
# Introducción Data Mining 2022

Caso Nº1 - IDM 1

Docentes: Eduardo Poggi
Gaston Pezzuchi

Cerutti Leandro
Garcia Rio Veronica
Levit Mariana
Marchetta Mariano Daniel

# Planteo del problema:

El método de comunicación actual para eventos, actividades y promociones no es efectivo, dado que es poco personalizado y tiene poca llegada a los clientes. Se necesita una forma de personalizar los mismos de acuerdo a los distintos perfiles de clientes. Como consecuencia de esto, una vez definidos estos perfiles podremos mejorar el servicio de mailing privado para el verano.

Del presente documento, podemos recuperar la siguiente información que consideramos será relevante para el caso de estudio.

Data sources:

- Datos demográficos.
- Cuestionario de ingreso.
- Uso de servicios (tipo de servicio, fecha y hora).
- Compra de productos mediante tarjetas de crédito.
- Cuestionario completado al finalizar la estadía.

Canales de comunicación:

- TV (durante la estadía)
- Mailing (posterior a la estadía)

Objetivos y consideraciones especiales:

- El sistema deberá brindar sugerencias (de eventos, promociones, etc.) a los huéspedes durante su estadía en el hotel, esta actividad deberá tomar en consideración solamente los anuncios sobre los cuales el huésped esté posiblemente interesado. Esto también tiene que ser realizado durante los primeros días del huésped en el lugar, es decir, cuando no cuenta con información sobre servicios o productos consumidos, y una vez que su estadía haya finalizado.

- El sistema deberá seleccionar qué hoteles le pueden interesar a antiguos huéspedes, con el objetivo de enviar ofertas por mail a los mismos. Se debe tener en cuenta que la mayoría de las veces, un huésped no elige el mismo hotel en el que ya estuvo.

# Set de datos

- Datos demográficos: Supondremos que estos datos son recolectados sobre alguna plataforma web, caso contrario (es decir que son registrados manualmente), será necesario digitalizar esta información. Formato mínimo esperado:

*Dni, país, provincia/estado, calle, altura, código postal, edad, teléfono, nombre, apellido.*

- Cuestionario de ingreso: Similar al punto anterior, pero en este caso estos datos son entregados por el huésped mediante una encuesta, si la misma es en formato físico, será necesario primero digitalizar esta información. Formato mínimo esperado:

*Dni, viaja en grupo familiar (si/no), cantidad de hijos, viaja por trabajo (si/no), email.*

- Información transaccional: Dado que la misma tarjeta es utilizada para compras como ingreso a lugares, supondremos que la información se almacena en una misma base de datos. Formato mínimo esperado:

*Dni, tipo de registro (consumo/acceso), monto (solo en caso de compra), hora, permanencia en lugares.*

- Cuestionario de salida: Supondremos la misma situación que el cuestionario de ingreso. Esta información puede servir como retroalimentación, para determinar si las ofertas realizadas fueron acordes o no, nos permitirá brindar futuras sugerencias por mailing. Formato mínimo esperado:

*Dni, volvería (1-5), recomendaría (1-5), nivel de satisfacción (1-5), encontró algo interesante en nuestras ofertas (1-5), comentarios, ¿Le interesaría recibir información sobre viajes? Negocio (T/F) Familia (T/F) Amigos (T/F).*

# Pre-procesamiento

- ## Integración de datos:

    - Las bases de datos se consolidarán en una única a través del DNI. Tomaremos el cuestionario de ingreso como fuente inicial para la integración.

- ## Limpieza de datos:
    - Datos demográficos:
        - Completar/Validar código postal, teléfono.

    - Información transaccional:
        - Eliminar registros con permanencia menor a 10 minutos en lugares recreativos (tiempo validado con el gerente del hotel). Asumiendo que ese tiempo no se considera uso de la instalación.

    - Cuestionario de ingreso:
        - Valores faltantes, se harán sustituciones para tener datos válidos. Ej.: caracteres especiales, modismos de los huéspedes.

    - Cuestionario de salida:
        - Completar valores faltantes, con valores medios para ese tipo de registro.

En esta etapa se realizará un análisis exploratorio de los datos, con el objetivo evaluar distribuciones de los distintos features y encontrar valores atípicos, para darles el tratamiento adecuado.

Por otro lado, analizaremos la correlación entre distintas variables, con el objetivo de eliminar variables redundantes, reduciendo así la dimensionalidad del conjunto de datos.

- ## Selección de datos:

    - Datos demográficos:
        - Se eliminan datos irrelevantes para el modelado (Calle, altura, provincia, teléfono).
    - Información transaccional:
        - Se eliminan todos aquellos que no hayan sido descriptos en la sección Set de datos.
    - Cuestionario de ingreso:
        - Utilizaremos los comentarios para inferir (sentiment analysis) las puntuaciones numéricas en caso de que no hayan sido ingresadas.

- Cuestionario de salida:
  - Utilizaremos los comentarios para inferir (sentiment analysis) las puntuaciones numéricas en caso de que no hayan sido ingresadas.

- Transformación de datos:

  - Agregación/Enriquecimiento:
    - Completar los días de estadía en el hotel a través de los registros en la base de datos del sistema de registro del hotel.
    - Crear nuevas variables que aporten información relacionando fechas de estadía con distancia a feriados, navidad, por ejemplo (extracción de características)

  - Discretización y binarización de datos:
    - La fecha de creación del registro se llevará a valores discretos (mes y año) porque es más valioso agrupar por mes.

    - El campo tipo, que referencia a si la acción fue de acceso o consumo, lo representaremos con valores de 1 en caso de acceso y 2 en caso de consumo.

    - El campo viaja en grupo familiar (si/no) y viaja por trabajo (si/no), lo representaremos con valores de 1 en caso afirmativo y 0 en caso negativo.

# Data mining

Para resolver el primer problema sobre decidir cuáles anuncios de TV van a ser enviados a los huéspedes durante **los primeros días** de su estadía planteamos:

- ## Clusterización:

Este algoritmo encontrará grupos (clases) en el conjunto de datos históricos. Utilizaremos para esto los datos brindados por la encuesta de ingreso, los datos demográficos y datos transaccionales (históricos). También tomaremos en cuenta los Formulario de Evaluación previos. Esta etapa nos permitirá identificar las clases y clasificar a un nuevo cliente en el paso siguiente.

Se recalcularán las clases cada una temporada, con el objetivo de agrupar conjuntos o identificar nuevos.

Por ejemplo, luego del análisis podrían definirse las clases: aventura, relax, familiar, etc. Cada clase definida por el hotel va a incluir un grupo de actividades acordes a los resultados obtenidos.

Analizando grupos etarios, procedencia, tipo de viaje podemos inferir la clase a la que pertenece un nuevo cliente.

- ## Árbol de decisión:

Al ingresar un nuevo cliente, se clasificará en un clúster aplicando un árbol de decisión, el cual nos permitirá asignar a esta persona en una de las clases encontradas en la fase anterior.

- ## Caso de uso:

Al ingresar una pareja al hotel, gracias a sus datos demográficos y cuestionario completado al ingreso, se predice su clasificación como perteneciente al clúster Relax, con lo cual, en las primeras horas de su estadía el hotel va a recomendarle los servicios correspondientes a este clúster, por ejemplo, servicio del Spa, Pileta y Masajes.

Para mejorar las recomendaciones a realizar a los huéspedes **los días subsiguientes**, vamos a recurrir mayormente a los datos transaccionales.

Proponemos realizar un **análisis de asociación** en función a las compras realizadas, ingreso a los distintos servicios/zonas, así como también ciertos datos demográficos. De esta forma,

obtendremos reglas de asociación con cierto nivel de confianza que nos permita definir patrones de conducta.

Por ejemplo, si obtenemos como regla de asociación que las parejas jóvenes suelen hacer paseos de aventura (caminata, rafting) y luego hacer uso del Spa, enviaríamos un aviso a todas las parejas jóvenes que realicen paseos de aventura sobre una visita al Spa.

Para resolver el problema de la cadena de mailing, proponemos la metodología de collaborative filtering systems.

La idea es utilizar los clústeres de huéspedes que disponemos para ofrecerles hoteles bien valorados para ese perfil en concreto, es decir, tomaremos los huéspedes de cada clúster y armaremos una matriz de similitud (por clúster) para calcular las recomendaciones.

Para esto es fundamental que una vez que el huésped presentó la encuesta de salida, se valide la clasificación realizada en primera instancia teniendo en cuenta todos los datos obtenidos posteriores a la información de ingreso, pudiendo resultar en una reclasificación del huésped, con el objetivo de asignarle la clase más certera posible.

Entendiendo que podemos encontrar huéspedes que realicen viajes en diferentes roles (viaje por trabajo o por vacaciones), utilizaremos la información registrada en la encuesta de salida consultando si está interesado en recibir propuestas para sus próximos viajes, con el fin de ofrecerle opciones en esos otros perfiles correspondientes.

# Post Procesamiento

Utilizaremos técnicas de visualización y representación para exponer el conocimiento extraído al cliente. Para esto emplearíamos:

1. **Gráfico de clúster**, permitiendo la visualización de las clases generadas y su volumen.

2. **Listado reglas de asociación**, permitiendo listar ciertos insights obtenidos.

3. **Matriz de ratings**, permitiendo visualizar las puntuaciones obtenidas por cada clase en cada hotel, y también representar posibles recomendaciones a realizar.

4. **Informe ejecutivo**, detallando principales resultados obtenidos, características del análisis, conclusiones y recomendaciones.

# Bibliografía


- Data Mining: Concepts and Techniques - Jiawei Han - Micheline Kamber
- Introduction to Data Mining – Pang-Ning Tan – Michael Steinbach – Vipin Kumar
- https://www.aprendemachinelearning.com/sistemas-de-recomendacion/