

**PROYECTO FINAL ANS – GRUPO 13.**

**Angel Benito\_Rebollo**

**Cristian Gomez**

**Esteban Tabares**

**Jose Hoyos**

**Resumen**

El mercado de servicios de streaming ha experimentado un crecimiento exponencial en los últimos años, estableciéndose como una revolución del entretenimiento “personalizado” y ofreciendo a los consumidores una amplia variedad de opciones. Sin embargo, esta proliferación de plataformas presenta un desafío significativo para los usuarios: la "sobrecarga de opciones". Este fenómeno puede llevar a la parálisis en la toma de decisiones o a la elección de servicios que no se ajustan a las necesidades individuales. Este problema queda perfectamente esbozado en palabras del nobel de economía Richard Thaler *"Demasiadas opciones pueden paralizar a los consumidores, llevarlos a tomar decisiones ineficientes, o incluso hacer que eviten tomar decisiones por completo."*

Con este contexto, surge la necesidad de una herramienta que ayude a los consumidores a navegar por este complejo panorama de opciones de streaming. El objetivo principal de nuestro trabajo es desarrollar una guía integral y accesible que permita a los usuarios evaluar de manera precisa y fundamentada las principales plataformas de streaming en función de factores clave como el costo de suscripción y la calidad del contenido en cuanto al puntaje de los usuarios, valoración en sitios especializados y número de votos. Para lograr esto, se recopilarán y analizarán datos de reseñas de usuarios, puntuaciones de películas y popularidad en sitios como IMDB y TMDB.

Esta guía no solo ayudará a los consumidores a tomar decisiones informadas, sino que también proporcionará información valiosa a los proveedores de servicios de streaming, destacando áreas de mejora y oportunidades de crecimiento en un mercado altamente competitivo.

En resumen, nuestro trabajo pretende ser una herramienta esencial para maximizar el valor de la inversión de las personas en servicios de streaming, proporcionando una evaluación equilibrada y rigurosa de las opciones disponibles en el mercado.

**Introducción**

¿cómo determinar cuál servicio ofrece la mejor relación calidad-precio en función de las necesidades y preferencias individuales? La vasta oferta de contenido, combinada con distintas estructuras de precios y diversidad de títulos, puede resultar abrumadora para los usuarios al momento de tomar una decisión informada y es nuestro deseo poder ayudar a los millones de usuarios de estas plataformas en la toma de una decisión que los satisfaga.

El cliente potencial de esta propuesta son los millones de usuarios de plataformas de streaming que desean maximizar el valor de su suscripción, obteniendo el mejor balance entre costo, calidad y variedad de contenido. A nivel organizacional, las empresas que operan en este espacio también pueden beneficiarse de este análisis, ya que proporcionará información valiosa sobre cómo se perciben sus servicios en comparación con la competencia y qué áreas podrían mejorar para captar más usuarios.

El problema que buscamos resolver pertenece al área de análisis no supervisado (ANS) dentro del aprendizaje automático. Específicamente, utilizaremos técnicas de clustering y reducción de dimensionalidad para agrupar y analizar las plataformas de streaming según factores críticos. Estas técnicas esperamos nos permitan descubrir patrones ocultos y relaciones subyacentes entre las distintas plataformas, ofreciendo una evaluación objetiva y profunda que guiará tanto a consumidores como a proveedores en la toma de decisiones más informadas y estratégicas.

**Materiales y Métodos:**

**Datos Utilizados:**

Para este proyecto, se utilizaron dos conjuntos de datos principales:

1. **Catálogo de Streaming**:
   * **Fuente**: Archivo Excel catalogo\_streaming.xlsx.
   * **Descripción**: Contiene información detallada sobre películas y series disponibles en diversas plataformas digitales.
   * **Variables**:
     + title: Título de la película o serie.
     + platform: Plataforma digital (Netflix, Amazon Prime, etc.).
     + type: Tipo de contenido (IMAGE, SHOW).
     + imdbVotes: Número de votos en IMDB.
     + imdbScore: Puntuación media en IMDB.
     + tmdbPopularity: Puntuación de popularidad en TMDB.
     + tmdbScore: Puntuación media en TMDB.
     + runtime: Duración en minutos.
     + genres: Géneros de la película.
     + director: Director de la película.
     + year\_range: Rango de años de lanzamiento.
     + url: URL de la página de la película en la plataforma.
2. **Precios de Streaming**:
   * **Fuente**: Archivo Excel Precios Streaming COP.xlsx.
   * **Descripción**: Contiene información sobre los precios de suscripción de las plataformas digitales en pesos colombianos (COP).

**Pantalla de computadora con letras

Descripción generada automáticamente con confianza media**

Figura 1. Estadísticas descriptivas variables “Object”.

Pantalla de computadora con letras

Descripción generada automáticamente con confianza media

Figura 2. Estadísticas descriptivas variables numéricas.

**Proceso de Limpieza de Datos:**

El proceso de limpieza de datos incluyó los siguientes pasos:

1. **Transformación de Géneros**: Se transformó la columna de géneros de un string a un array y se expandió con explode() para que cada género tuviera su propia fila. Esto facilita el análisis y la clusterización basada en géneros específicos, permitiendo también evaluar la relación calidad-precio de cada plataforma por género.
2. **Eliminación de Columnas Irrelevantes**: Se eliminaron las columnas year\_range, url y runtime con drop() porque esta variable tenía problemas de typo y no era relevante para el estudio, mientras que url y runtime no aportaban información útil para la clusterización basada en las calificaciones y popularidad de las películas. Esto simplificó el conjunto de datos y redujo el ruido.
3. **Escalado de Datos**: Se escalaron con StandarScaler() las columnas numéricas (imdbVotes, imdbScore, tmdbPopularity, tmdbScore) para que todas tuvieran una media de 0 y una desviación estándar de 1. Esto es crucial para la clusterización, ya que asegura que todas las variables contribuyan de manera equitativa al cálculo de distancias. Las votaciones (imdbVotes) fueron especialmente importantes, ya que una película con muchas votaciones tiene un peso diferente a una con pocas, lo que puede cambiar los resultados de K-means. Solo se escalaron las variables numéricas porque las categóricas no son adecuadas para el cálculo de distancias en K-means.

**Algoritmo Utilizado:**

Se utilizó el algoritmo de K-means para la clusterización de las películas. Los pasos fueron los siguientes:

1. **Reducción de Dimensionalidad**: Se utilizó PCA para reducir las dimensiones a 3 componentes principales. Esto ayuda a simplificar el problema y a visualizar mejor los clusters, ya que la reducción de dimensionalidad permite observar las relaciones entre las variables de manera más clara y detectar patrones que podrían no ser evidentes en un espacio de mayor dimensión.
2. **Clusterización con K-means**: Se aplicó el algoritmo K-means para agrupar las películas en 2 clusters, con el objetivo de identificar el cluster con las mejores puntuaciones en las variables numéricas (imdbVotes, imdbScore, tmdbPopularity, tmdbScore). Estas variables se seleccionaron porque representan la popularidad y la calidad de las películas según las calificaciones de los usuarios en IMDB y TMDB. Esto es crucial para K-means y nuestro objetivo, ya que queremos identificar grupos de películas altamente valoradas y populares, lo que nos permitirá evaluar la calidad de las plataformas de streaming en función de estas métricas.
3. **Cálculo de Distancias Euclidianas**: Se calculó la distancia euclidiana de cada película al centroide del mejor cluster, lo que permite evaluar qué tan cerca está cada película del grupo de películas mejor valoradas.
4. **Evaluación de Plataformas**: Se evaluaron las plataformas según la distancia promedio al mejor cluster y se normalizaron los precios, lo que permite identificar la plataforma con la mejor relación calidad-precio.

**Resultados y Discusión:**

**Implementación del Algoritmo:**

Para la implementación del algoritmo, se utilizó PCA para reducir las dimensiones a 3 componentes principales, ya que estas explican más del 90% de la varianza. Esto simplificó el problema y permitió visualizar mejor los clusters. Posteriormente, se aplicó el algoritmo K-means para agrupar las películas en 2 clusters. Se probó con varios números de clusters y se decidió utilizar 2 porque se agrupan bien entre las mejores películas basándonos en las variables numéricas (imdbVotes, imdbScore, tmdbPopularity, tmdbScore) y las que no.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figura 3. Clusteres con K-means variables numéricas.

Se tomaron los precios básicos de las suscripciones en COP y se utilizó la siguiente fórmula para calcular la relación calidad-precio:

Nuestra variable resultado es quality\_price\_ratio, donde un **puntaje menor** indica una mejor relación calidad-precio. Los resultados mostraron que la mejor plataforma generalmente es Netflix con una calidad-precio de -0.789, seguida por Paramount Plus con -0.416, Amazon con -0.158, Max con 0, Apple con 0.669 y Disney última con 0.939.

Captura de pantalla de computadora

Descripción generada automáticamente con confianza media

Figura 4. Variables calculadas y ranking de plataformas.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figura 5. Gráfica relación Distancia Euclideanda y Calidad-Precio.

**Comparación con Algoritmos Similares:**

Se comparó K-means con DBSCAN, pero se determinó que K-means era más adecuado debido a su interpretabilidad y porque sabíamos nuestro cluster objetivo (películas con mayores puntajes). DBSCAN no fue útil en este caso debido a la naturaleza de los datos y la necesidad de identificar clusters específicos de alta calidad.

**Justificación del Algoritmo Elegido:**

K-means fue elegido porque permite identificar claramente grupos de películas altamente valoradas y populares, lo que es crucial para evaluar la calidad de las plataformas de streaming. Además, es un algoritmo eficiente y fácil de interpretar, lo que facilita la toma de decisiones basada en los resultados.

**Limitaciones:**

Las principales limitaciones de esta implementación incluyen:

* **Acceso a Datos**: Las restricciones regionales pueden afectar la disponibilidad de datos.
* **Moneda**: Solo se consideró el costo en COP, lo que puede no ser representativo a nivel global.
* **Tiempo**: El proyecto se realizó en 4 semanas, lo que es un tiempo limitado para un análisis exhaustivo.

**Análisis y Estudios Posteriores:**

En el futuro, podrían desarrollarse estudios que abarquen más regiones y monedas para averiguar las mejores plataformas en términos de calidad-precio. Además, se podría profundizar en cuáles plataformas son mejores respecto a tipo o género de contenido, proporcionando una visión más detallada y específica.

**Referencias:**

* [IMDB](https://www.imdb.com)
* [TMDB](https://www.themoviedb.org)
* [Netflix tech blog](https://netflixtechblog.com/)
* [Estudio Percepcion Consumidor Servicio Streaming](https://repositorio.usm.cl/server/api/core/bitstreams/d67c307d-1a49-4c90-b54d-335457c3938d/content)
* [Comparativa Streaming](https://www.xataka.com/basics/comparativa-disney-netflix-hbo-movistar-prime-video-apple-filmin-catalogo-funciones-precios)
* [Recomendaciones de Netflix](https://netflixtechblog.com/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429)