibilita obtener resultados enfocados en los objetivos que se fijen para el análisis.

El sistema de data mining construido a partir de las diferentes etapas del proceso KDD, decidirá cuáles anuncios de TV van a ser enviados a los huéspedes durante su estadía, considerando si el motivo del hospedaje es por trabajo o vacaciones, teniendo en cuenta el rango etario y franjas horarias en las que exista mayor probabilidad de que los huéspedes se encuentren en sus habitaciones.

Durante los primeros días de la estadía de un huésped, serán enviados anuncios generales relacionados a los servicios del hotel, eventos que tengan lugar dentro de la misma semana de la estancia del huésped, así como ofertas y descuentos teniendo siempre en consideración todos los criterios enunciados.

Finalmente, la selección de hoteles para el mailing privado durante el verano será realizada mediante la consideración de las preferencias anteriores de los huéspedes, como así también ofertas para fines de semana cortos donde posiblemente tenga algún tipo de viaje por razones de trabajo.