CASO 1 - HOTELES

Introducción a Data Mining – 2022 virtual

Grupo 11

Horacio Basilio
Micaela Bassan
Fernando Bogorni

Alumnos: Bassan M.,
Basilio H.
Bogorni F.
Cohorte: 2022-2023 Virtual

01.-Introducción 2
02.- Objetivos 2
03.- Metodología 3
1) Limpieza de datos: 3
2) Integración de datos: 4
3) Selección de datos 4
4) Transformación de datos 4
5) Procesamiento de datos: 4
6) Evaluación de patrones 9
7) Presentación del conocimiento 9
04.- Supuestos y Premisas 9
05.-Aplicación de solución 10
Resolución objetivo 1 - Publicidad TV 10
Resolución objetivo 2 – Campaña de Mailing 13
07.-Conclusiones 14

Alumnos: Bassan M.,
Basilio H.
Bogorni F.
Cohorte: 2022-2023 Virtual

01.-Introducción

En un mundo que es rico en datos, pero pobre en información es fundamental que las empresas adquieran una mejor comprensión del contexto comercial de su organización, como sus clientes, el mercado, el suministro, los recursos y los competidores.

A través de la minería de datos podemos llegar a convertir una gran cantidad de datos en conocimiento.

En el presente trabajo buscamos resaltar la importancia de la minería de datos para una cadena hotelera y haciendo uso de la metodología KDD lograr responder las preguntas de negocio que apunten a:

- Mejorar la eficiencia en la toma de decisiones: analizar grandes volúmenes de datos y obtener información relevante que nos ayude a tomar decisiones más acertadas y eficientes en cuanto a la gestión de los hoteles.
- Aumentar la rentabilidad: identificar tendencias y patrones en el comportamiento de los clientes, lo que nos permitiría ofrecer servicios y promociones que se ajusten a sus preferencias y necesidades, lo que a su vez puede aumentar la rentabilidad.
- Mejorar la experiencia del cliente: conocer mejor a los clientes y ofrecerles una experiencia personalizada y adaptada a sus preferencias, lo que puede aumentar su satisfacción y fidelización.
- Fortalecer la competitividad: obtener información valiosa sobre el mercado y la competencia, y esto nos permite adaptarnos a las necesidades y preferencias del mercado y mantenernos en un entorno cada vez más competitivo.

02.- Objetivos

Según los requerimientos de la cadena hotelera, se realizará un proyecto de Data Mining que permita conocer a sus clientes para poder sugerir mediante publicidades direccionadas y correo personalizados diferentes ofertas, varias orientadas a que obtengan la mejor estadía y divertimentos posibles que **fidelicen** su relación con la marca de la empresa que posee diversos establecimientos en varios países. Si bien se sabe que un pasajero no tiende a elegir el mismo hotel, se espera afianzar la marca en sus conciencias como símbolo de calidad, excelente estadía y atención personalizada.

Toda la información acerca de sus huéspedes en los diferentes hoteles es recolectada a la llegada de estos. Durante el check-in 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, gimnasios, casinos, restaurant, salones, 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 y servicios en las 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.

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. En resumen, nuestro modelo de análisis de los datos debe poder predecir y cumplir con los siguientes ítems:

A. 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.

B. Decidir sobre la selección de hoteles para el mailing privado durante el verano.

03.- Metodología

Como primera medida, se selecciona una metodología para realizar Análisis inteligente de Datos, conocida como el proceso KDD (En español, descubrimiento de conocimiento en bases de datos). Este proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. Esta tarea implica generalmente preprocesar los datos, hacer minería de datos (*data mining*) y presentar resultados.

Para esta metodología, se realizan los siguientes pasos:

1) **Limpieza de datos:** la realizamos para eliminar ruido y datos inconsistentes. En la etapa de preprocesamiento/limpieza (*data cleaning*) se analiza la calidad de los datos, se aplican operaciones básicas como la remoción de datos ruidosos, se seleccionan estrategias para el manejo de datos desconocidos (*missing* y *empty*), datos nulos, datos duplicados y técnicas estadísticas para su reemplazo.

   En esta etapa, es de suma importancia la interacción con el usuario o analista. Los datos ruidosos (*noisy data*) son valores que están significativamente fuera del rango de valores esperados; se deben principalmente a errores humanos, a cambios en el sistema, a información no disponible a tiempo y a fuentes heterogéneas de datos. Los datos desconocidos *empty* son aquellos a los cuales no les corresponde un valor en el mundo real y los *missing* son aquellos que tienen un valor que no fue capturado. Los datos nulos son datos desconocidos que son permitidos por los sistemas gestores de bases de datos relacionales (sgbdr). En el proceso de limpieza todos estos valores se ignoran, se reemplazan por un valor por omisión, o por el valor más cercano, es decir, se usan métricas de tipo estadístico como media, moda, mínimo y máximo para reemplazarlos.

2) **Integración de datos:** Donde podemos tomar datos de diferentes fuentes y unificar. En este caso, por ejemplo, tomamos la encuesta completada en el momento del check in, luego los registros obtenidos de la tarjeta para el uso de amenities y consumos y a todo esto le sumamos el cuestionario final de nivel de servicio para tener todo unificado en el datawarehouse.

3) **Selección de datos:** En la etapa de selección, una vez identificado el conocimiento relevante y prioritario y definidas las metas del proceso KDD, desde el punto de vista del usuario final, se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos o una muestra representativa de este, sobre el cual se realiza el proceso de descubrimiento. La selección de los datos varía de acuerdo con los objetivos del negocio.

4) **Transformación de datos:** En este punto se transforman los datos y se consolidan en formas apropiadas para la minería de datos, por ejemplo, realizando operaciones resumen o agregación.

5) **Procesamiento de datos:** Aplicamos métodos inteligentes para extraer patrones de datos: como clustering, arboles de decisión, o técnicas para identificar patrones de comportamiento. Las técnicas de minería de datos crean modelos que son predictivos o descriptivos. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo, dependientes o clases, usando otras variables denominadas independientes o predictivas, como por ejemplo predecir para nuevos clientes si son buenos o malos basados en su estado civil, edad, género y profesión, o determinar para nuevos estudiantes si desertan o no en función de su zona de procedencia, facultad, estrato, género, edad y promedio de notas. Entre las tareas predictivas están la clasificación y la regresión. Los modelos descriptivos identifican patrones que explican o resumen los datos; sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos, como identificar grupos de personas con gustos similares o identificar patrones de compra de clientes en una determinada zona de la ciudad. Entre las tareas descriptivas se cuentan las reglas de asociación, los patrones secuenciales, los *clustering* y las correlaciones.
   Por lo tanto, la escogencia de un algoritmo de minería de datos incluye la selección de los métodos por aplicar en la búsqueda de patrones en los datos, así como la decisión sobre los modelos y los parámetros más apropiados, dependiendo del tipo de datos (categóricos, numéricos) por utilizar.

Dentro de la minería de datos se encuentran diferentes tipos de tareas, las cuales pueden considerarse como un tipo de problema para ser resuelto por un algoritmo de minería de datos. Entre las tareas de minería de datos más importantes están la clasificación, segmentación o *clustering*, asociación y patrones secuenciales.

Clasificación

La clasificación de datos permite obtener resultados a partir de un proceso de aprendizaje supervisado. Es, además, el proceso por medio del cual se encuentran propiedades comunes entre un conjunto de objetos de una base de datos y se los cataloga en diferentes clases, de acuerdo con el modelo de clasificación.

Este proceso se realiza en dos pasos: en el primero se construye un modelo, en el cual cada tupla de un conjunto de tuplas de la base de datos tiene una clase conocida (etiqueta), determinada por

uno de los atributos de la base de datos llamado *atributo clase*. El conjunto de tuplas que sirve para construir el modelo se denomina *conjunto de entrenamiento* y se escoge de manera random del total de tuplas de la base de datos. A cada tupla de este conjunto se denomina *ejemplo de entrenamiento.* En el segundo paso se usa el modelo para clasificar. Inicialmente, se estima la exactitud del modelo utilizando otro conjunto de tuplas de la base de datos, cuya clase es conocida, denominado *conjunto de prueba*. Este conjunto es escogido al azar y es independiente del conjunto de entrenamiento. A cada tupla de este conjunto se denomina *ejemplo de prueba*.

La exactitud del modelo, sobre el conjunto de prueba, es el porcentaje de ejemplos de prueba que son correctamente clasificadas por el modelo. Si la exactitud del modelo se considera aceptable, se puede usar para clasificar futuros datos o tuplas para los cuales no se conoce la clase a la que pertenecen. Se han propuesto varios métodos de clasificación: árboles de decisión, redes neuronales, Bayes, algoritmos genéticos entre otros.

El modelo de clasificación basado en árboles de decisión es un método de aprendizaje supervisado que construye árboles de decisión a partir de un conjunto de casos o ejemplos denominados *conjunto de entrenamiento* (*training set*) extraídos de la base de datos. También se escoge un conjunto de prueba, cuyas características son conocidas, con el fin de evaluar el árbol.

La calidad del árbol depende de la precisión de la clasificación y del tamaño del árbol. El método primero escoge un subconjunto del conjunto de entrenamiento y forma un árbol de decisión. Si el árbol no da la respuesta correcta para todos los objetos del conjunto prueba, se utiliza el proceso de descubrimiento de conocimiento en bases de datos de excepciones donde se adiciona al conjunto de entrenamiento y el proceso continúa hasta que se encuentra el conjunto de decisiones correctas. El eventual resultado es un árbol en el cual cada hoja lleva un nombre de la clase y cada nodo interior especifica un atributo con una rama correspondiente a cada posible valor del atributo.

Entre los algoritmos de clasificación para árboles de decisión se cuentan id-3 (Quinlan, 1986), c4.5 (Quinlan, 1993), Sprint (Shafer, Agrawal y Metha, 1996), sliq (Metha, Agrawal y Rissanen, 1996) y j48 (Hall, Frank y Witten, 2011). La idea básica de estos algoritmos es construir los árboles de decisión en los que:

- Cada nodo no terminal está etiquetado con un atributo.
- Cada rama que sale de un nodo está etiquetada con un valor de ese atributo.
- Cada nodo terminal está etiquetado con un conjunto de casos, los cuales satisfacen todos los valores de atributos que etiquetan el camino desde ese nodo al nodo inicial.

La aplicación de un atributo *A* como criterio de selección clasifica los casos en distintos conjuntos (tantos como valores discretos del atributo). Se trata de construir el árbol de decisión más simple que sea consistente con el conjunto de entrenamiento T. Para ello hay que ordenar los atributos relevantes, desde la raíz a los nodos terminales, de mayor a menor poder de clasificación. El poder de clasificación de un atributo *A* es su capacidad para generar particiones del conjunto de entrenamiento que se ajuste en un grado dado a las distintas clases posibles; de esta forma se introduce un orden en dicho conjunto. El orden o el desorden (ruido) de un conjunto de datos son medibles. El poder de clasificación de un atributo se mide de acuerdo con su capacidad para reducir la incertidumbre o entropía (grado de desorden de un sistema). Esta métrica se denomina *ganancia de información*. El atributo con la más alta ganancia de información se escoge como el atributo que forme un nodo en el árbol.

El árbol de decisión se construye de la siguiente forma:

• Calcular la entropía que puede reducir cada atributo.
• Ordenar los atributos de mayor a menor capacidad de reducción de entropía.
• Construir el árbol de decisión siguiendo la lista ordenada de atributos.

La ganancia de información obtenida por el particionamiento del conjunto T, de acuerdo con el atributo *A* se define como: $Gain(T,A)=I(T)-E(A)$

Donde es la entropía del conjunto *T*, compuesto de *s* ejemplos y *m* distintas clases *Ci (i=1,m)* y se calcula: $\mathrm{Obj}(T) = - p * log_2 (p_i)$

Donde, $p_i = s_i/s$ es la probabilidad que un ejemplo cualquiera pertenezca a una clase C_i y si es el número de ejemplos de *T* de la clase C_i.

E(A) es la entropía del conjunto *T* si es particionado por los *n* diferentes valores del atributo *A* en *n* subconjuntos, $\{S_1, S_2, ... S_n\}$, donde S_j contiene esos ejemplos de *T* que tienen el valor a_j en *A* y s_{ij} el número de ejemplos de la clase *Ci* en el subconjunto *Sj*.

E(A) se calcula:

$E(A)=s_{ij}/s * I(S_{ij})$
Donde, s_{ij} el número de ejemplos de la clase *Ci* en el subconjunto *Sj*

$I(S_{ij})= - p_{ij} log^2(p^{ij})$

Donde $p_{ij}=s_{ij} / |s^j|$ es la probabilidad de que un ejemplo de *Sj* pertenezca a la clase *Ci*.

En otras palabras, *Gain (T, A)* es la reducción esperada de la entropía causada por el particionamiento de *T* de acuerdo con el atributo *A.*

Finalmente, las reglas de clasificación se obtienen recorriendo cada rama del árbol desde la raíz hasta el nodo terminal. El antecedente de la regla es la conjunción de los pares recogidos en cada nodo y el consecuente es el nodo terminal.

Segmentación o clustering

El proceso de agrupar objetos físicos o abstractos en clases de objetos similares se llama segmentación o *clustering* o clasificación no supervisada. Básicamente, el *clustering* agrupa un conjunto de datos (sin un atributo de clase predefinido) basado en el principio de maximizar la similitud intraclase y minimizar la similitud interclase. El análisis de *clustering* ayuda a construir particiones significativas de un gran conjunto de objetos basado en la metodología divide y conquista, la cual descompone un sistema de gran escala en pequeños componentes para simplificar el diseño y la implementación.

La meta del *clustering* en una base de datos es la partición de esta en segmentos o *clusters* de registros similares que comparten un número de propiedades y son considerados homogéneos. Los registros en diversos *clusters* son diferentes y estos últimos tienen una alta homogeneidad interna (dentro del *clúster*) y una alta heterogeneidad externa (entre *clusters*). Por homogeneidad se

entiende que los registros en un *clúster* están próximos unos a otros; allí la proximidad se expresa por medio de una medida, dependiendo de la distancia de los registros al centro del segmento. Por heterogeneidad se entiende que los registros en diferentes segmentos no son similares de acuerdo con una medida de similaridad.

La segmentación, típicamente, permite descubrir subpoblaciones homogéneas: por ejemplo, se aplica a una base de datos de clientes, para mejorar la exactitud de los perfiles, identificando subgrupos de clientes que tienen un comportamiento similar al comprar. En nuestro caso para agrupar a los hosts en grupos en los cuales podríamos enviar mails, según sus perfiles y preferencias.

El algoritmo de *clustering* segmenta una base de datos sin ninguna indicación por parte del usuario sobre el tipo de *clusters* que va a encontrar en la base de datos, y desecha cualquier sesgo o intuición por parte del usuario; así́ potencia el verdadero descubrimiento de conocimiento. Por esta razón, al método de segmentación o *clustering* se lo denomina *aprendizaje no supervisado*. Algunos de los algoritmos utilizados para *clustering* son: K-Means (Han y Kamber, 2001), Clarans (*Clustering Large Applications based upon Randomized Search*) (Ng y Han, 1994), y Birch (*Balanced Iterative Reducing and Clustering using Hierarchies*) (Zhang, Ramakrishnan y Livny, 1996).

El *clustering* se utiliza por ejemplo en el análisis de flujo de efectivo para un grupo de clientes que paga en un periodo del mes en particular, para hacer segmentación de mercado y para descubrir grupos de afinidades. También se utiliza para descubrir subpoblaciones homogéneas de consumidores en bases de datos de *marketing*.

Asociación

La tarea de asociación descubre patrones en forma de reglas, que muestran los hechos que ocurren frecuentemente juntos en un conjunto de datos determinado. El problema fue formulado por Agrawal et al. (1992), y a menudo se referencia como el problema de canasta de mercado (*market-basket*). En este problema se da un conjunto de ítems y una colección de transacciones que son subconjuntos (canastas) de estos ítems. La tarea es encontrar relaciones entre los ítems de esas canastas para descubrir reglas de asociación que cumplan unas especificaciones mínimas dadas por el usuario, expresadas en forma de soporte y confianza. Las cantidades de ítems comprados en una transacción no se toman en cuenta, lo que significa que cada ítem es una variable binaria que representa si un ítem está presente o no en una transacción.

Formalmente, sea I={i1, i2, ...,im} un conjunto de literales, llamados ítems; sea D un conjunto de transacciones, donde cada transacción T es un conjunto de ítems tal que T I. Cada transacción se asocia con un identificador llamado tid.

El significado intuitivo de una regla de asociación es que las transacciones de la base de datos que contienen X tienden a contener Y. La regla X Y se cumple en el conjunto de transacciones D con una confianza c si el c% de las transacciones en D que contienen X también contienen Y. La regla X Y tiene un soporte *s* en el conjunto de transacciones D si el *s*% de las transacciones en D contienen X Y.

La confianza denota la fuerza de la implicación y el soporte indica la frecuencia de ocurrencia de los patrones en la regla. Las reglas con una confianza alta y soporte fuerte son referidas como reglas fuertes (*strong rules*). El problema de encontrar reglas de asociación se descompone en los siguientes pasos:

• Descubrir los *itemsets* frecuentes, i.e., el conjunto de ítems que tienen el soporte de transacciones por encima de un predeterminado soporte *s* mínimo.

• Usar los *itemsets* frecuentes para generar las reglas de asociación para la base de datos.

Después de que los *itemsets* frecuentes son identificados, las correspondientes reglas de asociación se pueden derivar de una manera directa. Un ejemplo de una regla de asociación es "el 30% de las transacciones que contienen cerveza también contienen pañales; el 2% de todas las transacciones contienen a ambos ítems". Aquí el 30% es la confianza de la regla y el 2%, el soporte de la regla.

Según Han y Kamber (2001), existen varios criterios para clasificar las reglas de asociación, uno de estos es el de las dimensiones que estas abarcan. De acuerdo con este criterio, las reglas de asociación pueden ser unidimensionales y multidimensionales. Una regla de asociación es unidimensional, si los ítems o atributos de la regla hacen referencia a un solo predicado o dimensión. Por ejemplo, se tiene la siguiente regla de asociación:

Cerveza ^ papas fritas => pañales, que se puede reescribir como: Compra (cerveza) ^ compra (papas fritas) => compra (pañales), hace referencia a una sola dimensión: compra.

Una regla de asociación es multidimensional, si los ítems o atributos de la regla hacen referencia a dos o más criterios o dimensiones. Por ejemplo, está la siguiente regla de asociación:

Edad (30...39) ^ ocupación (ingeniero) =>compra (laptop), contiene tres predicados*: edad, ocupación y compra.*

Un uso clásico de asociaciones es el análisis de la canasta de mercado, en la cual la asociación es una lista de afinidades de productos. Por ejemplo, observar los pedidos individuales de clientes para suministros de oficina puede generar una regla: el 70% de los clientes que ordenan plumas y lápices también ordenan libretas.

Otras aplicaciones de reglas de asociación son los análisis de demandas médicas para determinar procedimientos médicos que se realizan al mismo tiempo o a lo largo de un periodo, para un diagnóstico en particular. También se aplican para el análisis de textos, diseño de catálogos, segmentación de clientes basado en patrones de compra, en mercadeo, entre otros.

Patrones secuenciales

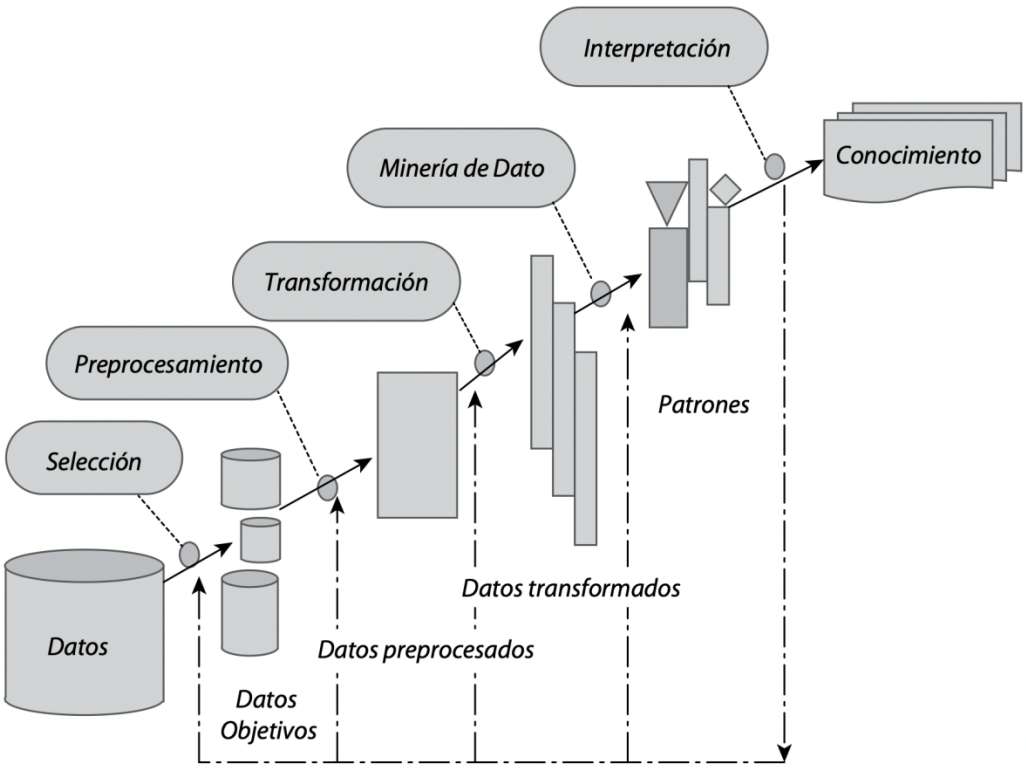
Los patrones secuenciales buscan ocurrencias cronológicas. Se aplica principalmente en el análisis de la canasta de mercado y su objetivo es descubrir en los clientes ciertos comportamientos de compra en el tiempo. El dato de entrada es un conjunto de secuencias llamado *data-secuencia*. Cada una de estas últimas es una lista de transacciones, en las que cada transacción es un conjunto de ítems (literales). Típicamente, hay un tiempo asociado con cada transacción.

Un patrón secuencial también se compone de una lista de conjuntos de ítems. El problema es encontrar todos los patrones secuenciales que cumplan con un soporte mínimo especificado por el usuario, en el cual el soporte es el porcentaje de *data-secuencias* que contiene el patrón. Por ejemplo, en una base de datos de una librería, cada *data-secuencia* puede corresponder a todas las selecciones de libros de un cliente y cada transacción, a los libros seleccionados por el cliente en una orden.

Un patrón secuencial puede ser "El 50% de los clientes que van al gimnasio también utilizan la pileta del hotel". La *data-secuencia* correspondiente al cliente, quien utilizó otros servicios conjuntamente o después utilizó otros, contiene este patrón secuencial. La *data-secuencia* puede también tener otros eventos en la misma transacción, así como uno de los tipos del patrón. Elementos de un patrón secuencial pueden ser conjuntos de ítems; por ejemplo, "'*Foundation*' y '*Ringworld*', seguido por '*Foundation* y *Empire*' y '*Ringworld Engineers*'". Sin embargo, todos los ítems en un elemento de un patrón secuencial deben estar presentes en una transacción simple para que la *data-secuencia* soporte al patrón.

6) **Evaluación de patrones:** En la etapa de interpretación/evaluación, se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones. Esta etapa puede incluir la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de los patrones útiles en términos que sean entendibles para el usuario. Por otra parte, se consolida el conocimiento descubierto para incorporarlo en otro sistema para posteriores acciones o, simple- mente, para documentarlo y reportarlo a las partes interesadas; también para verificar y resolver conflictos potenciales con el conocimiento previamente descubierto.

7) **Presentación del conocimiento:** Utilizamos herramientas de visualización (ejemplo Power BI o Tableau) para representar el conocimiento extraído a los usuarios.



04.- Supuestos y Premisas

Para el diseño y desarrollo de la solución se establecen las siguientes premisas y supuestos que permitirán encuadrar las técnicas y teorías aplicadas.

Se presupone que el hotel forma parte de un grupo hotelero internacional con presencia en las principales ciudades y destinos turísticos del mundo.

Se supone que la cadena de hoteles cuenta con software de Data Warehouse o un DataLake en la nube para procesar toda la información que se recabe de los clientes no solo en el propio hotel sino en toda la cadena a la que este pertenece.

Si bien el registro del check-in y check-out, se asienta en el sistema contable y de facturación del hotel, el uso de las tarjetas de acceso a la habitación y a todas las áreas de servicios y comodidades del hotel (amenities) como gimnasio y/o canchas de deportes, sauna, spa, piscina, salón comedor o desayunador, bar, estacionamiento, solárium, área de juegos, casino y todo extra con el que cuente este y otro hotel de la cadena, también se registran en el sistema de almacenamiento de eventos transaccionales de los hoteles que van a dicho Data Lake.

Todos los hoteles de la cadena tienen el mismo sistema de almacenamiento de la información local que es sumarizada o resumida y enviada a al sistema de almacenamiento central en la nube.

El sistema en la nube es consultado por el software programador de publicidades del canal privado de cada hotel y este el que tiene la posibilidad de armar el conjunto de publicidades que se van a proveer a cada dispositivo de televisión en las habitaciones de los clientes y las zonas comunes.

Al finalizar la estadía se invita al cliente a completar una pequeña encuesta en un monolito de informes computarizado que existe en cada recepción y que puede ofrecer un regalo de despedida para motivar a los pasajeros a completar la misma y contar con información veraz y fehaciente, motivada por el regalo. La encuesta es simple y las preguntas se contestan en el rango del 1 al 5 con la posibilidad de incluir algunas palabras preestablecidas que permita hacer un análisis lingüístico y una final libre para que el cliente pueda expresarse libremente.

05.-Aplicación de solución

Resolución objetivo 1 - Publicidad TV

A.- Decidir cuáles anuncios de TV van a ser enviados a los huéspedes durante su estadía; un subproblema de este primer problema es decidir cuáles anuncios van a ser enviados durante los primeros días de la estadía de un huésped.

1. Toma inicial de datos del cliente:

Al arribo del cliente al hotel se realiza el proceso de check in del mismo para la toma inicial de sus datos que, en caso de ya estar registrados en el sistema por viajes anteriores, actualizará los mismo para su utilización durante esta nueva estadía. Esta toma inicial de datos lanza el proceso de data mining para el ofrecimiento de publicidades en el sistema cerrado de televisión del hotel con las metodologías que se aplican a continuación.

2. Clasificar a los pasajeros del hotel:

Para poder direccionar la publicidad en el canal personalizado del hotel es necesario clasificar diferentes tipos de interés que sirvan para segmentar a los clientes y en diferentes categorías donde las publicidades ofrezcan propuestas acordes a los mismos. El mayor inconveniente que se enfrenta es que existe una gran variedad de ofertas y una gran cantidad de perfiles diferentes de pasajeros.

Se decide definir para resolver este inconveniente un método de clasificación por etiquetas o "tags" que, en su combinación de acuerdo a los datos colectados de los mismos en diferentes sectores de interés, registro de check in, historial de anteriores alojamientos y encuestas o comportamiento en otros hoteles de la cadena permitan identificar intereses anteriores y actuales de los mismos.

Si bien este método es para responder a un requerimiento específico dentro de cada hotel, en la asignación de publicidades para el sistema privado de TV, se contempla poder reutilizar esta clasificación o metodología en toda la cadena para el envío del mailing de sugerencias veraniegas de alojamiento, por lo cual los grupos definidos en principio son 3 pero podrán ir aumentando o variando dependiendo de las ofertas que tengan el propio hotel, donde el mismo se implemente o las posibilidades turísticas que tenga la ciudad o destino turístico donde se produzca el alojamiento del cliente.

Los grupos o etiquetas seleccionados y definidos son:

1.- Clase Social o S**egmento**: High Class – Premium – Turista – Business

2.- Grupo Etario o **Edad**: Mayor- Adulto – Joven – Niño

3.- Declaración definida del viaje o **motivación**: Aventura – Descanso – Trabajo – Familiar

Las asignaciones de la etiqueta de "**segmento"** se realiza mediante la información de checkin del pasajero y confirmada en el sistema por el recepcionista según forma de pago acordada (efectivo, tipo de tarjeta de crédito, canje, etc.), historial crediticio e historia de alojamientos anteriores.

La **edad** se divide según la fecha de nacimiento registrada entre: Mayor si > 60, Adulto entre 30 y 60, Joven entre 18 y 29 y Niño los menores de 18 años.

Por último, la etiqueta de **motivación** del viaje se toma de la declaración realizada por el propio pasajero para esta oportunidad de alojamiento.

3. Identificar patrones frecuentes de comportamiento histórico y actual para armado del perfil de clientes.

Si tomamos la información de huéspedes anteriores podríamos analizar patrones de comportamientos secuenciales o reglas de asociación, para por ejemplo decir, si el huésped viene por trabajo el 70% de las veces utilizará el gimnasio y el 90% de las veces cenará en nuestro restaurante. Y así, poder enviar descuentos de los amenities que usualmente usa y porque no relacionarlos. Por ejemplo: Si el viaje es familiar, el 90% de los huéspedes desayuna y luego va a la pileta o a actividades recreativas.

También podríamos utilizar clientes que suelen ir en determinado mes o temporada para enviarles descuentos de alojamiento y volver a atraerlos a nuestros hoteles.

4. Configurar un set de publicidades para diferentes perfiles de clientes predefinidos.

Para la configuración de un set de publicidades necesitamos asegurarnos ciertos pasos:

a. En base a la clasificación de clientes definida en el punto 1, las publicidades también se clasificarán por los tres atributos básicos, pero pudiendo pertenecer a más de uno a la vez, lo que hace que puedan ser seleccionadas de la base de publicidades si cumplen con estas condiciones y filtradas en modo random cambiando la lista de publicidades cada 30 minutos. También tendrán un atributo de amenities si es una publicidad relacionada directamente con algún servicio que brinda el hotel directamente.
b. Una vez identificados los diferentes clientes, debemos investigar sus necesidades y preferencias. Esto nos permitirá crear anuncios que sean relevantes y atractivos para cada grupo de clientes.
   De esta manera las publicidades que tengan relación con las preferencias de uso del hotel por parte del pasajero serán ponderadas con la sumarización de las preferencias del uso del hotel por parte de estos, para que tengan prioridad de visualización en la lista. Por ejemplo, si el cliente hace uso del bar varias veces al día, pero más que el gimnasio, se pasará primero la publicidad del programa de música en vivo del hotel en el bar antes que el cronograma de clases diversas en el gimnasio o la piscina.
c. Luego, debemos diseñar y crear los anuncios publicitarios de manera que sean efectivos para cada perfil de cliente. Esto puede incluir el uso de imágenes, videos y mensajes específicos para cada grupo de clientes.
d. Una vez que tengamos los anuncios listos estamos en condiciones de utilizar el canal elegido para llegar a los clientes, en este caso por medio del canal de TV privado y las campañas de mailing.
e. Finalmente, monitorear y medir el rendimiento de los anuncios publicitarios para asegurarnos de que están siendo efectivos en atraer a los diferentes perfiles de clientes. Esto nos permitirá hacer ajustes y mejorar los anuncios en el futuro.

En resumen, configurar un conjunto de anuncios publicitarios para diferentes perfiles de clientes de la cadena de hoteles requiere investigación, diseño efectivo de los anuncios, y selección de las plataformas publicitarias adecuadas. Además, es importante monitorear y medir el rendimiento de los anuncios para asegurarnos de que están siendo efectivos.

5. Armar un modelo de árbol de decisión para el ofrecimiento de publicidades.

Los árboles de decisión sirven para sugerir definitivamente que publicidad ofrecer a cada pasajero que encienda la tv de su habitación con la tarjeta de acceso. Esta decisión surge del atributo de clase que el pasajero posee y la misma puede ir modificándose a medida que este haga uso de las instalaciones donde utiliza la tarjeta de acceso o compras de manera dinámica.

Si bien el total de opciones de decisión de este árbol tiene 4x4x4 = 64 posibilidades, muchas publicidades se repiten en los grupos correspondientes a los nodos hoja o finales, dado que todos los seres humanos tenemos necesidades básicas comunes como alimentación, descanso y diversión.

A medida que el cliente va haciendo uso de las instalaciones del hotel y los comercios internos donde utiliza la tarjeta personal de acceso, estos eventos se van sumariando a su perfil, marcando un comportamiento de la actual estadía.

Resolución objetivo 2 – Campaña de Mailing

B. Decidir sobre la selección de hoteles para el mailing privado durante el verano.

1. Toma final de datos

Una vez que el cliente realiza el check out en uno de los hoteles de la cadena se dispara el proceso para decidir qué ofertas, paquetes o promociones contemplar en una campaña de mailing. A continuación, explicamos algunas técnicas utilizadas para esta campaña.

2. Clasificar de los hoteles que conforman la cadena según perfil de clientes para el envío de mailing:

Utilizando la técnica de clustering explicada anteriormente, ordenaremos todas las características de nuestros hoteles con los amenities correspondientes y sus calificaciones: de 5 a 3 estrellas.
En base a esto esperamos que el algoritmo nos agrupe en segmentos hoteleros donde podamos enviar los mismos tipos de mailing según los servicios ofrecidos.
La idea principal de esto es poder tener las ofertas de los hoteles agrupadas en el data lake, y que, según el perfil del cliente clasificado posteriormente, se envíen automáticamente mails centralizados recomendando ciertas actividades.

3. Clusterizar a los clientes para el envío de mailing de ofertas veraniegas.

La clusterización de datos de clientes de la cadena de hoteles consiste en conglomerarlos según sus características y preferencias similares. Esto permite segmentar a los clientes en grupos homogéneos y personalizar las ofertas y promociones en función de sus necesidades y gustos.

Por otro lado, cabe aclarar que esta técnica es diferente a la de clasificación de clientes recomendada en el punto 2 de la solución anterior en donde asignamos a cada cliente una categoría o clase específica en función de sus características y comportamientos que nos permitan identificar un perfil.

Clusterizando podemos construir una campaña de mailing para que los conjuntos obtenidos del modelo con recomendaciones específicas de ofertas, paquetes y opciones de alojamientos dentro de los hoteles de la cadena.

Ambas técnicas tienen aplicaciones relevantes en la gestión de clientes en la cadena de hoteles, pero la clusterización se enfoca en agrupar a los clientes en conjuntos similares y la clasificación en asignar a cada cliente a una categoría específica.

07.-Conclusiones

Para resolver los objetivos planteados y los subproblemas intermedios utilizamos el método KDD aplicando ciertas técnicas de data mining como ser: clasificación, clusterización, patrones de comportamiento y arboles de decisión. Establecemos herramientas de recolección de datos a partir del check in, el uso de la tarjeta del hotel, datos demográficos y el check-out para clasificar a los clientes en determinados perfiles (que se actualiza cada vez que recibimos nuevamente al huésped y se reclasifica). Esto nos permitió ayudar a la empresa a conocer mejor a sus clientes y ofrecerles una experiencia personalizada y adaptada a sus preferencias, lo que puede aumentar su satisfacción y fidelización y generar mayor rentabilidad. El subproblema del ofrecimiento de publicidades iniciales fue resuelto por la clasificación inicial del huésped al momento del check in sumado al posible historial de alojamientos anteriores.

Ante el desafío planteado confirmamos que la minería de datos es una técnica que permite analizar grandes volúmenes de datos que, si bien pueden ser históricos, también se van modificando en tiempo real y que al obtener información relevante que puede ser utilizada para mejorar la toma de decisiones en una organización, colabora con esta para mejorar el objetivo del negocio que es ofrecer una mejor estadía a nuestros clientes. En el caso de esta cadena hotelera, pudo ser utilizada para obtener información valiosa acerca de sus clientes, sus comportamientos, sus preferencias actuales y nuestras ofertas de servicios, lo que permite a la empresa tomar decisiones más acertadas y eficientes en cuanto a la gestión de la calidad de atención personalizada de nuestros huéspedes. Además, la minería de datos puede ayudar a la compañía a identificar tendencias y patrones en el comportamiento de sus clientes, lo que le permite ofrecer servicios y promociones que se ajusten a sus preferencias y necesidades, buscando de esta manera aumentar la fidelidad del cliente y por ende rentabilidad del negocio.