SPOTIFY PLAYLIST DATASET CHALLENGE

Por: Paula Velosa- Fabián López
Grupo 4 - Mineria de Datos 2025-1

MOTIVACIÓN

¿Por qué importa?

Playlists = huellas de gustos colectivos

MOTIVACIÓN

¿Por qué importa?

Mejor comprensión ⇒ **recomendaciones más personalizadas.**

Descripción del Conjunto de Datos

DIMENSIONALIDAD

Descripción del conjunto

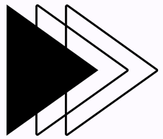
- **Playlists**: 10.000 registros.
- **Columnas**: 6 variables por Playlist - 8 variables por canción.
- **Pistas visibles**: Hasta 100 por playlist.
- **Formato anidado**: Cada playlist contiene una lista de canciones.

EL RETO DE LOS DATOS

EDA

10 000 playlists parciales (2010-2017).

6 variables por playlist + hasta 100 canciones.



ENRIQUECIMIENTO MULTIFUENTE

+ Spotify API, MusicBrainz, AcousticBrainz.

83 variables acústicas y semánticas finales

DESCRIPCIÓN DEL CONJUNTO

EDA

1. 281000 datos repartidos en 9000 grupos

2. 124 dimensiones, algunas vacías

3. En general los datos tienen bpm medio, una energía baja, una bailabilidad baja y alto volúmen

DESCRIPCIÓN DE VARIABLES

Nivel Playlist

Variable	Tipo	Descripción
pid	Entero	ID de la playlist
name	Texto	Nombre de la playlist (opcional)
num_holdouts	Entero	Canciones ocultas
num_samples	Entero	Canciones visibles
num_tracks	Entero	Total de canciones
tracks	Lista	Lista de canciones visibles

DESCRIPCIÓN DE VARIABLES

Nivel Canción (MUSICBRAINZ)

Variable	Tipo	Rango estimado	Descripción
pos	Entero	0 - 99	Posición en la playlist
track_name	Texto	---	Nombre de la canción
track_uri	Texto	22 caracteres	URI único en Spotify
artist_name	Texto	---	Artista principal
artist_uri	Texto	22 caracteres	URI del artista
album_name	Texto	---	Nombre del álbum
album_uri	Texto	22 caracteres	URI del álbum
duration_ms	Entero	30.000 - 600.000 ms	Duración de la canción

DESCRIPCIÓN DE VARIABLES

Nivel Canción (ACOUSTICBRAINZ, HIGH LEVEL), algunas

Variable	Tipo	Rango	Descripción
danceability	Categórico| Probabilistico	danceable, not_danceable	Indica si la pista resulta apta para bailar, basándose en ritmo, pulso y regularidad del compás.
gender	Categórico| Probabilistico	female, male	Género percibido de las voces principales en la grabación (femenina o masculina)
mood_happy	Categórico| Probabilistico	happy, not_happy	Evalúa la valencia emocional de la pista, si transmite una sensación alegre.
mood_party	Categórico| Probabilistico	mood_party	Categórico| Probabilistico

OBJETIVOS DE MINERIA

1. Filtrar playlists válidas
2. Comparar y agrupar por similitud
3. Revelar temáticas y moods

PREPARACIÓN DE LOS DATOS

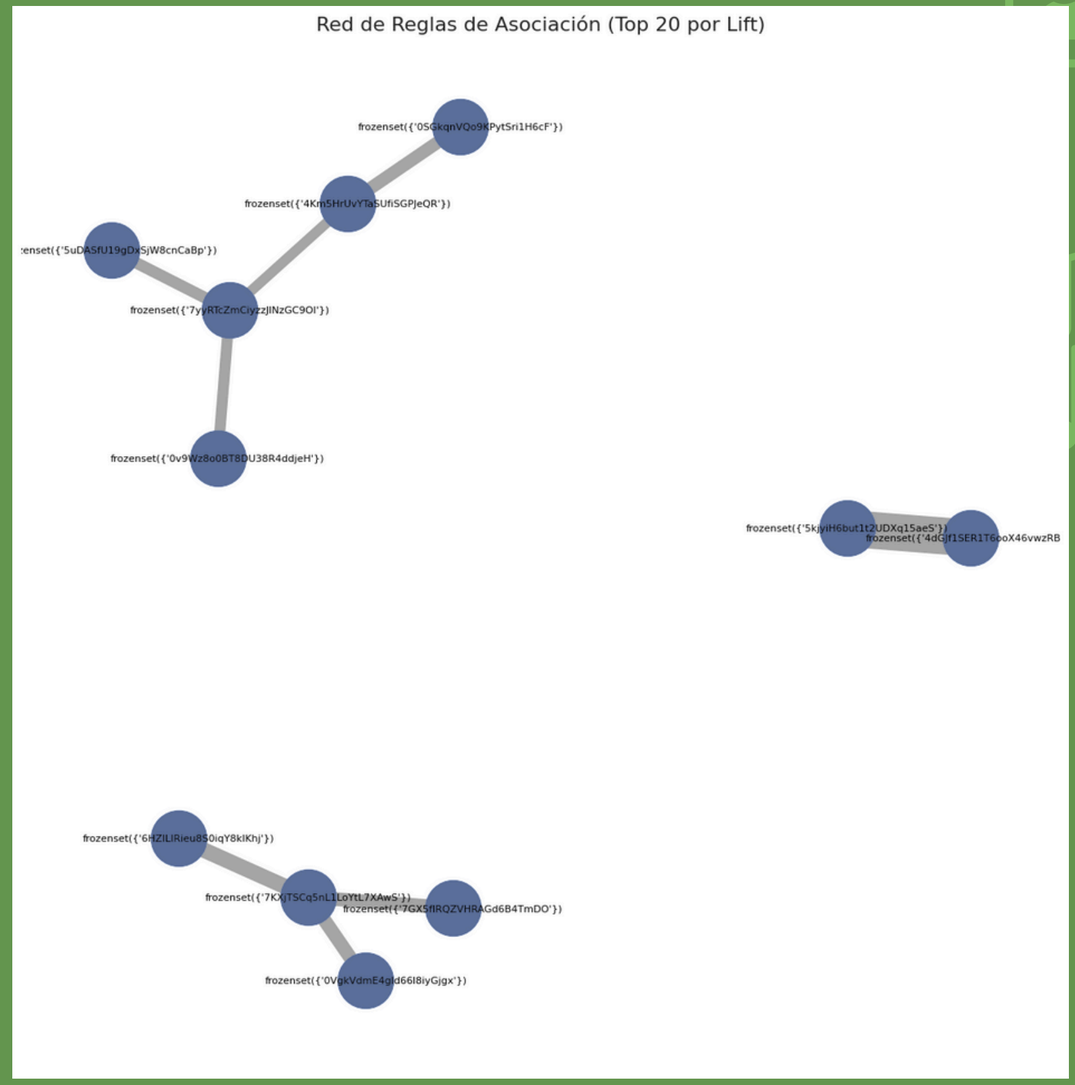
1. Escalado, winsorizing, imputación media/mediana.
2. Se imputaron los datos vacíos como “Desconocidos”
3. (Para clustering) 83 vars. → PCA + UMAP (-90 % var.).
4. Matriz transacciones para reglas asociación

REGLAS DE ASOCIACIÓN

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(7yyRTcZmCiyzzJlNzGC9Ol)	(0v9Wz8o0BT8DU38R4ddjeH)	0.025556	0.02037	0.009259	0.362319	17.786561
(7KXjTSCq5nL1LoYtL7XAwS)	(0VgkVdmE4gld66l8iyGjgx)	0.022593	0.017778	0.008889	0.393443	22.131148
(5kjyiH6but1t2UDXq15aeS)	(4dGJf1SER1T6ooX46vwzRB)	0.010741	0.013333	0.008519	0.793103	59.482759
(7KXjTSCq5nL1LoYtL7XAwS)	(6HZILIRieu8S0iqY8kIKhj)	0.022593	0.014444	0.008148	0.360656	24.968474
(7yyRTcZmCiyzzJlNzGC9Ol)	(5uDASfU19gDxSjW8cnCaBp)	0.025556	0.017037	0.008148	0.318841	18.714556

820 reglas filtradas (soporte ≥ 1 %), alta confianza.

REGLAS DE ASOCIACIÓN



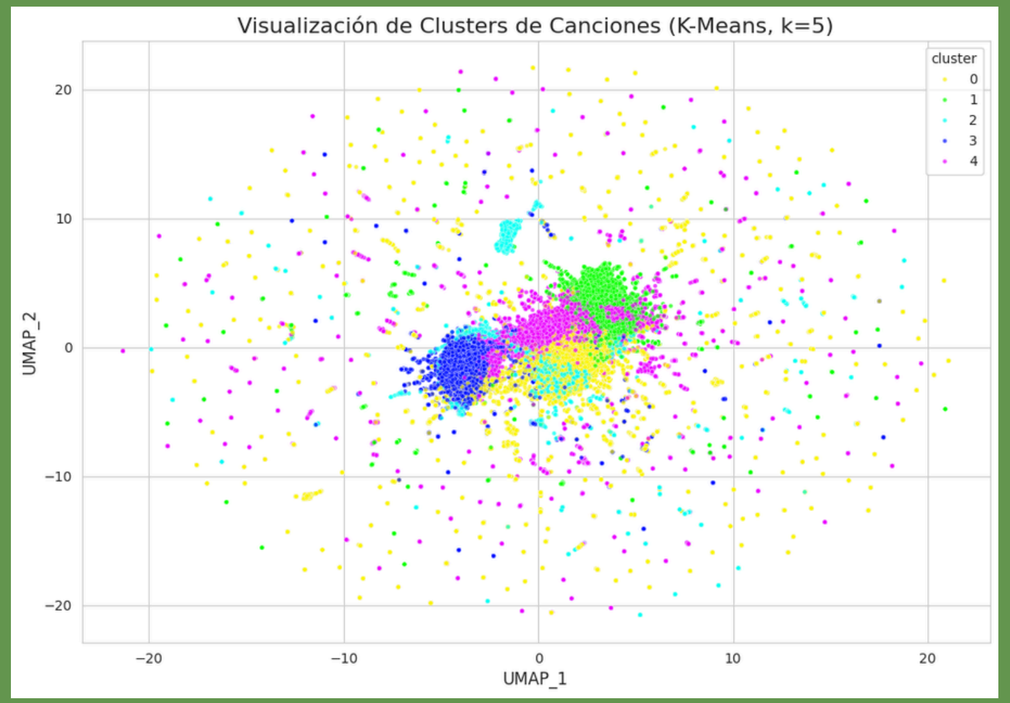
CLUSTERING

Tabla de comparativa de modelos

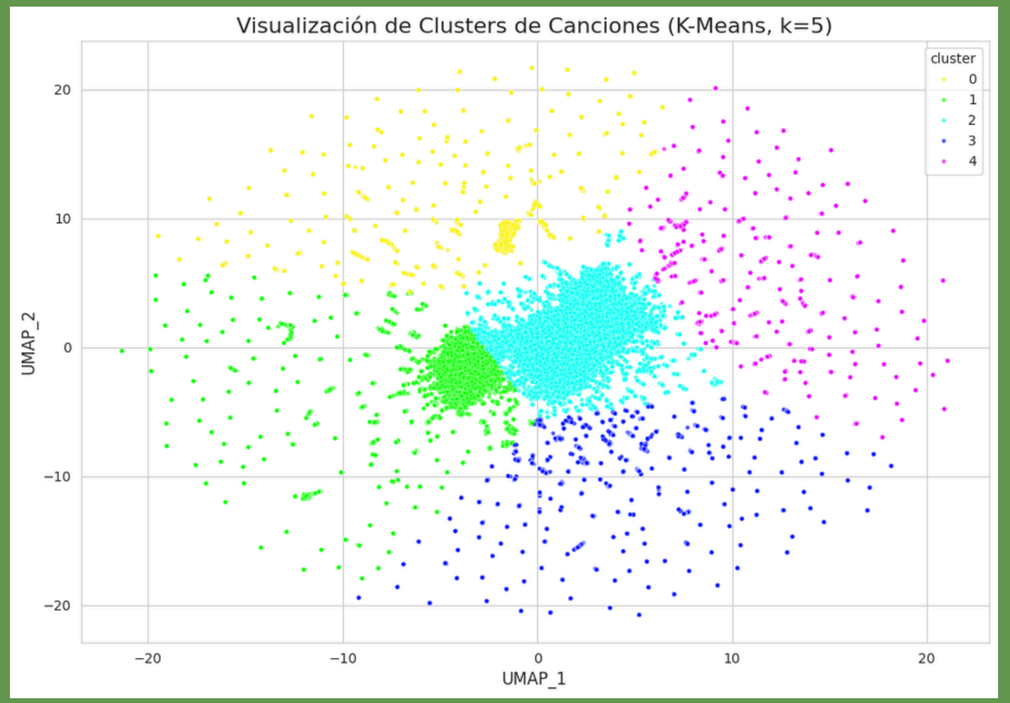
Algoritmo	Silhouette Score (más alto es mejor)	Davies-Bouldin Score (más bajo es mejor)
K-Means	0.141902	1.755768
DBSCAN	-0.13853	0.791794
MiniBatchKMeans	0.114168	1.930579

CLUSTERING

K-means



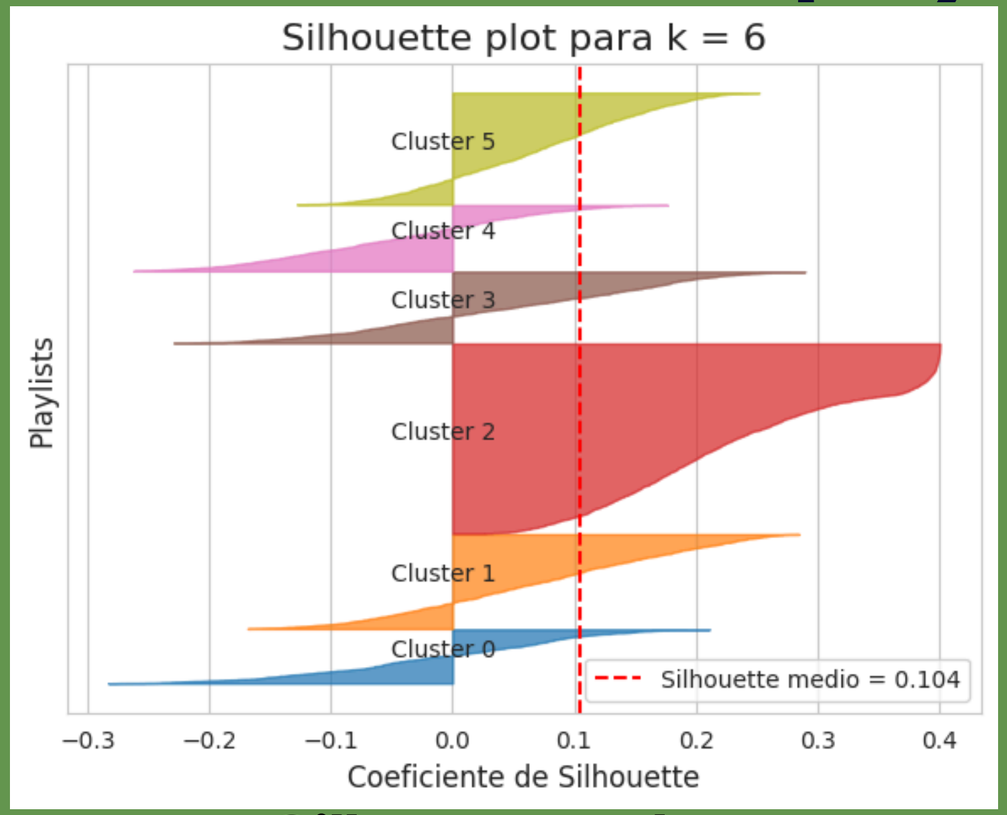
datos originales



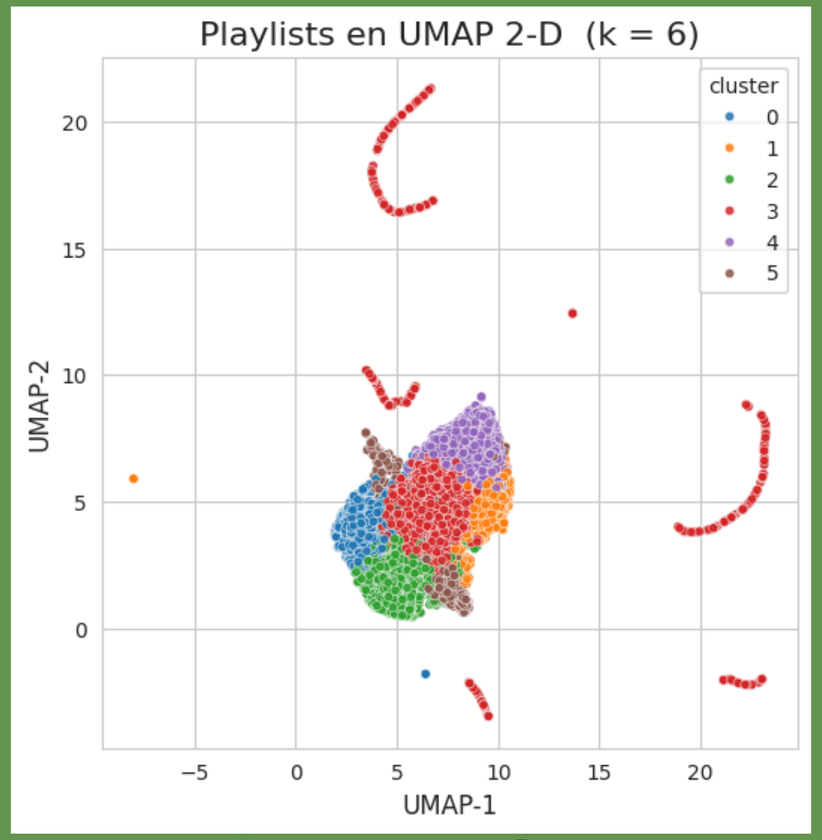
UMAP

CLUSTERING

Usando Feature de playlist



Silhouette plot



Scatterplot

CONCLUSIONES

CORTO PLAZO

- Etiquetar playlists por cluster.
- Features derivadas + embeddings.
- Clasificadores supervisados (SVM, RF, XGB).
- API REST de recomendación.
- Validación con usuarios reales.
- Ciclo feedback ↔ modelo.

LARGO PLAZO

Impacto potencial

Segmentación de usuarios por estado de ánimo.
Recomendadores multi-mood.
Curación editorial más precisa.