Introduccion a Data Mining – 2020
Caso IDM-1

1. Definir el problema u oportunidad de negocio

El punto de partida para cualquier proceso de Data Mining es definir el problema u oportunidad de negocio (hipótesis de investigación) y los objetivos analíticos en términos conceptuales, antes de especificar cualquier dato o medida. La idea en este punto es ver el problema en términos conceptuales, definiendo los conceptos e identificando las relaciones fundamentales a investigar.

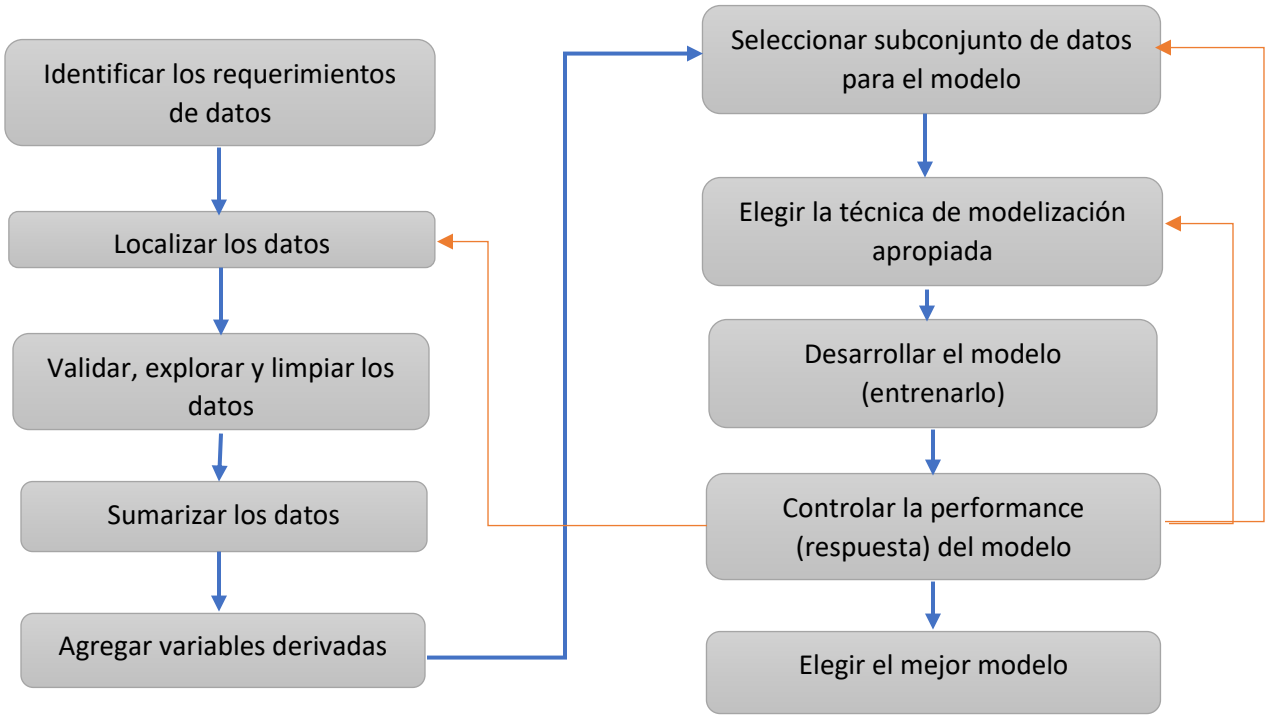
Para este caso práctico, se pretende a través de la implementación de un proyecto de data mining, extraer el máximo provecho de los datos registrados acerca de los huéspedes en los diferentes hoteles de una cadena alrededor del mundo. Más precisamente, la cadena desea conocer mejor a sus clientes de manera de poder informarlos sobre eventos especiales, promociones especiales, etc., durante su estancia como así también después de esta.
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.

2. Proceso KDD

Es en esta etapa donde el proceso de Data Mining tiene lugar en sí mismo y donde la información resultante es fundida para producir conocimiento.

TRANSFORMAR DATOS EN INFORMACIÓN



2.1 OBJETIVOS DE LA SEGMENTACIÓN DE CLIENTES

La cadena desea conocer mejor a sus clientes, este trabajo propone realizar el análisis de Perfiles de Clientes a través de un Modelo Descriptivo Multivariado orientado a la SEGMENTACIÓN DE CLIENTES, a partir de las variables que describen su comportamiento de consumo.

Conceptualmente, el objetivo fundamental del análisis cluster es la obtención de un conjunto de objetos en dos o más segmentos, basándose en su similitud para un conjunto de características especificadas. En el caso particular de este trabajo, el objetivo inicial de la segmentación es identificar aquellos clientes con gustos similares, para describir luego su perfil sociodemográfico y establecer la estrategia de comunicación de:

1) Próximos eventos, actividades y promociones especiales durante su estadía
2) Mailing personalizado luego de su estadía: publicaciones de una selección de sus hoteles a antiguos huéspedes, para obtener nuevas reservas.

2.2 SELECCIÓN DE LAS VARIABLES PARA EL ANÁLISIS CLUSTER

La selección de las variables utilizadas para caracterizar los sujetos a agrupar está directamente vinculada al objetivo del análisis cluster. Tanto en sentido confirmatorio como exploratorio, el resultado del modelo quedará restringido a las variables utilizadas en el proceso. Los segmentos derivados reflejan la estructura inherente de los datos sólo como definida por las variables seleccionadas. Se trata de un problema de Clustering o aprendizaje “no supervisado” ya que se pretende agrupar a los clientes según los indicadores de consumo que mejor los describen.

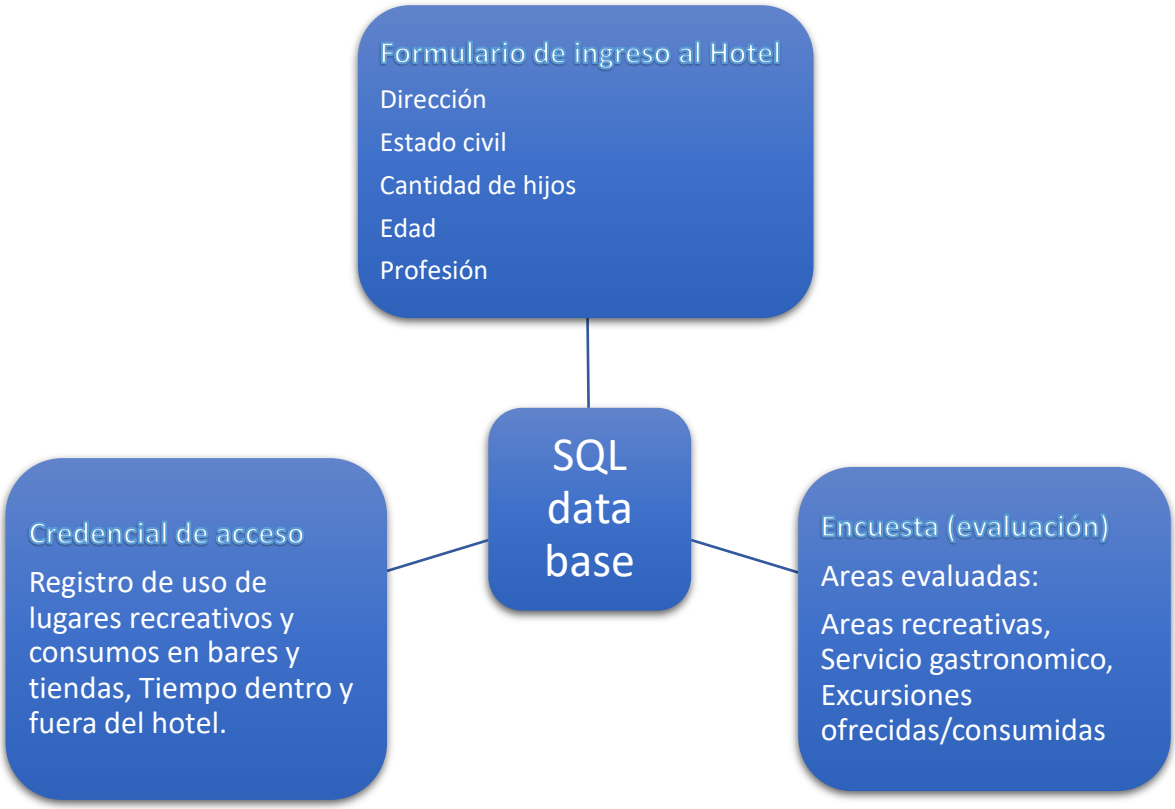
- **Uso de Piscina/Sauna**
- **Uso de Bares**
- **Uso de tiendas de compras (consumo de productos de las tiendas dentro del hotel)**
- **Entrada/Salida del hotel (Permite saber si el huésped este mas tiempo dentro del hotel usando los entretenimientos y atracciones o haciendo excursiones outdoors**
- **Datos demográficos (Lugar de residencia de los huéspedes)**
- **Consumo de bebidas (cantidad y tipo consumido)**

El ejercicio se planteará sin conocer de antemano las “reglas” de formación de los clusters, utilizando entonces una metodología “data driven” que no imponga a los datos condiciones “a priori”.

2.3 IDENTIFICACIÓN DE LAS FUENTES DE DATOS

Este proyecto cuenta con tres fuentes de datos, un formulario de ingreso, registros de uso de las instalaciones y servicios a través de la credencial de acceso, control de tiempo dentro y fuera del hotel, y por último la encuesta de satisfacción del cliente que se realiza al final de la estadía.

Se supone que esta información se consolidará en una base de datos relacional.



2.4 LIMPIEZA Y TRANSFORMACIÓN DE LOS DATOS

La primera etapa de preparación y limpieza de los datos insumirá una porción importante de recursos humanos y tiempo, pero existe plena conciencia de que los resultados del proceso dependen de la calidad de los datos, por lo que resulta necesario asegurar su consistencia antes de la aplicación del algoritmo de clustering.

En la primera etapa se estima que se llevarán a cabo las siguientes tareas de Data Cleaning:

- Agrupar transacciones de la misma cuenta que tenían el mismo participante -día-hora.
- Filtrar datos/transacciones inválidas
- Filtrar transacciones con fechas incorrectas.

2.5 SUMARIZACIÓN DE LOS DATOS Y AGREGADO DE VARIABLES DERIVADAS

En este caso se realizarán las siguientes sumarizaciones de datos por cliente:

- Consumo Bar
- Consumo Tiendas
- Consumo en excursiones
- Días de hospedaje en el año
- Cantidad de hoteles visitados en el año

Variables derivadas:

- Ratio uso piscina: Cantidad de días que se utilizó la piscina/cantidad de días hospedado
- Ratio uso Bar: Cantidad de días que se utilizó el bar/cantidad de días hospedado
- Ratio uso sauna: Cantidad de días que se utilizó sauna/cantidad de días hospedado
- Activación temprana de TV: uso de áreas recreativas y consumos en las primeras 24hs de estadía

Finalmente, en la siguiente tabla se observa el formato del data frame de datos que se propone para ser sometido al clustering.

Cliente	Ratio Piscina	Ratio Sauna	Ratio Bar	Activación TV	Consumo Bar	Consumo Tiendas	Consumo Excursiones
0001	0.5	0.2	0.7	0.6	1000	1500	NA
0002	0.7	0.9	0.5	0.8	569	1000	5000
0003	0.1	NA	NA	0.03	120	NA	NA

Nota: La variable Activación TV pretende medir la cantidad de tiempo que el TV esta encendido durante la estadía y a que horas.

- DETECCIÓN DE DATOS AUSENTES Y CASOS ATÍPICOS

La exploración univariada inicial de los datos permitirá la identificación de Datos Ausentes (NA) y Casos Atípicos (outliers) para los que se adoptará el siguiente criterio:

- Datos Ausentes (NA): Se estima que los datos ausentes están asociados al no uso o consumo de cierto servicio. Por lo que se puede convertir a cero el valor
- Casos Atípicos (Outliers): Los “casos atípicos” son observaciones de una combinación única de características identificables que les diferencia claramente de las otras observaciones. Los casos atípicos deben ser analizados en el contexto del estudio y evaluados en función de la información que proporcionan, antes de caracterizarlos como benéficos o problemáticos.

- DISCRETIZACIÓN DE LAS VARIABLES

Para poder definir tipos de clientes, se discretizarán las variables de la siguiente forma:

Variable / Valor Discreto	**1**	**2**	**3**	**4**
Ratio Piscina **(RP)**	(>=0.75)	(>=0.5, <0.75)	(>=0.25, <0.5)	(>=0, <0.25)
Ratio Sauna **(RS)**	(>=0.75)	(>=0.5, <0.75)	(>=0.25, <0.5)	(>=0, <0.25)
Ratio Bar **(RB)**	(>=0.75)	(>=0.5, <0.75)	(>=0.25, <0.5)	(>=0, <0.25)
Consumo Bar **(CB)**	>=1000	(>=750, <1000)	(>=500, <750)	(>=0, <500)
Consumo Tiendas **(CT)**	>=3000	(>=2000, <3000)	(>=1000, <2000)	(>=0, <1000)
Consumo Excursiones **(CE)**	<= 1	2	3	>3

A partir de estas transformaciones se obtiene entonces un conjunto de indicadores ordinales para cada tipo de cliente.

Tipo de Cliente	**Categoría ordinal**
Luxury	RP2, RS1, RB1, CB1, CT1, CE3
Family	RP1, RS4, RB3, CB3, CT2, CE2
Business	RP4, RS3, RB4, CB4, CT3, CE1
Green Traveler	RP3, RS2, RB2, CB2, CT4, CE4

2.6 DISEÑO DE LA INVESTIGACIÓN MEDIANTE EL ANÁLISIS CLUSTER

El análisis cluster agrupa a los clientes y a las variables en segmentos, de tal forma que los clientes del mismo segmento son más parecidos entre sí que a los clientes de otro segmento. El análisis cluster intenta maximizar la homogeneidad de los sujetos dentro de los segmentos. El criterio esencial de todos los algoritmos K-means es maximizar las diferencias entre los segmentos, relativa a la variación dentro de los mismos. Este tipo de algoritmos asigna los objetos una vez que el número de segmentos está especificado. Inicialmente se plantea la posibilidad de trabajar con 4 segmentos.

La interpretación de los de los resultados permitirá definir los tipos de clientes y las estrategias de marketing que son objeto de este estudio.

Suponiendo que los resultados nos permiten definir 4 tipos de clientes, entonces las estrategias de marketing podrían plantearse de la siguiente forma:

Tipo de Cliente	Anuncios de TV		Mailing privado
	Primeras 48hs Decisiones basadas en datos de estadías anteriores. La variable Activación TV permitirá definir si es viable esta técnica de marketing	**Estadía completa** Complementar datos anteriores con los datos recolectados en las primeras 48hs de estadía.	**Áreas de enfoque**
Luxury	Compras en tiendas	Actividades en zonas recreativas	Tiendas de interés, promociones de bebidas y actividades en las instalaciones del hotel
Family	Actividades para los niños dentro del hotel	Promocionar otros hoteles de la cadena en relación a actividades familiares	Actividades para niños en la piscina, promociones en bar y excursiones.
Business	Permanece más tiempo dentro de la habitación, pero no se recomienda bombardeo publicitario de servicios de ocio	Propaganda para próximas vacaciones familiares (en base de datos de la ficha de entrada)	Promociones sauna y tiendas, hoteles en zonas altamente urbanizadas, con espacios comunes para reuniones empresariales.
Green Traveler	Excursiones promocionales	Excursiones y promociones gastronomicas en el bar del hotel	Excursiones outdoors y Bar, hotels en zonas menos urbanizadas