UNIVERSIDAD
DE ANTIOQUIA

TRABAJO ANALÍTICA DE DATOS APLICADA EN MARKETING

SANTIAGO GÓMEZ BERRÍO
DIEGO ANDRÉS LUNA PATERNINA
MARIA CLARA SALAZAR DUQUE

JUAN CAMILO ESPAÑA LOPERA
ANALÍTICA DE DATOS

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
MEDELLÍN

2023

Análitica de Datos Aplicada en Marketing

Diseño de la solución:

La propuesta de solución como se muestra en la *figura 1* será llevada a cabo por el equipo de analítica. Esta consiste en extraer la información de los usuarios de la plataforma online, guardarla en una base de datos y alimentar el modelo de sistemas de recomendación que le permitirá al área de marketing generar estrategias relacionadas con los gustos de los usuarios o por contenido como tal, de manera general y personalizada con la información recolectada de acuerdo a su interacción en la plataforma online. Una vez obtenidas las estimaciones de lo que le podría gustar a los usuarios, se alimenta una base de datos de recomendaciones, para proceder a enviar las sugerencias a cada usuario diariamente, esto con el apoyo del área de tecnología y desarrollo de interfaz, teniendo la posibilidad de saber cuántos vieron las sugerencias, y así analizar si les gustó por medio de la calificación que les den a las películas. De esta manera, evaluar la satisfacción del usuario para ver si hay que ajustar alguno de los modelos de recomendaciones.

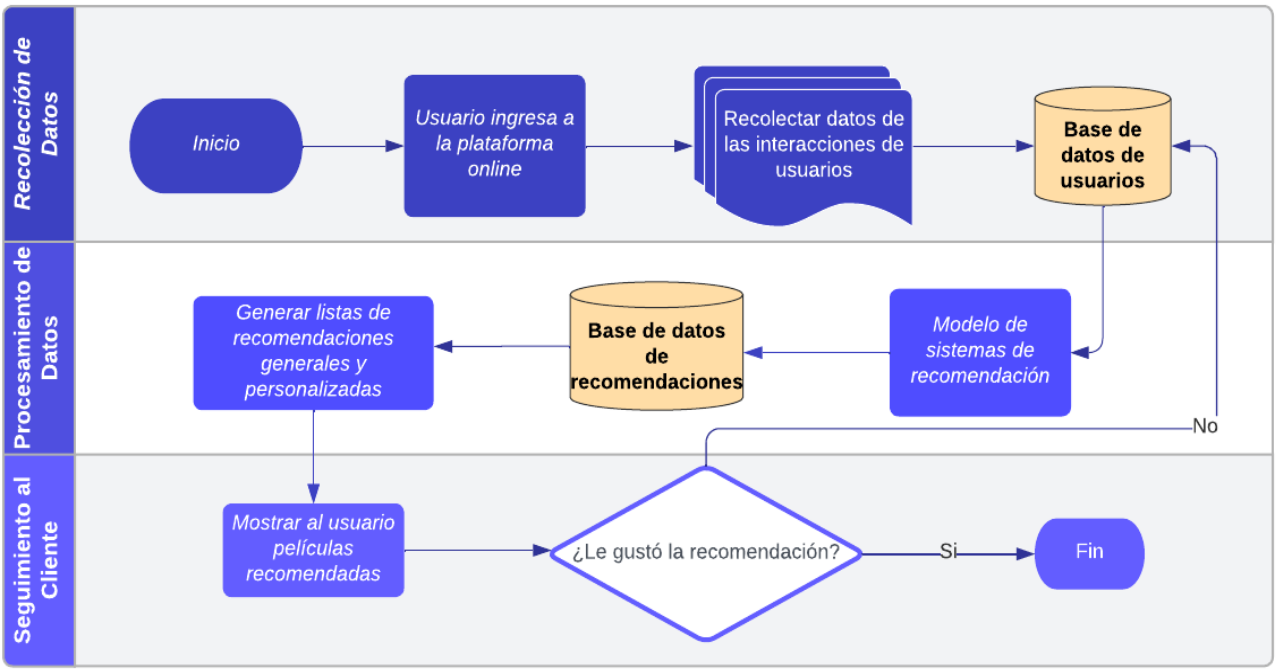

Figura 1. *Diagrama diseño de solución*

Limpieza y transformación:

Lo primero que se hizo con las tablas movies y ratings fue verificar si contenían información o valores nulos, en este caso ninguna de las dos tenía valores nulos, luego se verificaron los duplicados y posteriormente se transformó el formato de la fecha para que fuera más práctico, entendible y útil para poder realizar series de tiempo en el análisis exploratorio. También se quitaron los años de la variable título, para que solo apareciera el nombre de la película.

Análisis exploratorio:
Luego se pasó a hacer un análisis exploratorio de los datos donde se encontraron hallazgos relevantes como:

- La mayoría de películas tiene calificaciones entre 3 y 5 puntos.
- La película más vista fue Forrest Gump (1994) con 329 vistas, y en general las 10 películas más vistas tienen en promedio más de 220 visualizaciones.
- También se hizo una clasificación por géneros donde se encontró que los géneros más vistos son comedia, romance, guerra y drama con 329 vistas cada uno.
- Se encontró que 26.818 películas tienen una clasificación promedio de 4 mientras que 13.211 tienen una calificación promedio de 5 puntos de rating.
- El género con más películas es el de comedia con casi 3000 películas, mientras que el de menos es el musical.
- En las series de tiempos se pudo visualizar que a partir del año 2000 hubo un pico de visualización de películas en la plataforma, siendo julio es el mes de mayor concurrencia.
- Se identificó que el usuario con más vistas tiene 2698 vistas mientras que el usuario de menos vistas tiene 20.
- Se encontró que no hay relación entre el año en el que salió la película con su calificación, ya que esta no depende del tiempo, sino más bien de los gustos de cada usuario por eso se tienen películas muy antiguas calificadas con la puntuación más alta, como películas muy recientes también calificadas con la mejor puntuación.

Selección de variables:
Para una mejor manipulación de los datos, se unieron las tablas, conteniendo la calificación promedio de cada película, el número de vistas, el nombre de la película, los géneros y la movie id. Posteriormente, se separaron los géneros por medio de dummies ya que por cada película aprecian varios evitando una correcta clasificación. Todo esto se realizó para brindar recomendaciones no duplicadas.

Evaluación y análisis de modelos:
Para realizar los diferentes modelos de sistemas de recomendación se tuvieron en cuenta los siguientes:

- **Sistemas basados en popularidad**, el cual consiste en hacer rankings de las mejores películas, el número de visualizaciones y según la calificación de los usuarios.
- **Sistemas de recomendación contenido general**, de acuerdo a cualquier película vista por el usuario se le recomienda una nueva respecto a la correlación que tengan entre sí, es decir, que tengan características similares, por ejemplo si sale la segunda parte de una película anteriormente vista.
- **Sistemas de recomendación filtro colaborativo basado en películas:** En este sistema de recomendación, se tiene en cuenta una película con la cual se va a realizar la predicción, es decir, esta se ingresa de manera manual y arroja los resultados por medio de una función en la que se definen las correlaciones con las mismas películas y se presentan las que cuentan con un mayor puntaje. También se realizó una lista

desplegable en la que se tienen en cuenta todas las películas que servirá para observar diferentes recomendaciones de manera dinámica.

- **Sistema de recomendación basado en contenido KNN:** De acuerdo al anterior sistema y a manera de seguir profundizando en las recomendaciones, se tiene en cuenta este sistema recomendando las películas similares en género y año. Al realizar comparaciones con este y el anteriormente mencionado (basado en contenido), no se detectan diferencias por lo que se encuentran funcionando de manera similar y utilizando las mismas variables.
- **Filtro colaborativo basado en el usuario:** De acuerdo a este modelo, se realizan las predicciones de las recomendaciones evaluadas de acuerdo con el MAE, RMSE, fit time y test time, donde se evidencia que todos los modelos son similares en estos resultados. Por lo tanto, para seleccionar el algoritmo se tiene en cuenta el más equilibrado en las cuatro métricas evaluadas, siendo el KNN BASIC. Si bien es cierto, el modelo se adapta también a los costos de ejecución, ya que al verificar, es el que menos toma tiempo de procesamiento.
- **Filtros y contenidos:** Utiliza la información de las calificaciones de los usuarios para calcular las películas más recomendadas en un género seleccionado. La selección del género se realiza a través de una lista desplegable y el sistema actualiza automáticamente las recomendaciones en función de la selección del usuario. Posteriormente, el usuario selecciona un género de una lista desplegable, se realiza una consulta SQL para obtener todas las películas que pertenecen al género seleccionado y para calcular el puntaje de cada película en función de las calificaciones de los usuarios que también vieron películas en ese género. Después, se ordenan las películas según su puntaje y se muestran las 10 películas más recomendadas. Si el usuario selecciona un género diferente, se repiten los pasos anteriores para mostrar las películas recomendadas para ese género. Este código primero obtiene todos los géneros de la base de datos y crea un conjunto de géneros únicos. Luego, muestra una lista de géneros disponibles y le pide al usuario que seleccione el número correspondiente al género que le interesa. Después de que el usuario selecciona el género, el código ejecuta para obtener las 10 películas más recomendadas del género seleccionado. Finalmente, muestra los resultados al usuario.

Selección de modelos y despliegue:

Para hacer el despliegue del sistema de recomendaciones lo primero que se necesita es diseñar un espacio determinado en la plataforma, de esta manera en cuanto el usuario entre lo primero que verá será una pestaña con el nombre de tendencias en donde aparecerán las 15 películas más vistas en el último día, para esto se usa el sistema de recomendación KNN para generar dicha recomendación y mediante una función o tarea de automatización que ejecute el modelo de recomendación en batch diariamente, a las tres de la madrugada, este horario es definido con la intención de hacer la actualización en horas muertas de la plataforma evitando sobrecargar los servidores, la tarea tomaría como entrada los datos actualizados de la plataforma de películas y produciría un archivo tipo CSV con las recomendaciones para cada usuario, que luego podrían ser enviadas al servidor de la plataforma y ésta notificará a los usuarios correspondientes.

Mientras que las recomendaciones por filtro y contenido de cada usuario se haría con un entrenamiento en batch diario que se encargue de recoger los datos más recientes y generar recomendaciones de las 5 películas más recomendadas para cada usuario por día, crea un archivo .pkl que se envía al servidor de la plataforma la cual actualiza las recomendaciones diariamente a las 3:00 a.m., horario con bajo tráfico en la plataforma, donde al día siguiente al usuario le aparecerá no solo la pestaña de tendencias donde estarán las películas recomendadas para todos los usuarios en general, sino que también le saldrá un nueva pestaña llamada "Para mí", anteriormente diseñada por el equipo de desarrolladores, donde encontrará una lista de películas recomendadas especialmente para él o ella según sus gustos y preferencias.

Conclusiones:

De acuerdo a los modelos desarrollados, se puede decir que todos cuentan con una capacidad de recomendación útil para el usuario. Por lo tanto, esta selección depende de la empresa, puesto que todos tienen enfoques diferentes para brindar las recomendaciones. En este caso se escogieron los sistemas de recomendación KNN y recomendación por contenidos puesto que presentan menor tiempo de procesamiento y se adaptan mejor a la solución planteada.

Adicionalmente, es muy importante que la empresa encargada tenga un desarrollo web que permita gestionar la plataforma, revise las tendencias y la interacción que tienen los usuarios con las películas, para el ingreso y salida de cada una de ellas.

Finalmente, la empresa se encuentra utilizando técnicas de machine learning para nutrir su plataforma, por lo que se recomienda que más adelante se creen modelos más robustos con técnicas de Deep learning que permitan recolectar otros tipos de datos para seguir alimentando los diferentes modelos a realizar.