****

**Análisis de Grandes Volúmenes de Datos (Gpo 10)**

**David Nava Jiménez - A01168501**

**Edwin David Hernández Alejandre - A01794692**

**Jorge Fernando Bonilla Diaz - A01793935**

**Nombre del entregable:**

Avance de proyecto 1: Sistema de Recomendación

**Domingo 12 de Mayo de 2024**

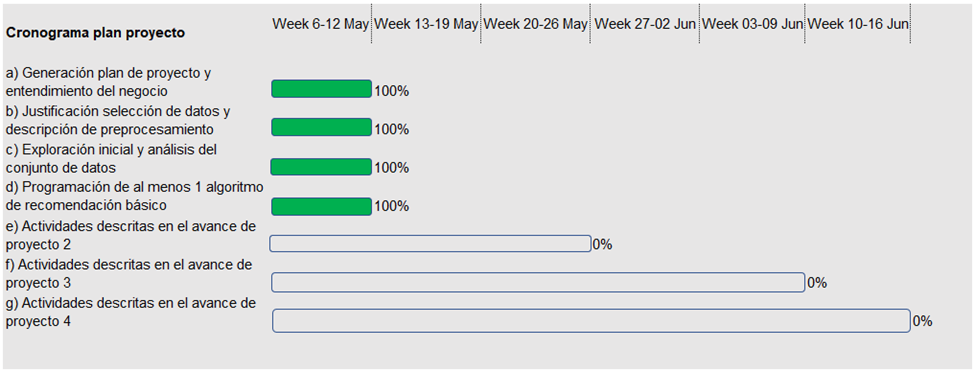
**Plan de Proyecto:**

El proyecto que estaremos desarrollando será un sistema de recomendación inteligente, utilizando modelos de aprendizaje automático supervisado, modelos de descomposición de valores singulares (SVD) y usando medidas de similitud como Similitud Euclidea, Coseno o Jaccard.. Los cuales nos ayudarán a realizar recomendaciones de peliculas.

El plan de proyecto se basará en la metodología CRISP-DM (Hotz,2024), la cual consiste en las siguientes fases:

1. Entendimiento del negocio. Qué es lo que necesita el negocio?
2. Entendimiento de los datos. Con qué datos contamos? Los datos están limpios?
3. Preparación de los datos. Cómo organizamos los datos para construir el modelo?
4. Qué técnicas de modelado aplicaremos?
5. Evaluación. Cuál es el mejor modelo que se adapta a los objetivos del negocio?
6. Cómo nuestros stakeholders acceden a los resultados?

El **cronograma** se irá actualizando con las actividades de cada entrega, una vez se tenga visibilidad de las mismas. Por el momento el cronograma al día de hoy es el siguiente:



Entendimiento del negocio:

Como antecedentes, se realizó la investigación de Netflix y como su modelo de negocio ha revolucionado la industria del entretenimiento en la era digital que vivimos actualmente.

Netflix (de acuerdo con su sitio web) fue fundada en 1997 en California, por Reed Hastings y Marc Randolph, donde se les ocurre la idea para alquilar un DVD por correo, enviándoselo a si mismos, llegando este intacto y de ahí nace la semilla de Netflix, siendo este un sitio para alquilar mediante una suscripción DVD ilimitados sin fecha de entrega o penalizaciones por retrasos.

Posteriormente en el año 2000, realizan la presentación de un sistema personalizado de recomendación de películas que usa las valoraciones de los suscriptores de títulos anteriores para predecir sus futuras elecciones, en el año 2005 lanzan la función perfiles, que permite a los suscriptores crear listas para distintos usuarios o estados de animo

Lo que requiere el negocio es contar con herramientas de toma de decisiones basadas en datos y para que el negocio de video por streaming siga siendo sostenible, tendrá que continuar su innovación en sistemas de big data para identificar patrones de consumo de peliculas y generar mayor satisfacción al cliente que se traduce en mayores ingresos y mayor rentabilidad.

Justificación de selección del conjunto de datos utilizado - Entendimiento de los datos:

Se va a utilizar una base de datos de Kaggle denominada “Them ovies dataset”, la cual ocuparemos para el proyecto. Esta Base de Datos, contiene metadatos de las 45.000 películas incluidas en el conjunto de datos completo de MovieLens. El conjunto de datos incluye películas estrenadas en julio de 2017 o antes. Los datos incluyen el reparto, el equipo, las palabras clave del argumento, el presupuesto, los ingresos, los carteles, las fechas de estreno, los idiomas, las productoras, los países, los recuentos de votos de TMDB y los promedios de votos. Este conjunto de datos también contiene archivos con 26 millones de valoraciones de 270.000 usuarios para las 45.000 películas. Las valoraciones están en una escala del 1 al 5 y se han obtenido de la página oficial de GroupLens.

Se compone de los siguientes archivos:

movies\_metadata.csv: El archivo principal de metadatos de películas. Contiene información sobre 45.000 películas incluidas en el conjunto de datos MovieLens. Incluye carteles, fondos, presupuesto, ingresos, fechas de estreno, idiomas, países de producción y empresas.

keywords.csv: Contiene las palabras clave de los argumentos de nuestras películas MovieLens. Disponible en forma de objeto JSON.

credits.csv: Contiene información sobre el reparto y el equipo de todas nuestras películas. Disponible en forma de objeto JSON encadenado.

links.csv: El archivo que contiene los ID de TMDB e IMDB de todas las películas que aparecen en el conjunto de datos de Full MovieLens.

Links\_small.csv: Contiene los ID de TMDB e IMDB de un pequeño subconjunto de 9.000 películas del conjunto de datos completo.

ratings\_small.csv: El subconjunto de 100.000 valoraciones de 700 usuarios sobre 9.000 películas.

Para este primer ejercicio, ocuparemos el archivo denominado movies\_metadata.csv, el cual contiene las siguientes columnas:

1. Adult (varchar): Es un valor binario que define si la película es +18 o no.
2. belongs\_to\_collection(varchar): Define si la pelicula pertenece a una serie de películas,
3. budget(big int): Define el presupuesto en USD para la pelicula
4. genres(varchar): Es un valor tipo lista que dicta la(s) categoría(s) que contiene la pelicula
5. homepage(varchar): Pagina web de la pelicula
6. id(int): Valor indice que asigno MovieLens
7. imdb\_id(varchar): Valor índice que relaciona esta tabla con links y links\_small
8. original\_language(varchar): Lenguaje de la película.
9. original\_title(varchar): Nombre de la película oficial
10. overview(varchar): Descripcion de la pelicula
11. popularity(float): Indice de popularidad de la pelicula
12. poster\_path(varchar): Nombre del archivo que contiene la imagen del poster de la pelicula
13. production\_companies(varchar): Valor tipo lista que contiene el nombre de las compañías que produjeron la película.
14. production\_countries(varchar): Valor tipo lista que contiene el nombre de los países donde se produjo la película.
15. release\_date(date): Fecha en la cual se realizo el estreno de la pelicula
16. revenue(big int): Valor en USD de lo recaudado por la pelicula
17. runtime : N/A
18. spoken\_languages(varchar): Valor tipo lista que contiene los idiomas que se hablan en la película:
19. status(varchar): Estatus de la pelicula
20. tagline(varchar): Eslogan de la pelicula
21. title(varchar): Nombre de la pelicula
22. video(varchar): Valor binario que dicta si la película tuvo un trailer.
23. vote\_average(float): Promedio de calificacion de la pelicula
24. vote\_count(int): Numero de calificaciones que tuvo la pelicula

**Los pasos de preprocesamiento que se seguirán, son los siguientes (realizado en jupyter-notebook):**

1. **Identificar si existen valores faltantes en el conjunto de datos.**
2. **Identificar si existe suficiente cardinalidad de las variables independientes.**
3. **Para las variables numéricas visualizar la distribución con histogramas y las correlaciones.**
4. **Se van a escalar las variables numéricas para posteriormente determinar si es viable su transformación.**
5. **Identificar si es viable la codificación en las variables categóricas.**
6. **Visualizar las gráficas de variables con las transformaciones antes aplicadas.**

**Repositorio GitHub privado por equipo con la información de los integrantes del equipo:**

**https://github.com/drekkel/TC4034.10**

**Referencias**

**Vergnou, B. (2021). Spotify Recommendation. Kaggle. Retrieved 05 07, 2024, from https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/data**

**Hotz, N. (28 de April de 2024). Data Science Process Alliance- What is CRISP DM. Recuperado el 01 de May de 2024, de** [**https://www.datascience-pm.com/crisp-dm-2/**](https://www.datascience-pm.com/crisp-dm-2/)

**Sébastien Ronteau, L. M. (2023). Beyond Digital Ubiquity: The Digital Business Model Iron Triangle. In L. M. Sébastien Ronteau, Digital Business Models: The New Value Creation and Capture Mechanisims of the 21st Century (pp. 1-10). Berlin: De Gruyter. Retrieved from https://0-eds-p-ebscohost-com.biblioteca-ils.tec.mx/eds/ebookviewer/ebook/bmxlYmtfXzM0NzI2MzRfX0FO0?sid=43ed574d-f0dc-473c-beab-6fed1e37ae23@redis&vid=1&hid=/&format=EB**