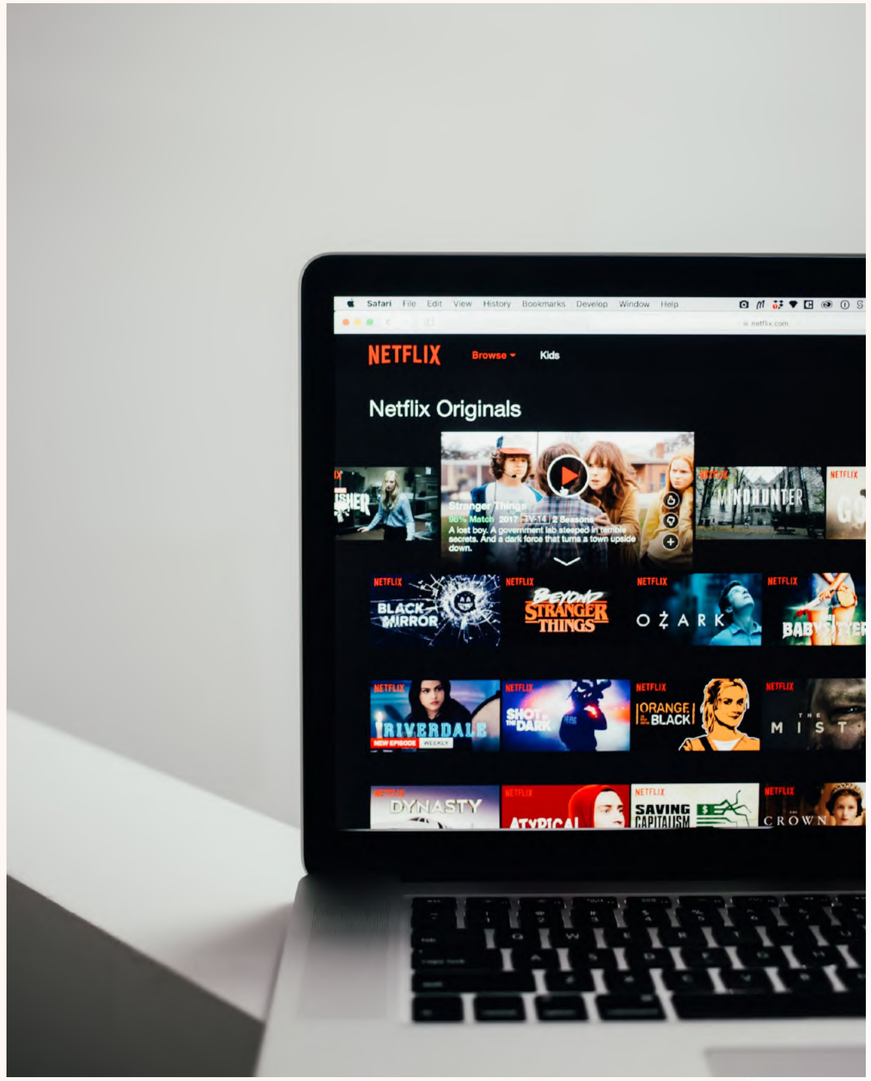
Data Science: Películas

Equipo 12:

Gabriela Miranda

María López de Armentia

Julio 2025



Objetivo

Analizar el catálogo de Peliculas para entender patrones de distribución de contenido por tipo, género, país, y evolución en el tiempo, con el fin de obtener insights que puedan apoyar decisiones editoriales, de marketing o recomendaciones

Netflix - TV Shows & Movies

01 Elección de dataset

02 Exploración y Transformación de datos

03 Visualización de datos

Dataset

Dataset: Análisis Exploratorio de Datos

Análisis de Features

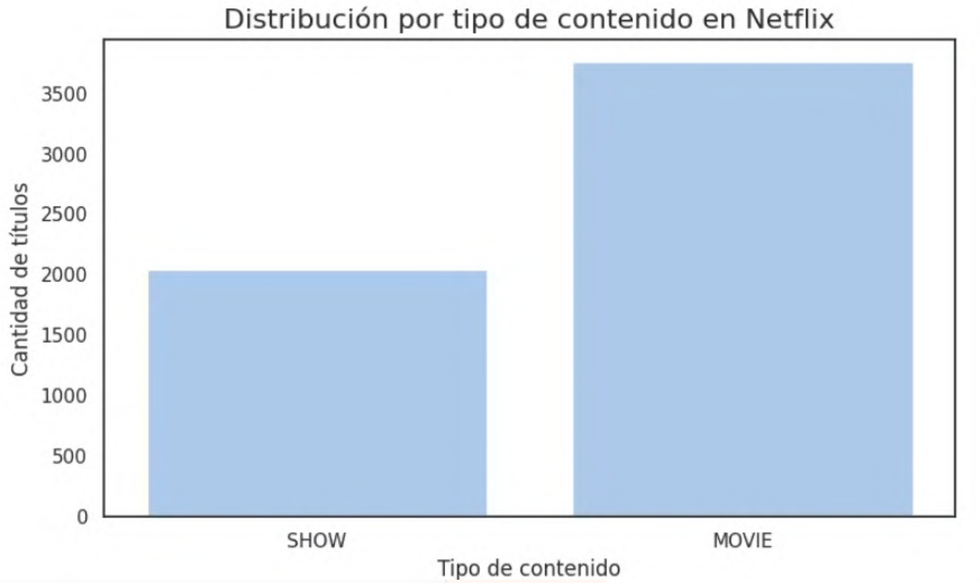
- id: Identificador único del título en el dataset.
- title: Nombre del título (película o serie).
- type: Tipo de contenido: MOVIE o SHOW.
- description: Breve sinopsis del título.
- release_year Año en que fue lanzado originalmente.
- age_certification: Clasificación por edad (e.g. PG, R, TV-MA). Puede estar vacía en algunos casos.
- runtime Duración en minutos (para películas) o promedio por episodio (para series).
- genres Lista de géneros asociados al contenido (formato tipo lista Python).
- production_countries País o países de origen de la producción (también en formato lista).
- seasons Número de temporadas (solo aplicable a series, NaN en películas). imdb_id ID del título en IMDb (útil para hacer joins externos si fuera necesario).
- imdb_score Calificación promedio en IMDb (escala de 0 a 10).
- imdb_votes Número total de votos recibidos en IMDb.
- tmdb_popularity Índice de popularidad según TMDb (valor continuo). Métrica de visibilidad
- tmdb_score Calificación promedio en TMDb (escala de 0 a 10).

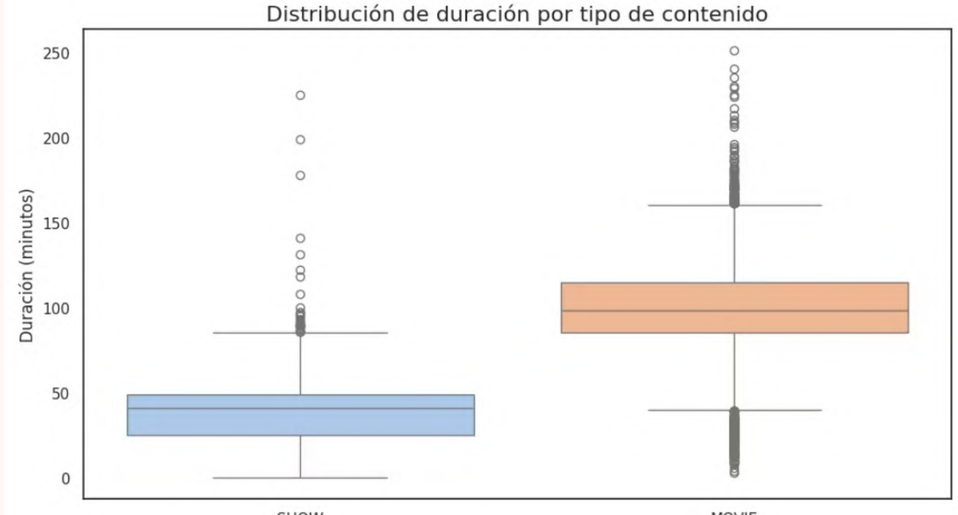
15 columnas - 5806 registros

Limpieza de valores nulos

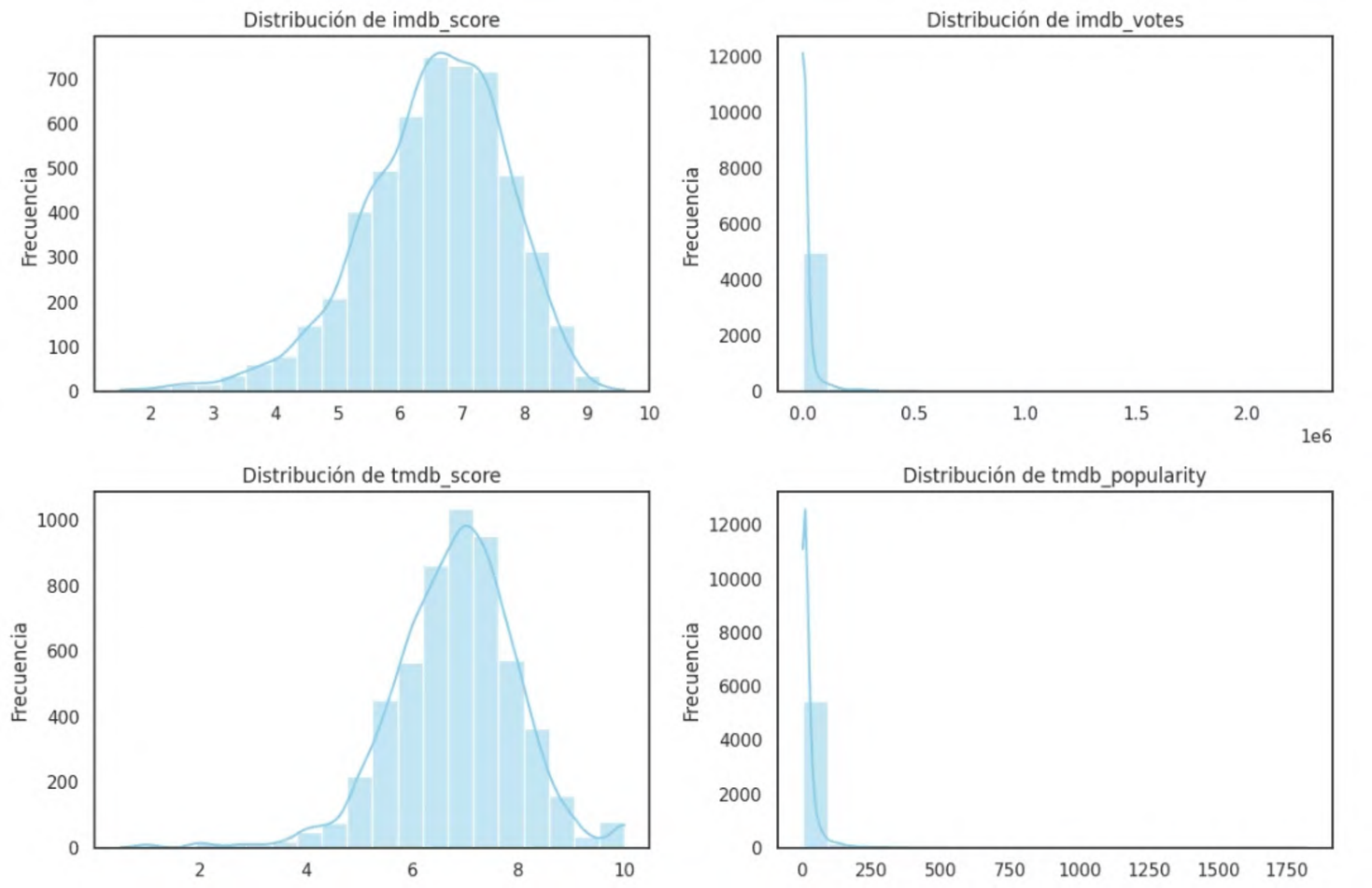
	nulos	porcentaje
id	0	0.00
title	1	0.02
type	0	0.00
description	18	0.31
release_year	0	0.00
age_certification	2610	44.95
runtime	0	0.00
genres	0	0.00
production_countries	0	0.00
seasons	3759	64.74
imdb_id	444	7.65
imdb_score	523	9.01
imdb_votes	539	9.28
tmdb_popularity	94	1.62
tmdb_score	318	5.48

Dataset: Análisis Exploratorio de Datos

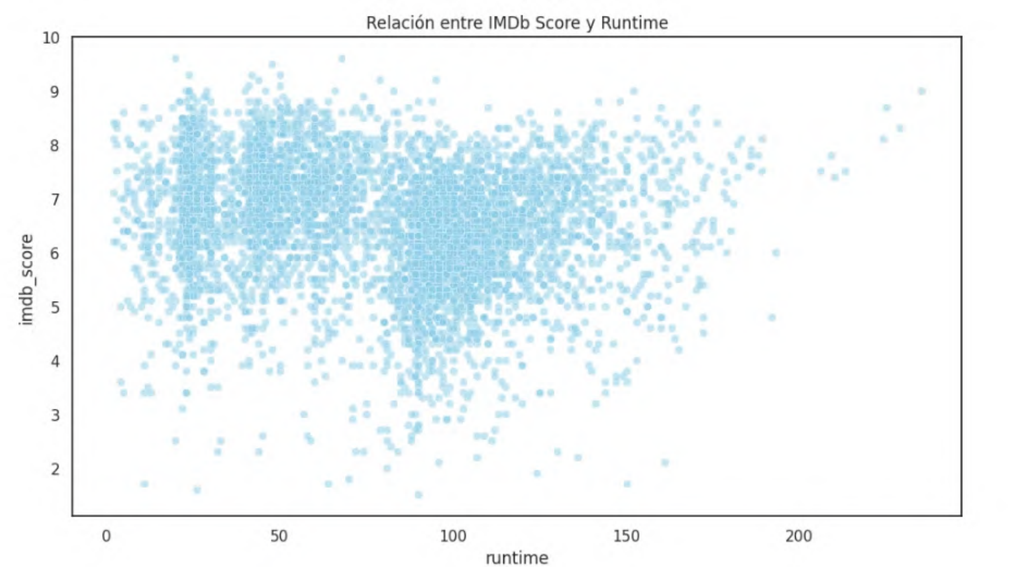


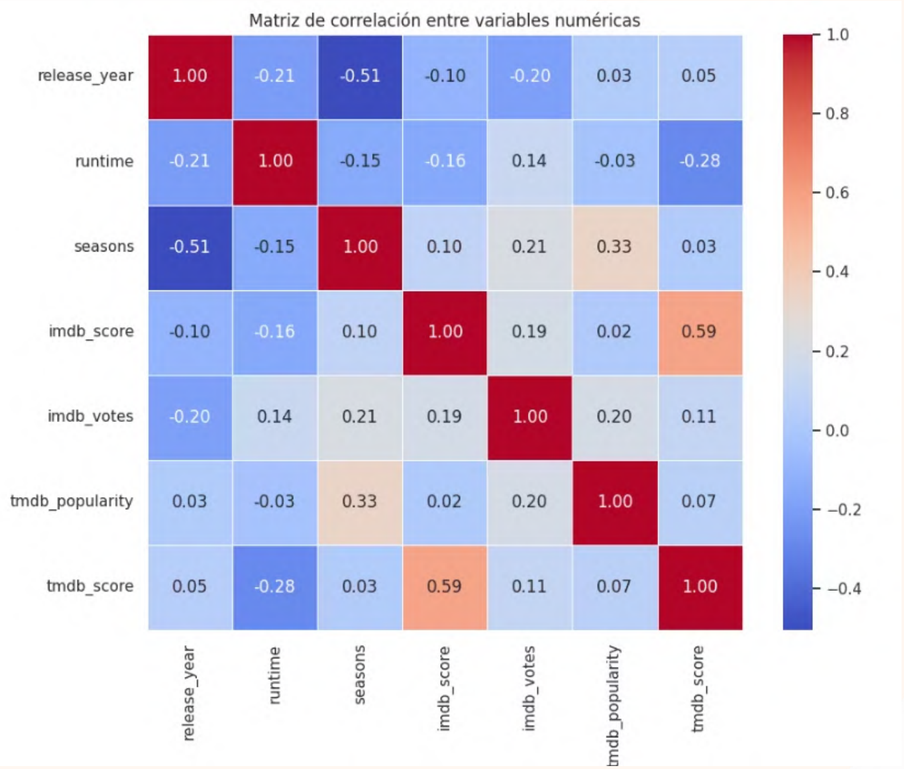


Dataset: Análisis Exploratorio de Datos



Dataset: Análisis Exploratorio de Datos





Dataset: Nulos / Outliers

Temporadas	65% valores nulos se corresponden a películas
Duración	Películas >200 min= OK Películas =0, reemplazo por mediana Cortometrajes / miniseries / Documentales especiales= OK (única temporada)
Calificación / Votos / Popularidad	Sólo se eliminan los que todos los valores son nulos, si sólo falta un dato completamos con la mediana

Aprendizaje Supervisado

Modelo de clasificación

Objetivo

Entrenar un modelo de clasificación para predecir si un título está "bien valorado" según su puntuación IMDb (variable continua que convertimos en clases):

- Baja calidad (<6.0)=0
- Media calidad (6.0–7.9)=1
- Alta calidad (≥8.0)=2

Normalización y estandarización

- Eliminación features irrelevantes
- One hot encoding (género, países)

	release_year	runtime	seasons	imdb_score	imdb_votes	tmdb_popularity	tmdb_score	type_enc	age_certification_ord	action	animation	comedy	crime	documentation	drama	european	family	fantasy	history	horror	music	reality	romance	scifi
1	1976	113	0.0	8.3	795222.0	27.612	8.2	0	8	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
2	1975	91	0.0	8.2	530877.0	18.216	7.8	0	5	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
3	1979	94	0.0	8.0	392419.0	17.505	7.8	0	8	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	1973	133	0.0	8.1	391942.0	95.337	7.7	0	8	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
5	1969	30	4.0	8.8	72895.0	12.919	8.3	1	7	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0

Modelo de clasificación

DEFINICIÓN DE MODELO

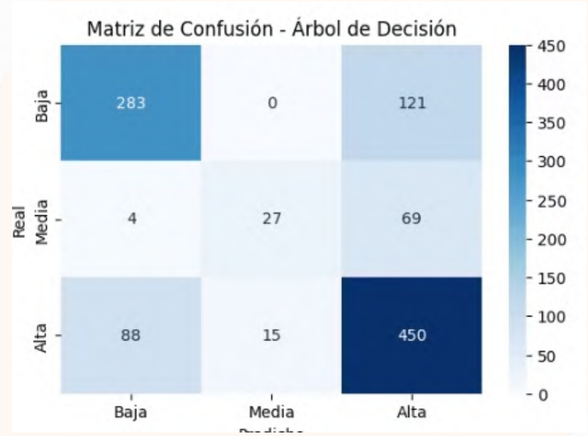
Variable objetivo: imbd score. → transforma en variable categórica (baja, media y alta)
Variables predictoras: todas las demás

Separamos el conjunto en entrenamiento y prueba

ARBOL DE DECISIÓN

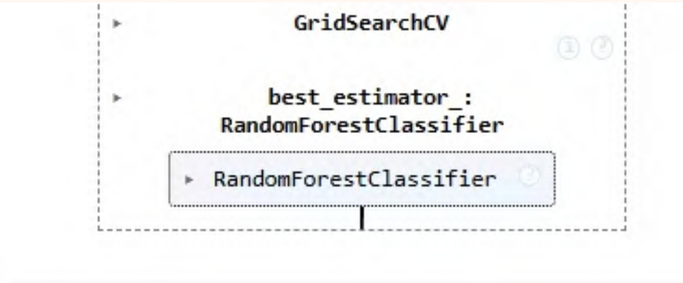
```
Accuracy: 0.7190160832544938
Reporte de Clasificación:
              precision    recall  f1-score   support

           0       0.75      0.70      0.73       404
           1       0.64      0.27      0.38       100
           2       0.70      0.81      0.75       553

    accuracy                           0.72      1057
   macro avg       0.70      0.59      0.62      1057
weighted avg       0.72      0.72      0.71      1057
```

MATRIZ DE CONFUSIÓN



Modelo de clasificación

RANDOM FOREST



Accuracy en test: 0.7568590350047304
Reporte:

	precision	recall	f1-score	support
0	0.78	0.78	0.78	404
1	0.69	0.29	0.41	100
2	0.75	0.82	0.78	553
accuracy			0.76	1057
macro avg	0.74	0.63	0.66	1057
weighted avg	0.75	0.76	0.75	1057

Aprendizaje No Supervisado

Segmentación de contenido

Objetivo

Se busca aplicar técnicas de clustering para segmentar el catálogo de películas y series en función de sus características con el fin de descubrir patrones latentes que permitan entender mejor la estructura del contenido disponible y facilitar su análisis o recomendación.

SEGMENTAR CONTENIDO SEGUN CARACT. TECNICAS

Agrupar obras similares en duración, temporadas, puntajes y votos.

DETECTAR PERFILES POR GÉNERIO Y DISTRIBUCIÓN

Analizar influencia de géneros y disponibilidad por región

DESCUBRIR PATRONES EN VALORACIÓN

Explorar agrupamientos basados en popularidad y puntuaciones

IDENTIFICAR CLUSTERS ATÍPICOS

Encontrar grupos pequeños que no sigan los patrones principales

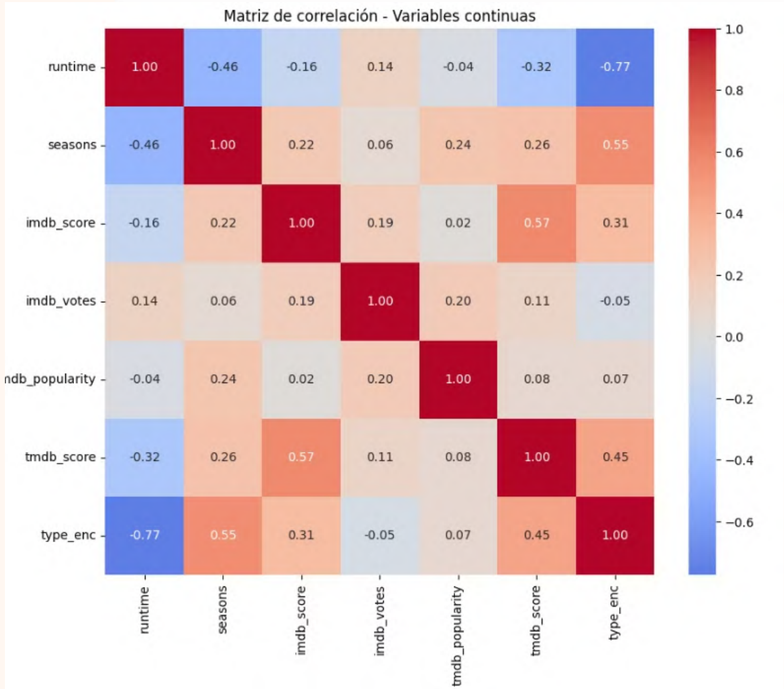
SIMPLIFICAR CATÁLOGO DE RECOMENDACIONES

Reducir recomendaciones según clusters más representativos

Procesamiento de datos

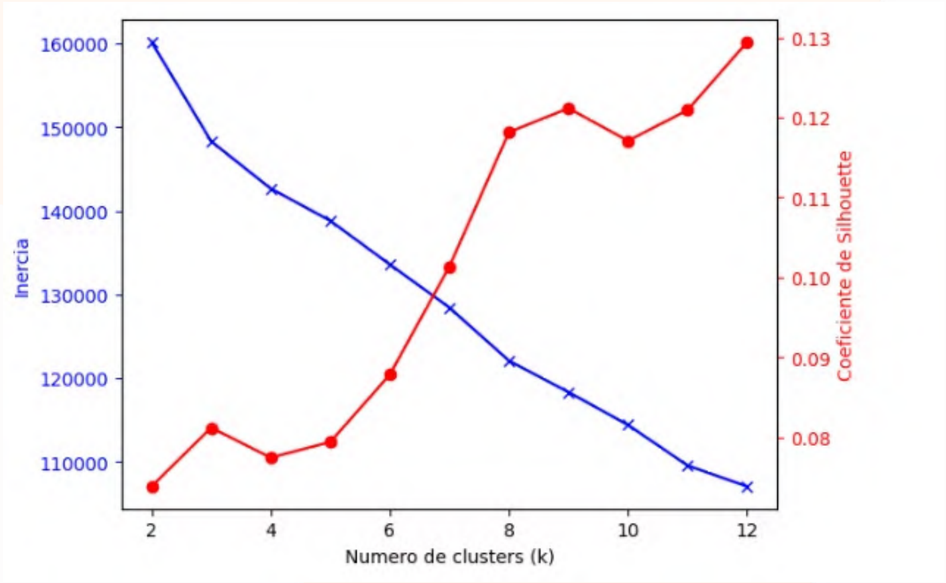
- Se mantienen las variables continuas: Runtime, seasons, imdb_score, imdb_votes, tmdb_score, tmdb_popularity. Variables binarias: action, comedy, drana y todas las pertenecientes a genero se mantenien.
- Las variables referidas a regiones se agrupan para simplificar el modelo.
- Se eliminan las features de menor importancia
- Se escalan los datos con Standard Scaler

MATRIZ DE CORRELACIÓN

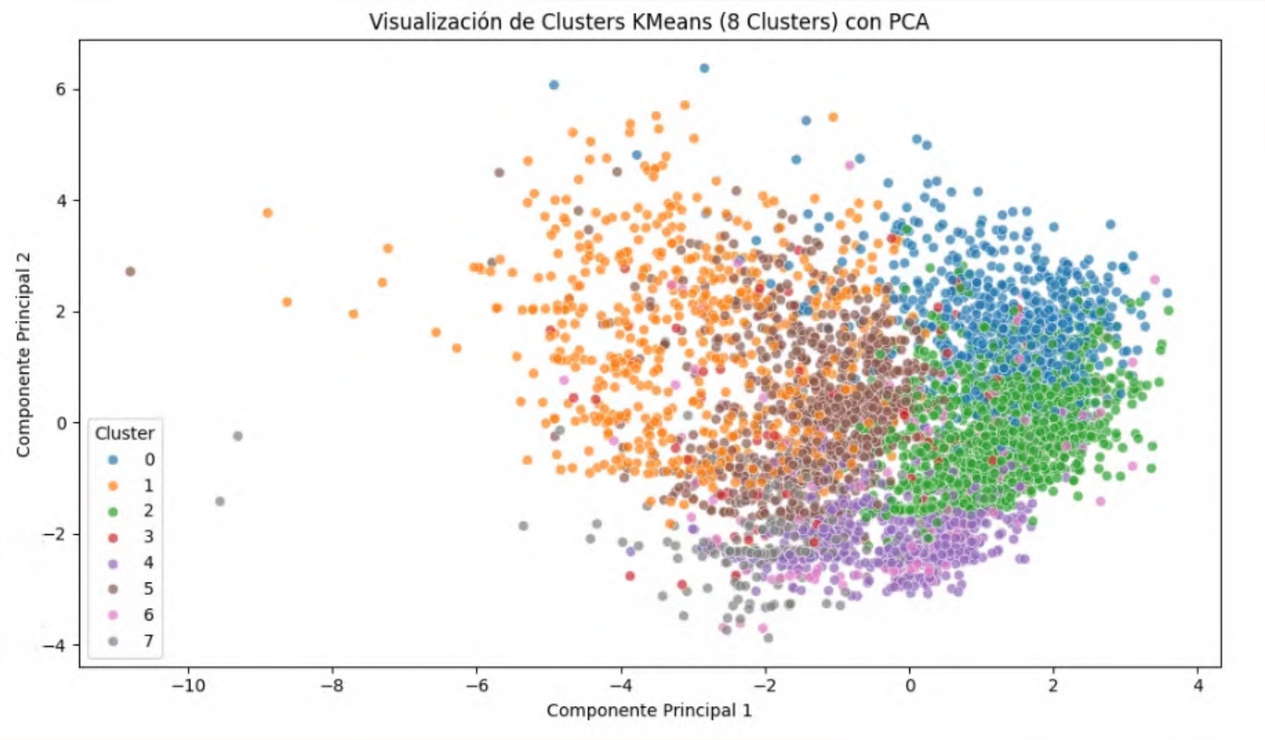


K- MEANS

MÉTODO DEL CODO- COEFICIENTE DE SILHOUETTE



PCA (K=8)



K- MEANS

CONCLUSIÓN

Se obtuvieron 8 clusters con el algoritmo Kmeans con las siguientes características:

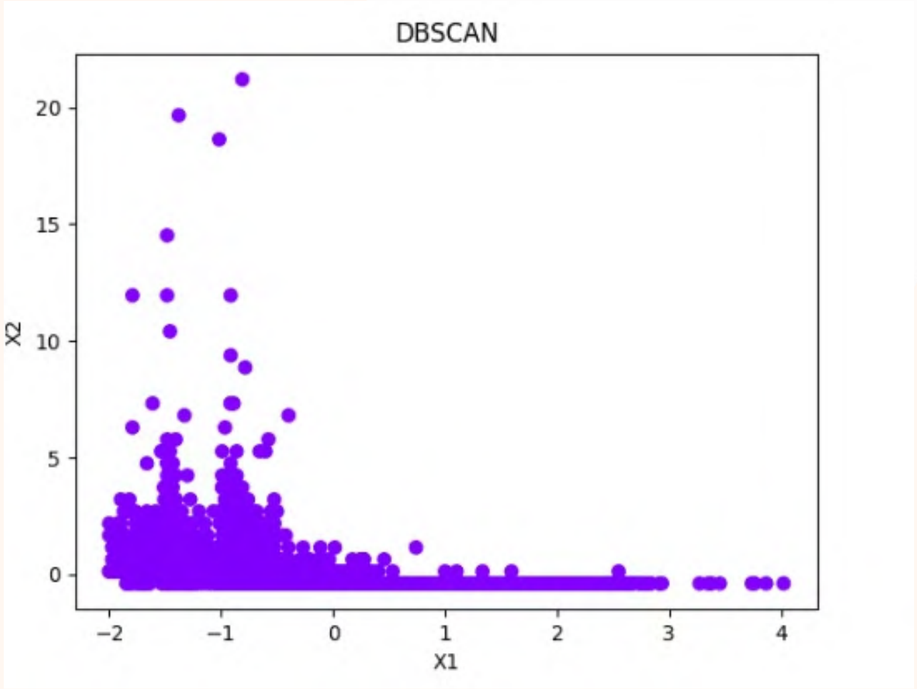
- **Cluster 0**: Películas europeas de mediana-larga duración (109 min) con fuerte presencia de drama (66%) y thriller, escasa animación o comedia, y una puntuación IMDb media (6.03); representan un cine más serio e introspectivo, popular especialmente en Europa.
- **Cluster 1**: Series animadas de tipo familiar y fantástico con casi 2 temporadas promedio, alta presencia en América (51%) y géneros como animation (90%), family (58%) y fantasy, ideal para público infantil o general, con buena puntuación (6.65 IMDb) y enfoque de entretenimiento ligero.
- **Cluster 2**: Películas románticas y dramáticas distribuidas globalmente (Europa, América y Asia), de duración media (103 min), con participación destacada de los géneros drama (60%), romance (32%) y comedia; tienen bajo nivel de votos y popularidad, pero mantienen una buena puntuación IMDb (6.16).
- **Cluster 3**: Series variadas con duración media (68 min), centradas en drama, crime y thriller, con algo de horror y producción 100% de Oceanía, lo que sugiere una oferta regional con tono serio y narrativas policiales; puntuación alta (6.69 IMDb) y más de una temporada promedio.
- **Cluster 4**: Documentales (99%) de duración media (71 min), producidos principalmente en América (74%), con muy buena calificación IMDb (7.07) y baja cantidad de temporadas (0.45); representan contenido educativo o informativo puro, con escasa mezcla de géneros.
- **Cluster 5**: Series cortas (2.1 temporadas) con muy alto puntaje IMDb (7.22), orientadas a drama (80%), romance, y con presencia importante de Europa y Asia; sus historias son profundas y humanas, con buena recepción y enfoque más emocional que comercial.
- **Cluster 6**: Películas globales de tono diverso, con foco en drama (52%), comedia y algo de horror y thriller, con duración media (82 min), nivel de votación alto (~26k) y producción mayoritariamente americana (57%); tienen una buena puntuación (6.74 IMDb) y abarcan varias emociones y géneros.
- **Cluster 7**: Reality shows (100%) con más de 2 temporadas, producción principalmente americana (67%), duración corta (~42 min), popularidad baja pero puntuación estable (6.40 IMDb); representan un formato bien definido de entretenimiento televisivo basado en lo real.

DBSCAN

Permite identificar clusters de forma arbitraria y detectar outliers sin necesidad de fijar K.

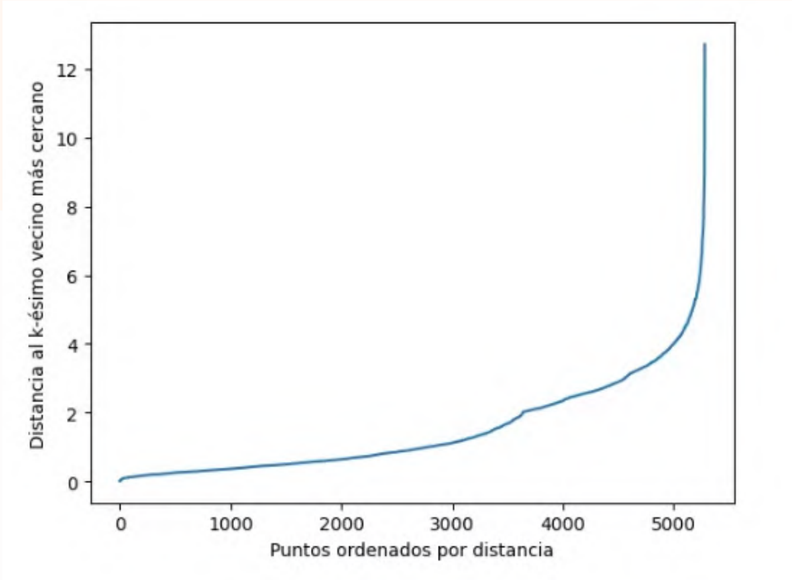
- Se ajustan eps y min_samples.
- Se compara su performance usando Silhouette Score.
- Se grafican los resultados para observar outliers.

RESULTADO INICIAL: 1 CLUSTER Y 5195 PUNTOS DE RUIDO

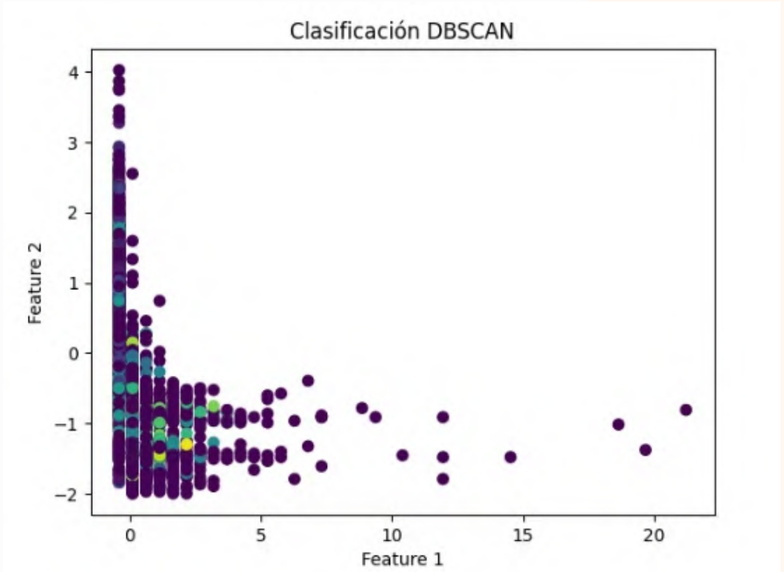


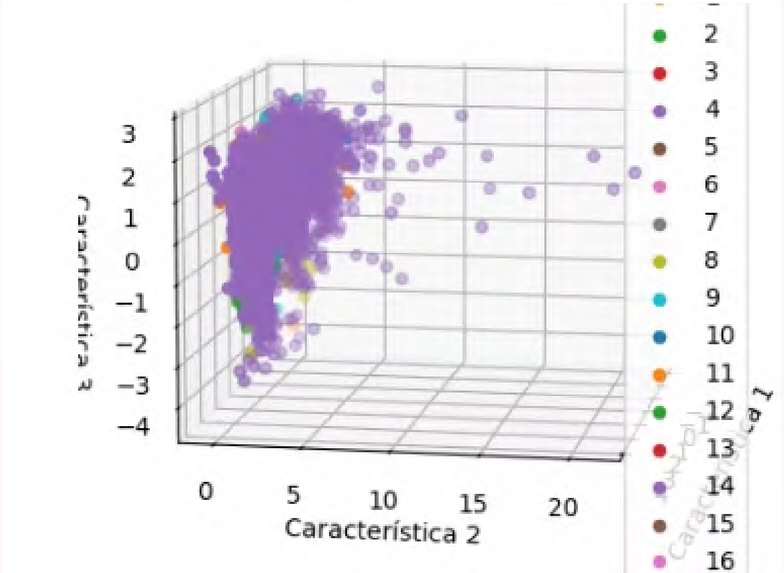
DBSCAN

METODO DEL VECINO MAS CERCANO (EPS=2)



RESULTADO CON EPS=2: 64 CLUSTERS Y 3335 PUNTOS DE RUIDO





DBSCAN

CONCLUSIÓN

Se corrieron dos versiones:

- Una con eps = 0.3, que arrojó 1 clusters pero con demasiados puntos de ruido (5195).
- Otra con eps = 2.0, que detectó 64 clusters y redujo los puntos de ruido a 3335.

Ambos resultados muestran que DBSCAN logró detectar agrupamientos sin necesidad de definir cuántos clusters había previamente. Sin embargo, el alto nivel de ruido indica que el dataset tiene muchas observaciones atípicas o que el espacio de características es muy disperso, por lo cual la información obtenida no es de mucha utilidad.

Muchas gracias!!