Examen ACUE 2024

# 2) Ejercicio del Taxi

Tenemos que las probabilidades de que un coche sea de cada compañía son:

* Verde: 0.7
* Azul: 0.2
* Rojo: 0.1

Y la probabilidad de que el testigo diga Azul y sea de cierto color es:

* Verde: 0.2
* Azul: 0.8
* Rojo: 0.2

Entonces, la probabilidad de que el coche sea Azul si el testigo lo identifica como Azul puede ser calculada mediante el teorema de Bayes, el cual he implementado en código Python. El resultado es una probabilidad del 0.67 de que sea el Azul.

# 3) Preferencias de usuario

Las preferencias de usuario basadas en el contenido se obtienen calculando las frecuencias de cada articulo en el total de compras del usuario. En este ejemplo, se han comprado 9 elementos en total y 1 cerveza, por lo que la preferencia del usuario por la cerveza es de 1/9. Los resultados totales, resueltos mediante código, son:

* Milk -> 0.222222 (2/9)
* bread -> 0.333333 (3/9)
* butter -> 0.222222 (2/9)
* beer -> 0.111111 (1/9)
* diapers -> 0.111111 (1/9)

# 4) Proyecto Data Science

La fase del proyecto más importante es el **Business Understanding**, ya que establece los objetivos y el alcance del proyecto de forma que se desarrolle un producto que satisfaga las necesidades del cliente. Si en esta fase no se comprenden las necesidades y requisitos del cliente y no se integran con las capacidades y limitaciones técnicas del equipo de desarrollo, el proyecto estará abocado al fracaso.

La fase del proyecto que mas tiempo requiere es el **Data Preparation**, ya que incluyen recolectar los datos (que pueden estar fragmentados en múltiples y no compatibles fuentes), su limpieza, transformación e integración, es decir, el proceso de ETL. Depende de como el cliente proporcione los datos, pero lo normal es que requieran de una gran inversión de labor para transformarlos en una fuente de calidad.

Los perfiles más idóneos para cada fase del proyecto serían:

1. **Business Understanding**: Aquí entrarían los analistas, sean de negocio, de requisitos de software o de economía ya que son estos analistas los que están preparados para comprender al cliente y sus necesidades de negocio y traducir estas a requisitos y funcionalidades del producto a desarrollar.
2. **Data Understanding**: Aquí resaltan los Data Scientist que pueden ser ingenieros informáticos, matemáticos y estadísticos, ya que deben analizar los datos que el cliente posee y buscar en ellos que procesos de ETL necesitan, que formato deben obtener para después en Modeling alimentar los modelos y que patrones pueden identificarse en ellos.
3. **Data Preparation**: Aquí entrarían los Data Engeneer sobre todo de ingeniería informática, que implementan lo estipulado por los Data Scientist mediante una pipeline ETL. Este paso puede llegar a ser muy técnico y tedioso y los ingenieros informáticos son los mejores formados para abordarlo.
4. **Modeling**: Aquí vuelven los Data Scientist, en este caso como los datos ya están preparados podrían seguir los ingenieros informáticos pero también podrían volver los matemáticos y los estadísticos ya que se reduce la complejidad del uso de los datos.
5. **Evaluation**: Aquí sigue siendo un Data Scientist pero más orientado a los matemáticos y estadísticos ya que reconocer errores y rendimiento de los modelos predictivos requiere de buenos conocimientos de la teoría tras los cálculos, conocer el qué se está calculando y el porqué para determinar si es lo mas correcto para cada caso de uso.
6. **Deployment**: Este es el ámbito del ingeniero informático, los DevOps sean con kubernetes, dockers u otras soluciones más corporativas como AWS son áreas de conocimiento en las que el ingeniero informático se forma.

# 1) Recomendación de tapas

Para resolver este problema primero vamos a observar los datos que tenemos y extraer información de ellos.

Podemos ver que la tapa “t1” es de tipo Daring, su popularity es 39, su media es 4.39, la media de su restaurante 4.28 y tenemos las reseñas de 37 usuarios que la han probado.

Por el otro lado, el usuario “u31” ha puntuado 10 tapas de 10 restaurantes distintos, 8 de las cuales son de carácter “Traditional” y 2 son “Daring”.

No tenemos los suficientes datos para calcular la similitud entre usuarios y generar así una puntuación basada en vecinos. Tampoco creo que sea de calidad una respuesta basada en las valoraciones del usuario a tapas “Daring” ya que solo ha probado 2, y la solución a parte de ser trivial (una media de los dos valores que daría 4.5) no está basada en suficientes datos.

Podríamos modelar la relación entre la popularidad de las 10 tapas que ha tomado el usuario y su puntuación, para extraer un valor para la tapa “t1” pero no creo que sea una respuesta de valor ya que se basa en un valor ajeno al usuario, la “popularidad” de la tapa.

Por lo tanto, implemento un recomendador colaborativo basado en las opiniones de la comunidad, es decir, calculo la media de todas las valoraciones de los usuarios que han consumido la tapa, el cual es 4.43.

Para poder expandir los datos, lo primero sería tener más valoraciones para cada usuario que no es el u31, por ejemplo 10 para cada usuario, y así poder buscar vecinos mediate similitud con los cuales crear un recomendador basado en vecinos.

También tener mas datos sobre las tapas permitiría crear recomendadores más precisos. Valores como dulce o salado, precio o peso podrían permitir obtener patrones de preferencia de los usuarios.

Con todo esto, implementaría un recomendador basado en vecinos.