Introducción a la Ciencia de Datos

Maestría en Ciencia de Datos y Aprendizaje Automático

Facultad de Ingeniería, UdelaR

11/07/2024

Tarea 3

Joana Auriello, Pablo Molina

Introducción

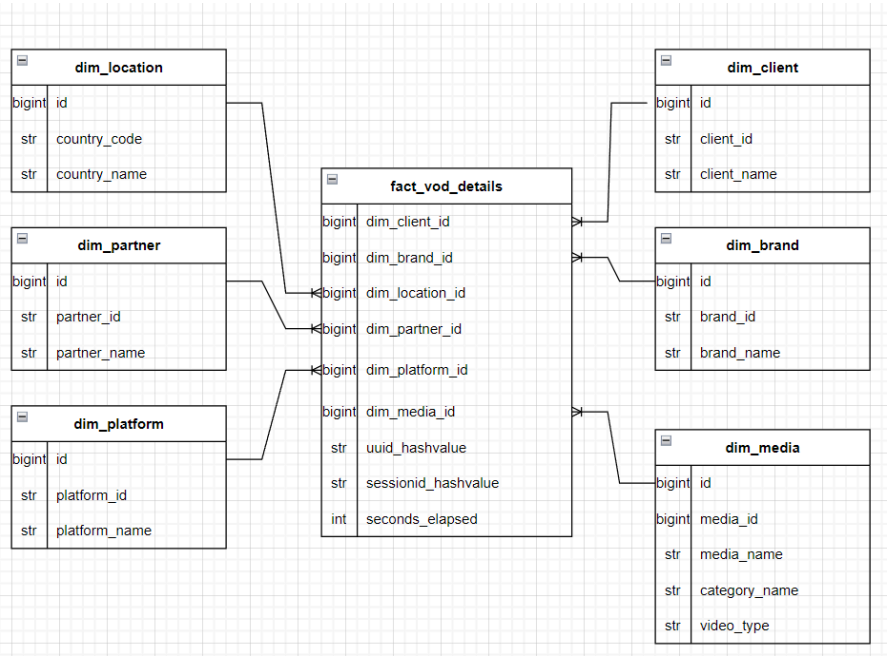
En este informe, se aborda la problemática de la segmentación de usuarios en plataformas de streaming con el objetivo de mejorar la personalización de contenido y las estrategias de retención. Utilizando datos provenientes de los logs de aplicaciones de servicios de streaming, que incluyen información sobre dispositivos, sistemas operativos, plataformas y ubicaciones geográficas. Este estudio se centra en analizar la calidad de los datos y aplicar métodos de ciencia de datos para segmentar a los usuarios basándose en sus patrones de visualización y otras características.

Datos: descripción y calidad

Los datos a analizar surgen de los logs de aplicaciones de servicios de streaming. De estos logs se extraen datos vinculados a los dispositivos, sistemas operativos, plataformas, ubicaciones geográficas, entre otros, de los usuarios que están consumiendo el contenido disponible en estas apps de streaming. Estos logs son procesados a través de pipelines ETL en la nube y se cargan a un data warehouse periódicamente.

Por razones de granularidad de los datos y privacidad, a pesar de que mantenemos una muestra representativa, en este trabajo se presenta solamente una parte del esquema completo con el que se trabaja usualmente.

Modelo de datos



En este modelo, la tabla fact_vod_details actúa como tabla de hechos y contiene los datos principales de visualización, incluyendo uuid_hashvalue(llave primaria), seconds_elapsed y sessionid_hashvalue. Las llaves foráneas en esta tabla (dim_client_id, dim_brand_id, dim_location_id, dim_partner_id, dim_platform_id, dim_media_id) se relacionan con las

tablas dimensionales correspondientes. La tabla dim_client incluye id (llave primaria), client_id y client_name. La tabla dim_brand contiene id (llave primaria), brand_id y brand_name. La tabla dim_location tiene id (llave primaria), country_code y country_name. La tabla dim_partner incluye id (llave primaria), partner_id y partner_name. La tabla dim_platform contiene id (llave primaria), platform_id y platform_name. Finalmente, la tabla dim_media incluye id (llave primaria), media_id, media_name, category_name y video_type. Estas relaciones permiten la integración de datos detallados sobre clientes, marcas, ubicaciones, sistema operativo, plataformas y títulos

A partir de este modelo se genera un dataset que contiene las siguientes columnas:

- client_id, corresponde al número de cliente, dueño del canal.
- client_name, nombre del cliente.
- brand_id, código del canal
- brand_name, nombre del canal
- country_code, código único del país.
- country_name, nombre del país.
- partner_id, codigo del dueño de la plataforma.
- partner_name, nombre del dueño de la plataforma
- platform_id, codigo del sistema operativo.
- platform_name, nombre del sistema operativo.
- media_id, codigo del producto que fue visto.
- media_name, nombre del producto que fue visto.
- category_name, categoria del canal.
- video_type, tipo de video visto (película, episodio, clip, tráiler)
- UUID_HASHVALUE, código para identificar usuarios, se crea a partir de atributos del dispositivo, geográficos y la IP (solo se usa cuando se tiene restricciones para compartir la información) esto se hace para poder identificar el usuario sin comprometer los datos personales.
- seconds_elapsed, tiempo total que la persona permanece en la aplicación.
- sessions, número total de sesiones, definido como count distinct de sessionid_hashvalue, siendo este último un identificador basado en uuid_hashvalue y un componente de tiempo.

Calidad de datos

Los datos cuentan con cierto tratamiento y transformaciones para su calidad, algunos de los siguientes puntos listados ya que fueron corregidos o están en proceso de corrección:

- **Inconsistencia en las Categorías:** Dado que cada cliente define sus propias categorías en su aplicación, es posible encontrar categorías similares con nombres diferentes, como "Comedia" y "Comedy". Esto puede dificultar el análisis agregado y la comparación entre usuarios. Existe una categoria universal de categorías denominadas IAB (Interactive Advertising Bureau), para solucionar este problema se puede matchear las categorias proveniente de cada brand con el estándar internacional.

- **Datos Faltantes o Incompletos:** Puede haber **registros con valores faltantes** en campos clave como client_name, brand_name, country_name, etc. Esto afectaría la precisión de los análisis y reportes.
- **Duplicación de Datos:** Podrían existir **duplicados** en la tabla fact_vod_uu_details o en las tablas dimensionales, lo que inflaría los resultados de conteos y sumas.
- **Valores Nulos o Incorrectos en Llaves Foráneas:** Las llaves foráneas en fact_vod_uu_details pueden tener valores nulos o incorrectos, impidiendo la correcta relación con las tablas dimensionales y afectando la integridad referencial.
- **Problemas de Integridad Referencial:** Falta de correspondencia entre las llaves primarias de las tablas dimensionales y las llaves foráneas en la tabla de hechos puede llevar a registros huérfanos o a la imposibilidad de realizar uniones correctas.
- **Formato de Datos Inconsistente:** Diferencias en el formato de datos, como fechas en diferentes formatos o códigos de país inconsistentes (USA vs US), pueden complicar la consolidación y análisis de datos.

Para abordar estos problemas, se podría implementar procesos de limpieza y estandarización de datos, como la normalización de categorías, la validación de llaves foráneas, la deduplicación de registros y la implementación de controles de calidad de datos durante la recolección y carga de datos.

Problemática a resolver con herramientas del curso

Problemas/Preguntas a Resolver

Basándonos en sus patrones de visualización y características demográficas, surgen los siguientes problemas que pueden ser resueltos con datos:

- ¿Cómo podemos segmentar a los usuarios para mejorar la personalización de contenido y las estrategias de retención?
- ¿Podemos predecir el tiempo que permanecerán en la aplicación?
- ¿Es posible automatizar el proceso de mapeo entre las categorías estándar y las utilizadas por cada aplicación?

Descripción de la implementación de la solución

Métodos

Para resolver este problema, se puede seguir un proceso estructurado que involucre varias etapas de la ciencia de datos, aplicando conceptos y herramientas presentadas en el curso.

1. Recolección y Exploración de Datos

- **Recolección**: Usar la consulta proporcionada para recolectar datos relevantes de las tablas fact_vod_uu_details, dim_client, dim_brand, dim_location, dim_partner, dim_platform y dim_media.
- **Exploración**: Realizar un análisis exploratorio de datos (EDA) para entender las distribuciones, valores atípicos y patrones generales. Visualizar datos usando histogramas, gráficos de barras y matrices de correlación para identificar relaciones importantes entre variables.

2. Limpieza y Calidad de Datos

- **Limpieza**: Tratar los valores faltantes y duplicados. Normalizar las categorías de medios para evitar inconsistencias. Validar y corregir las llaves foráneas para mantener la integridad referencial.
- **Calidad de Datos**: Implementar controles de calidad para asegurar que los datos sean precisos, completos y consistentes.

3. Segmentación y Aprendizaje Automático

- **Segmentación de Clientes**: Utilizar métodos de aprendizaje no supervisado como K-means clustering para segmentar a los usuarios en grupos basados en sus patrones de visualización y características demográficas.
- **Predicción de duración de la sesión:** Utilizando Random Forest, este algoritmo se puede utilizar tanto para problemas de clasificación como de regresión, su faceta de regresión será la que nos permita a partir de las características del usuario estimar el tiempo que este permanecerá en la aplicación.
- **Automappeo de categorias**: En este caso es posible la utilización de medidas de distancia que junto con un algoritmo de predicción nos permita comparar el parecido entre nuestras categorías y el estándar, definiendo un threshold de certeza podemos automatizar que se remplacen aquellas categorías que tienen una distancia menos o un mayor parecido con las categorías del sistema de clasificación estándar.

4. Visualización e Interpretación de Resultados

- **Visualización de Resultados**: Crear gráficos de dispersión, diagramas de cajas y gráficos de radar para visualizar los segmentos de usuarios. Usar mapas de calor para visualizar la correlación entre las variables y los clústeres formados.
- **Interpretación**: Interpretar los resultados para entender las características y comportamientos de cada segmento. Identificar patrones comunes y diferencias clave entre los segmentos.

5. Aplicaciones Prácticas

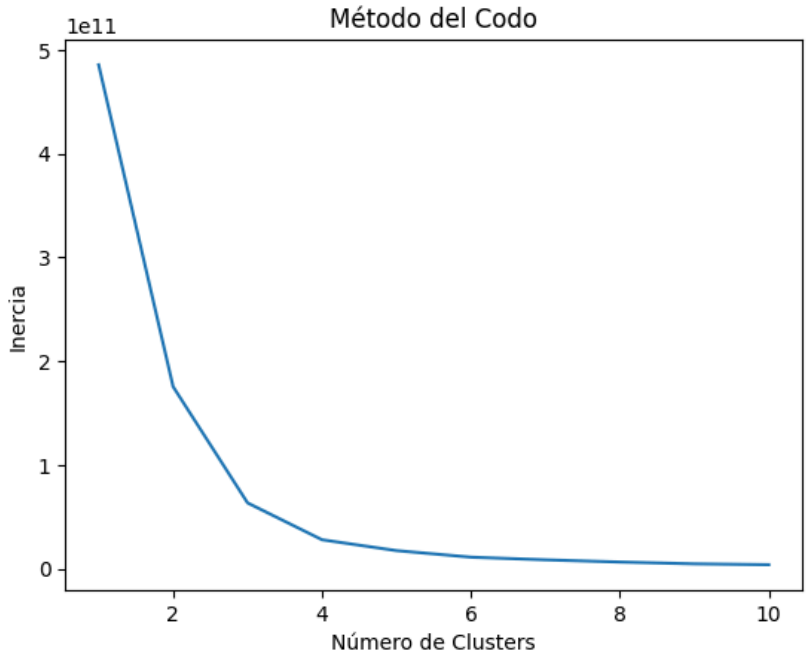
- **Personalización de Contenido**: Utilizar los segmentos de clientes para desarrollar recomendaciones personalizadas de contenido.
- **Estrategias de Retención**: Diseñar estrategias de retención específicas para cada segmento basado en sus características y comportamientos.

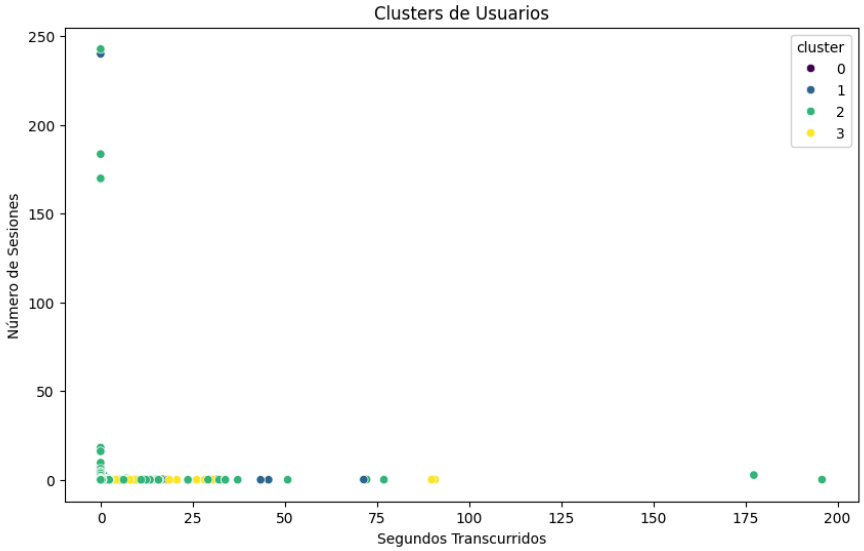
Anexo: Implementación

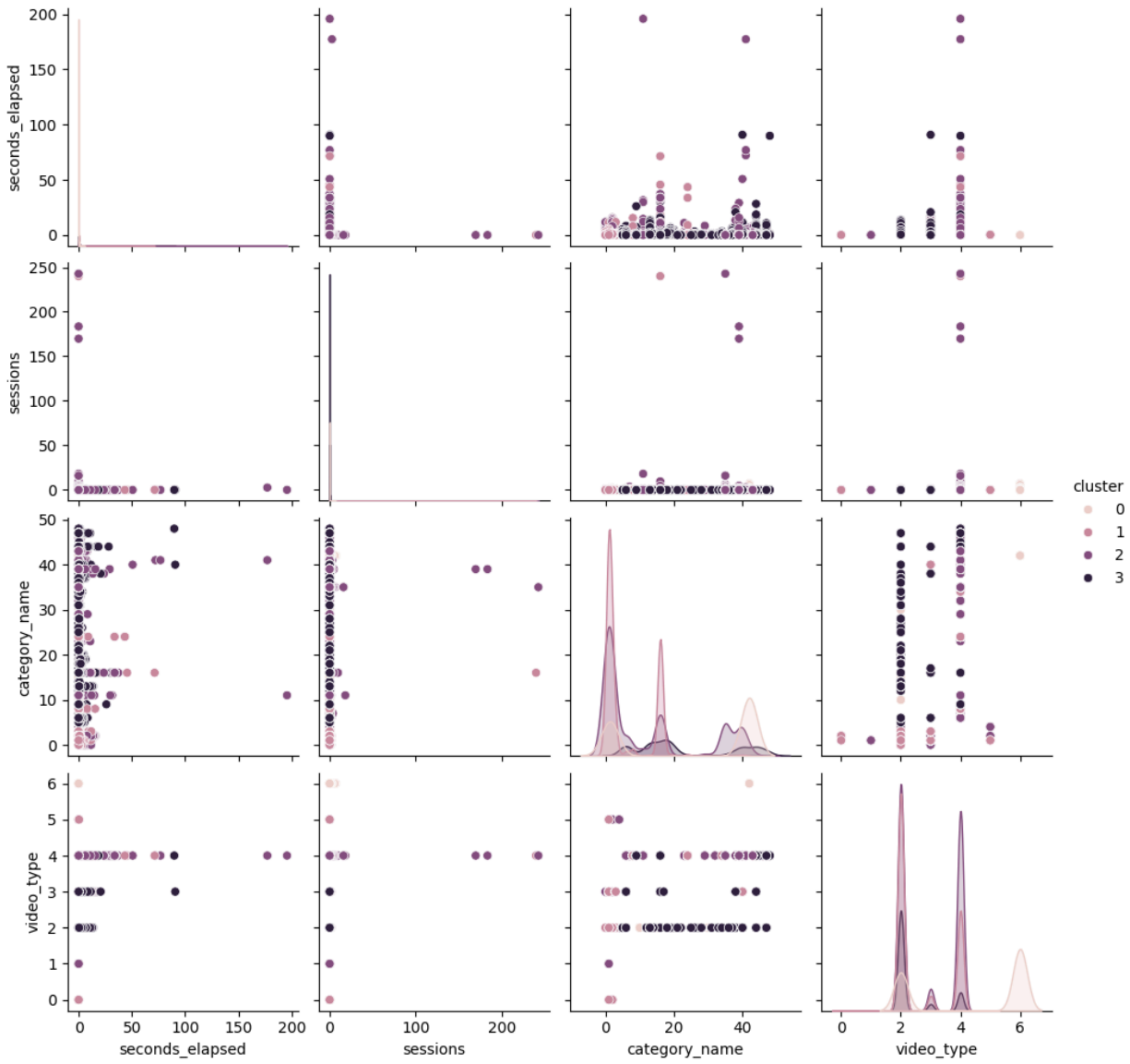
A modo de prueba buscamos lograr una prueba de concepto de cómo funcionaria el análisis de K-means en la base de datos planteada.

Visualizaciones

Siguiendo el método del codo identificamos que el número ideal de clusters sería 4.







Las principales diferencias entre clusters están dadas por las categorías y los tipos de videos que se ven.

n=4

cluster	client_id	brand_id	country_code	partner_id	platform_id	media_id	category_name	video_type	seconds_elapsed	sessions
0	0	1.954065	109.843488	50.686975	1.320845	315.270463	26.344578	4.464142	-0.08586	-0.004419

1	0	1.998967	95.401462	46.217162	1.069932	3985.892734	5.851944	2.648742	-0.037079	0.001193
2	0	1.764959	106.282952	56.130338	1.131405	2497.916466	12.624513	2.939894	-0.000224	0.005801
3	0	1.996013	112.517399	45.787697	1.68895	5584.190224	23.345692	2.345588	0.248858	-0.016577

n=3

cluster	client_id	brand_id	country_code	partner_id	platform_id	media_id	category_name	video_type	seconds_elapsed	sessions
0	0	1.998152	100.071328	46.450271	1.260963	4443.17775	10.802647	2.572941	0.045648	-0.003644
1	0	1.954623	109.847692	50.676345	1.32085	315.719216	26.334441	4.463135	-0.085815	-0.004428
2	0	1.771743	105.987627	55.513435	1.109591	2524.300533	12.287326	2.924514	-0.000816	0.005634