Patrón de fondo

Descripción generada automáticamente con confianza media

Universidad Austral - Maestría en Ciencia de Datos

Caso #1

**Autor: Andres Montes de Oca**

Fecha: 30 de Noviembre de 2022

Introduccion al

Data Mining

Explicación de los pasos del proceso de KDD, comparándolo con el ejemplo de la Cadena de Hoteles

En el siguiente documento se explicaran los 7 pasos en el proceso de Konwledge Discovery in Data, haciendo una analogía con el Caso Practico #1.

1. **Data Cleaning**: En este paso se depura el DataSet de ruido y de datos inconsistentes que ensucian mi datos. Un claro ejemplo de esto seria lidiar con los valores nulos que encontramos en los datos luego de que el husped llena el cuestionario de ingreso al hotel. Muchas veces estos cuestionarios no están 100% completados y los mismos tienen varios sino todos sus valores nulos. Si son solo algunos datos faltantes, aquí se puede decidir entre varias opciones. Se pueden completar los datos faltantes con la media/mediana si son datos cuantitativos (dependiendo si se observan outliers o no), o con la moda en caso de datos cualitativos. En cambio si todos los datos están faltantes, probablemente la mejor idea seria dropear los mismos, ya que no me aportarían ningún insight al modelo. En el formulario de evaluación que el husped llena al check-out, también habría que limpiarlo con alguna de estas opciones.
2. **Data Integration**: Aquí es cuando unimos o combinamos las diferentes fuentes de datos en una sola. En nuestro caso de ejemplo se utilizaría cuando el husped hace el ingreso, ya que deberíamos checkear con la base de datos de clientes pasados que tenemos para confirmar si ya tenemos información sobre él, para poder ofrecerle productos y servicios que ya hayan sido de su interés. Otra cosa que se me ocurre, aunque no se si éticamente seria correcto, seria hacer una integración de datos recopilados desde las redes sociales publicas del huésped, también para poder ofrecerle cosas de su interés.
3. **Data Selection**: Aquí definimos que variables de todas las disponibles que tenemos, vamos a utilizar en la creación del modelo. Esto si bien es mucho prueba y error, creo que es importante focalizarnos en cual es el propósito puntual que queremos atacar. En nuestro caso de ejemplo no es lo mismo los datos que voy a usar como input cuando el usuario llega al hotel por primera vez (ya que no tengo datos pasados), que los datos que voy a usar durante la estadía del huésped en el hotel (los gastos en servicios y ammenities visitados), y serian todavía distintos a los datos que usare para el sistema de recomendaciones del mailing enviado a potenciales huéspedes para la próxima temporada.
4. **Data Transformation**: Acá es cuando consolidamos o transformamos la data de una manera apropiada, para que nos sea mas sencilla la obtención de insights, o que los mismos sean mas fácilmente interpretables. Una técnica muy conocida para esto es PCA, en la cual en un dataset con variables altamente correlacionadas puede ser reducido a unas pocas, con muy poca perdida de información. En nuestro caso de ejemplo se me ocurre que en esta etapa podría caer los procesos automáticos de nuestro Data Warehouse, en el cual sumarizamos según distintos cortes, la información de todas las variables que disponemos de nuestros datos. Esto luego simplificaría la complejidad en el minado de insights de nuestros datos
5. **Data Mining:** Es el paso principal en el cual se implementan los diferentes algoritmos para identificar los patrones. En nuestro ejemplo se me ocurren varias aplicaciones: Al principio cuando el huésped llega al hotel por primera vez, si no tenemos muchos conocimientos previos de sus preferencias, en el formulario de ingreso podríamos preguntar muy pocas preguntas obligatorias. Las mismas nos van a ayudar en la generación de un modelo de Clustering con el cual podremos categorizar a nuestro husped en una categoría, la cual se usaría para el envio de las publicidades en el canal privado de TV. Otra cosa que se me ocurre seria la implantación de un algoritmo de Clasificación, usado para identificar a que hoteles de mi cadena un potencial husped recurrente puede estar interesado en ir, y enviarle así un correo mas personalizado. También se me ocurre, que para huéspedes ya conocidos, se podría llegar a implementar un algoritmo de Regresión en cual identifique las fechas potenciales en las cuales nuestro cliente puede tomarse sus próximas vacaciones.
6. **Pattern Evaluation:** Aquí utilizamos distintas medidas de evaluación sobre los modelos creados. No veo como machear esto con el caso, pero se me ocurre que podría llegar a ser los valores del Accuracy de nuestro sistema de clasificación, el cual nos indica el porcentaje de acierto en la selección de hoteles para las próximas vacaciones.
7. **Knowledge Presentation:** Visualizacion de los patrones e insights descubiertos en los datos, generalmente utilizando graficas que son mas fácil de interpretar. Estos podrían ser los distintos reportes generados a partir de los datos de entrada de cada huésped, como así también el mailing automatico o semi-automatico enviado a los potenciales huéspedes.