# Resolución Caso N°1

# Introducción al Data Mining

# Alumno: Cesar Germán Santamaria

**Objetivo:** Los cambios en las preferencias de los huéspedes influyen en el rendimiento de la cadena hotelera creando así la necesidad de que el management considere la aplicación de métodos que ayuden a predecir las preferencias y necesidades de los huéspedes con el objetivo de obtener mejoras en la planificación y en el desarrollo de los servicios a ofrecer.

**Solución propuesta:** consiste en la aplicación de Inteligencia de Negocios para la cadena de hoteles, desarrollando un modelo de Data Mining que permita identificar grupos de huéspedes con el fin de realizar la predicción de las necesidades y preferencias de éstos para de esta manera ofrecer un servicio personalizado en todos los hoteles de la firma.

**Supuestos e información suministrada:** Actualmente los hoteles cuentan con sistemas que permiten llevar un registro de datos sobre los servicios extra (piletas, bares, etc.) que utilizan los huéspedes como así también sobre las transacciones que realizan dentro del hotel, pero **no cuentan con un data warehouse**. Por otra parte, se propone recabar información complementaria mediante **web scraping** de páginas como Booking o TripADVISOR y de redes sociales que permitan obtener las palabras para definir la opinión de huéspedes que han completado su estadía en alguno o varios de los hoteles.

**Estado de situación:** El sistema que compila los datos almacena altas cantidades de información relevante de los huéspedes, pero no se cuenta con un módulo de Inteligencia de Negocios que sirva de apoyo para mejorar las estrategias del negocio.

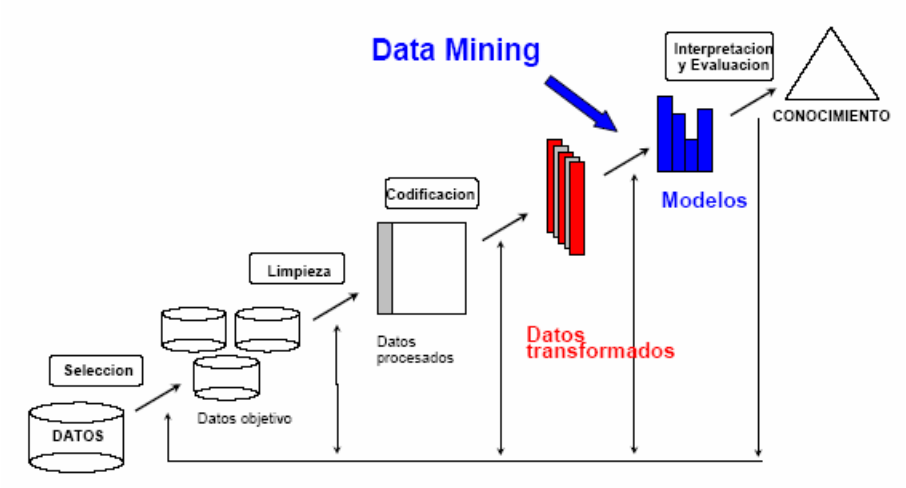
**Regulaciones y normativa:** Antes de pasar al próximo punto se considera que es importante tener en cuenta que el uso y tratamiento de todo este tipo de información está fuertemente vinculada con la privacidad de las personas. Existen diversas regulaciones en nuestro país tanto a nivel nacional como provincial que se refieren a estos aspectos. La figura de habeas data brinda la posibilidad a las personas de conocer, actualizar y rectificar los datos sobre ellas que exista en cualquier tipo de entidad, y también la de pedir la eliminación, actualización o confidencialidad de estos. Más allá de esto, en varias ocasiones no se respeta este derecho y se utiliza información sin el consentimiento de la persona, y en otras ocasiones para un propósito diferente para el cual fueron solicitados.

***KDD (knowledge discovery from data)***

El proceso KDD es un proceso iterativo, el cuál consta de los siguientes pasos como componentes del proceso:

1. Limpieza de datos.
2. Integración de los datos.
3. Selección de datos.
4. Transformación de los datos.
5. Minería de datos.

6. Evaluación de los patrones obtenidos.
7. Presentación de los resultados.




**Primera Etapa: Limpieza de los datos (Data cleaning to remove noise and inconsistent data)**

En esta etapa se evalúan la calidad de los datos de la base (relational data base + transaction data) bajo el supuesto de que no se dispone de un data warehouse. Por lo tanto, se debe acceder a la base de datos para cada uno de los hoteles de la cadena y tratar aquellos que puedan tener valores que dificulten o hagan menos confiable el Data Mining. Más específicamente, el tratamiento de valores nulos, faltantes, errores de tipeo, equivocaciones, duplicación de datos y outliers mediante la eliminación o el reemplazo con valores autogenerados mediante **ingeniería de datos** y también eliminar, comas, puntos, acentos, etc para los datos demográficos y los obtenidos mediante **web scraping.**

A modo de ejemplo, para la detección de outliers el gráfico de boxplot permite identificar dos tipos de valores extremos, los moderados y los severos en base a 1.5 y 3.0 veces la distancia intercuartil por encima o por debajo del tercer y primer cuartil respectivamente.

Se propone agregar en el cuestionario de ingreso preguntas de selección múltiple que luego puedan calificarse y sirvan de base para que el sistema las pueda utilizar como base de aprendizaje junto con el resto de la información disponible. Por ejemplo, las preguntas podrían ser sobre los servicios que le parecen más y menos interesantes que ofrece el hotel, sí desea recibir publicidad en la TV sobre las actividades del hotel con opciones de respuesta intermedias y condicionado a esta pregunta si desea recibir ofertas y actividades en su celular durante su estadía o posterior.

**Segunda Etapa: Integración de los datos (Data integration where multiple data sources may be combined)**

Luego de la limpieza de datos, se integran las fuentes de datos, es decir, la base de los hoteles (datos demográficos, datos de transacciones y datos del formulario de evaluación) y los obtenidos del web scraping en un **data warehouse** que servirá de fuente para poder realizar las consultas de información necesarias para la elaboración de los modelos.

**Tercera Etapa: Selección de los datos (Data selection where data relevant to the analysis task are retrieved from the database)**

Una vez finalizada la etapa dos se procede a seleccionar mediante criterios objetivos las variables relevantes del negocio y, en caso de variables nominales, mediante una matriz de correlaciones determinar la relación entre las mismas de forma de poder aplicar alguna técnica de reducción de datos cómo puede ser **análisis de componentes principales (PCA)**.

Del web scraping se debe contar del total de las palabras y su participación relativa, las que se encuentran en todas las opiniones, es decir las que son comunes o más frecuentes en base al criterio de preferencia de actividades ofrecidas por el hotel que los huéspedes elijen cómo prioritarias y el orden de prioridad. Esto dará una idea de lo que más destacan los huéspedes, ya sea en cuanto a lo positivo o a lo negativo. También, para aquellos huéspedes que han pasado estadías en otros hoteles de la firma y que a su vez vacacionan en verano o tienen intenciones de hacerlo.

Una herramienta de visualización interesante es la nube de palabras que básicamente representa la frecuencia de los términos más utilizados asignando un tamaño y una ubicación en base a su importancia.


Las 5 palabras que más se repiten son a simple vista y sin tratamiento son:

1. Desayuno
2. Habitaciones
3. Vista
4. Personal
5. Centro

En esta selección se puede observar que hay palabras de similar significado que deberían ser agrupadas como una sola, por ejemplo, habitación y habitaciones.

**Cuarta Etapa: Transformación de los datos (Data transformation where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)**

Dado que los datos pueden tener diferencias de escala y de medición la estandarización de los mismos (suponiendo distribución normal) o la discretización permiten expresar los datos en términos de desvíos o convertirlos de variables de continuas a discretas.

Para los datos obtenidos de la web se segmentará el texto y se eliminan aquellas palabras que no tienen un significado por sí mismas, como es el caso de preposiciones, artículos y algunos verbos no relevantes. Finalmente, a modo de simplificar el análisis, se agrupan palabras similares que tienen el mismo origen.

Luego se debe realizar la distinción entre palabras positivas y negativas mediante un diccionario que contenga un listado de términos a los que se les asigna una puntuación en una escala de valores de, por ejemplo -5 a 5. Las palabras que demuestran un sentimiento negativo son calificadas con valores de -5 a -1 y los sentimientos positivos son calificados con valores de 1 a 5. Entonces lo que determina si una frase es positiva o negativa es la suma de los valores de las palabras que se encuentran en el diccionario. En los casos en los que las opiniones no contengan ningún término del diccionario se tomarán como neutras, tomando valor 0.

**Quinta Etapa: Minería de datos (Data mining an essential process where intelligent methods are applied to extract data patterns)**

Esta etapa consiste en encontrar reglas de asociación en el comportamiento de los datos de los huéspedes, es decir, reglas que permitan encontrar asociaciones de implicancia importantes entre los servicios elegidos por los huéspedes, para saber así, qué servicios determinan la utilización de otros servicios.

Del web scraping suponemos que podemos obtener resultados sobre las palabras que más se repiten (vista, desayuno y habitaciones), la connotación positiva de cada palabra, los conceptos comunes a todos los hoteles sin importar su ubicación (bueno, comodidad, agradable). Del mismo modo, para las palabras de connotación negativa (problema, solo, dudas) los términos comunes y diferentes para todos y cada uno de los hoteles. Para el caso de la asociación de palabras mediante **bi-gramas**, lo más relevante surge de conceptos como pileta-climatizada, hermosa-vista, precio-calidad o wi-fi, esto puede mostrar una idea de aquellas características que los visitantes están buscando cuando eligen hospedarse en los hoteles de la cadena y cuáles de estas opciones priorizan durante su estadía. Esta información sirve de base para que el sistema aprenda y junto con los datos agregados a la encuesta inicial sobre preferencias poder decidir que anuncios y con qué frecuencia se les deben enviar o no a los huéspedes que inician su estadía en los hoteles.

Por otra parte, mediante las técnicas de clasificación y agrupamiento **(clustering)**, se pueden identificar los tipos de huéspedes y armar conjuntos de acuerdo a una lista de características (variables sobre consumo de servicios y estadías en hoteles de la cadena) seleccionadas con apoyo de un experto del negocio. Una vez que el experto del negocio valide los grupos identificados, se

procederá a etiquetar a los huéspedes con dichos grupos. La idea es que se agrupen solo por la semejanza de sus atributos, así se podrán obtener nuevos tipos de huéspedes y dirigir la estrategia publicitaria a un cierto grupo quienes tengan preferencias bien definidas dadas por la técnica de agrupamiento aplicada, evitando de esta forma realizar inversiones en publicidad que no sean efectivas y enviar promociones sobre los diferentes hoteles de la cadena en base al interés y los viajes y preferencias vacacionales del huésped.

El siguiente paso es aplicar una técnica de clasificación mediante **árboles** para obtener un modelo capaz de predecir qué avisos enviar durante los primeros días de estadía y que hotel sería el más adecuado en temporada de verano. Es decir, se espera que la predicción reciba como entrada un conjunto de casos en que se conoce la respuesta y entregue como salida un modelo que servirá para dar respuesta a la misma pregunta, pero con otros valores de los atributos. Por ejemplo, el management desea clasificar los huéspedes de acuerdo con las transacciones realizadas con la credencial durante su estadía, la respuesta a esta pregunta estaría dada por algunas características específicas del huésped (atributos), tales como: nacionalidad, edad, profesión, temporada de visita, etc. El modelo entregado debería servir para distinguir qué atributo es el más influyente en la respuesta obtenida, así como también, qué respuestas se obtendrían si se modifican los valores de los atributos.

**Sexta Etapa: Evaluación de los patrones obtenidos (Pattern evaluation to identify the truly interesting patterns representing knowledge based on interestingness measures)**

Las reglas de asociación tienen como objetivo encontrar relaciones o patrones interesantes entre un gran conjunto de datos. Para este caso, podría darse que el análisis de las transacciones realizadas con la credencial arrojara el resultado de que los huéspedes que solicitan habitaciones con vista al mar, tiendan a consumir preferentemente un determinado menú. Esto podría inducir al management a planear estrategias relacionadas al menú y al envío de información por la TV destinado a estos huéspedes objetivo, de manera que se ayude a maximizar el beneficio y customizar el servicio.

Además, se debe analizar con expertos los resultados o las asociaciones obtenidas por el sistema de manera de que cumplan con una lógica razonable y además evaluar la frecuencia relativa de cada patrón respecto del total de patrones observados afín de poder determinar su importancia relativa. Y aquellas asociaciones que pasen la revisión son aquellas seleccionables para enviarles a clientes antiguos promocionando los hoteles durante la temporada de verano.

**Séptima Etapa: Presentación de los resultados (Knowledge presentation where visualization and knowledge representation techniques are used to present mined knowledge to users)**

Finalmente, para la presentación de los resultados se utilizarán gráficos generados por el sistema del tipo de torta, de barras, de línea, cubos de datos multidimencionales y tablas multidimensionales, incluyendo crosstabs. La idea es que sean de fácil comprensión y que reflejen claramente los resultados obtenidos.