***Abstract***—**Este trabajo presenta el desarrollo de un pipeline de minería de datos orientado al análisis de similitud entre playlists utilizando el challenge set del Million Playlist Dataset (MPD) de Spotify. Con el fin de enriquecer los datos y capturar una representación más completa de las características musicales, se integró información adicional proveniente de la Spotify Web API, MusicBrainz y AcousticBrainz. El proyecto sigue el enfoque completo del proceso KDD (Knowledge Discovery in Databases),**

**incluyendo la integración y limpieza de datos, un análisis exploratorio detallado y la aplicación de técnicas de preprocesamiento orientadas a la reducción de dimensionalidad y la normalización de atributos. Posteriormente, se implementan métodos de agrupamiento (clustering) y métricas de similitud para identificar patrones latentes y estructuras de afinidad entre playlists. Los atributos considerados abarcan tanto propiedades acústicas (como danceability, energy, valence y tempo) como metadatos relevantes (géneros, popularidad, artistas), permitiendo una caracterización más rica de los gustos musicales. Este enfoque busca aportar herramientas para la comprensión de preferencias colectivas, con aplicaciones potenciales en sistemas de recomendación y análisis de comportamiento musical.**

**Index Terms—AcousticBrainz, análisis exploratorio, artistas,**

**características acústicas, clustering, dimensionalidad, energía,**

**géneros, KDD, Million Playlist Dataset, MusicBrainz, normalización, popularidad, preprocesamiento, recomendación, similitud, Spotify,**

I. INTRODUCCIÓN

En la era del consumo digital de música, las plataformas de streaming como Spotify han transformado la forma en que los usuarios descubren y organizan sus preferencias musicales. En este contexto, las playlists no solo representan colecciones arbitrarias de canciones, sino también manifestaciones estructuradas de gustos individuales y colectivos. Comprender las relaciones entre playlists, especialmente en términos de similitud, resulta fundamental para el diseño de sistemas de recomendación más eficientes y personalizados.

Este trabajo se enmarca en el análisis del challenge set del Million Playlist Dataset (MPD), un subconjunto de 10.000 playlists parciales provisto por Spotify para tareas de predicción y recuperación de contenido musical. Para enriquecer este conjunto de datos y obtener una representación más completa de cada ítem musical, se incorporó información proveniente de tres fuentes externas: la Spotify Web API, MusicBrainz y AcousticBrainz. Esta fusión de datos permitió acceder a atributos acústicos (como danceability, energy, valence y tempo), metadatos contextuales (como género, artista, popularidad) y descriptores semánticos adicionales.

El enfoque metodológico adoptado sigue el ciclo completo del

proceso de descubrimiento de conocimiento en bases de datos

(KDD), abarcando desde el análisis exploratorio y el preprocesamiento hasta la aplicación de técnicas de agrupamiento (clustering) y medición de similitud. El objetivo principal es identificar patrones latentes y agrupaciones naturales de playlists que compartan afinidades musicales, aportando así al entendimiento de las preferencias de los usuarios y a la mejora de sistemas de recomendación en entornos musicales.

II. DESCRIPCIÓN DE LOS DATOS

El conjunto de datos a considerar es un subconjunto de 10.000 playlist parciales provenientes del challenge set del Million Playlist Dataset (MPD) de Spotify. Cada registro incluye detalles como el título de la playlist, el título de las pistas y metadatos adicionales como la última fecha de edición y el número de playlists editadas. Los datos de este dataset abarcan playlists públicas creadas por usuarios estadounidenses desde enero de 2010 hasta noviembre de 2017. Cada playlist contiene 6 variables y una lista de hasta 100 canciones, de las cuales tienen cada una 8 atributos.

TABLE I  
DESCRIPCIÓN DE VARIABLES NIVEL PLAYLIST

| **Atributo** | **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 |

Adicionalmente, para enriquecer el dataset se agregan datos provenientes de MusicBrainz y AcousticBrainz, donde hay información adicional sobre las canciones en forma de descripciones acústicas y metadatos generados automáticamente a partir del análisis de las señales musicales. Estos serán atributos adicionales que buscan capturar elementos musicales como género, estado de ánimo, ritmo y timbre.. A cada canción registrada en los playlist del conjunto de datos original se le asocian los nuevos datos, que consisten en 22 variables adicionales. El dataset puede utilizarse para tareas de minería de datos como clasificación de géneros, detección de estados de ánimo, análisis de estilos musicales y segmentación de audiencias.

TABLE II  
DESCRIPCIÓN DE VARIABLES NIVEL CANCIÓN (DATASET ORIGINAL DEL CHALLENGE)

| **Atributo** | **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 |

TABLE III  
DESCRIPCIÓN DE VARIABLES NIVEL CANCIÓN (MUSICBRAINZ)

| **Atributo** | **Tipo** | **Descripción** |
| --- | --- | --- |
| mbid | Texto | MusicBrainz ID único para la canción |
| genre\_mb | Texto | Género según MusicBrainz |
| bpm | Real | Tempo promedio en beats por minuto |
| energy | Real | Intensidad percibida de la canción (escala relativa) |
| danceability\_ll | Real | Medida estimada de bailabilidad basada  en modelo de regresión logística |
| loudness | Real | Nivel relativo de volumen |
| rating\_value | Entero | Calificación media |
| rating\_votes | Entero | Número de votos asociados |

TABLE IV  
DESCRIPCIÓN DE VARIABLES NIVEL CANCIÓN (ACOUSTICBRAINZ, HIGH LEVEL)

| **Atributo** | **Tipo** | **Valores posibles** | **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). |
| genre\_dortmund | Categórico| Probabilistico | alternative, blues, electronic, folkcountry, funksoulrnb, jazz, pop, raphiphop, rock | Asigna la canción a uno de los grandes géneros musicales entrenados en el modelo de Dortmund. |
| genre\_electronic | Categórico| Probabilistico | ambient, dnb, house, techno, trance | Clasifica subgéneros dentro de la música electrónica según el modelo especializado. |
| genre\_rosamerica | Categórico| Probabilistico | cla (Classic), dan(dance), hip(hiphop), jaz(jazz), pop, rhy(rhythm and blues), roc(rock), spe (spoken word) | Género estimado por el modelo Rosamerica; cubre estilos que van de clásica hasta spoken word. |
| ismir04\_rhythm | Categórico| Probabilistico | ChaChaCha, Jive, Quickstep, Rumba-American, Rumba-International, Rumba-Misc, Samba, Tango, VienneseWaltz, Waltz | Identifica el ritmo/danza dominante según la taxonomía usada en ISMIR 2004. |
| genre\_tzanetakis | Categórico| Probabilistico | "blu": blues, "cla": classical (clásica), "cou": country, "dis": disco, "hip": hip hop, "jaz": jazz, "met": metal, "pop": pop, "reg": reggae, "roc": rock | Clasificación de género basada en el conjunto GTZAN de Tzanetakis; cubre los 10 géneros clásicos. |
| mood\_acoustic | Categórico| Probabilistico | acoustic, not\_acoustic | Indica si la pista suena predominantemente acústica (instrumentos no electrónicos). |
| mood\_aggressive | Categórico| Probabilistico | aggressive, not\_aggressive | Mide la energía y la agresividad percibida en la interpretación y producción. |
| mood\_electronic | Categórico| Probabilistico | electronic, not\_electronic | Señala la presencia dominante de texturas y elementos de producción electrónica. |
| mood\_happy | Categórico| Probabilistico | happy, not\_happy | Evalúa la valencia emocional de la pista, si transmite una sensación alegre. |
| mood\_party | Categórico| Probabilistico | party, not\_party | Determina si la pista es adecuada para un ambiente de fiesta o club. |
| mood\_relaxed | Categórico| Probabilistico | relaxed, not\_relaxed | Indica si la pista tiene un tempo y atmósfera tranquilos y relajantes. |
| mood\_sad | Categórico| Probabilistico | sad, not\_sad | Mide la carga emocional melancólica o triste de la canción. |
| moods\_mirex | Categórico| Probabilistico | Cluster1 - Cluster5 | Agrupa la canción en uno de cinco clusters de “estado de ánimo” definidos en la tarea MIREX de 2007. |
| tonal\_atonal | Categórico| Probabilistico | atonal, tonal | Clasifica la música según la presencia de una tonalidad clara (tonal) o ausencia de estructura tonal (atonal). |
| voice\_instrumental | Categórico| Probabilistico | instrumental, voice | Indica si predominan instrumentos sin canto (instrumental) o si hay voces destacadas (voice). |
| timbre | Categórico| Probabilistico | bright, dark | Describe la cualidad sonora general: “brillante” (agudos y presencias) o “oscura” (graves y resonancias). |

Los datos de los atributos expuestos en la tabla IV están disponibles en forma de una matriz de probabilidad: cada atributo tiene un diccionario donde se listan las probabilidades de cada categoría.   
Frente a todo el conjunto extendido de datos que se tiene se puede decir que se tiene una enorme dimensionalidad: 23 atributos por canción y 6 por playlist. Incluso se puede llegar a considerar variable ya que cada playlist puede albergar un número distinto de canciones.

III. OBJETIVO DE MINERÍA DE DATOS

El objetivo general de aplicar minería de datos sobre este conjunto extendido es descubrir agrupaciones y patrones de similitud musical entre playlists del challenge set. Entre los objetivos específicos encontramos:

●​ Filtrar playlists con al menos una canción visible.

●​ Comparar playlists entre sí usando artistas, canciones

y nombres.

●​ Agrupar playlists similares (clustering).

●​ Visualizar y analizar temáticas o estilos musicales

implícitos.

IV. METODOLOGÍA

**Proceso de Descubrimiento de Conocimiento (KDD):**

* **Selección e Integración de Datos:**

Para este estudio se cargó el challenge\_set completo del Million Playlist Dataset (10.000 playlists) y se filtraron aquellas con al menos una muestra, quedando unas 9.000 listas de reproducción. A cada pista se le incorporaron casi 120 atributos acústicos extraídos de AcousticBrainz, que tras un proceso de aplanamiento y limpieza se redujeron a 83 variables clave. Además, se aprovecharon metadatos de Spotify Web API y MusicBrainz (artist\_uri, track\_name, album\_uri) para enriquecer y unificar los registros mediante el identificador *track\_id*.

* **Preprocesamiento y Limpieza:**

El Data Frame resultante combina información de playlist y características acústicas en registros “canción–playlist”. Se eliminaron duplicados para asegurar unicidad en el clustering y columnas enteras sin datos (por ejemplo genre\_mb). Los valores faltantes numéricos se imputaron según la media de cada playlist —con mediana global de respaldo— y los categóricos se llenaron con “desconocido” o con la media de la mayor probabilidad según el atributo en cuestión. Para mitigar outliers se aplicó Winsorizing al 1 %–99 %, tras lo cual las variables continuas se escalaron con StandardScaler. Finalmente, se discretizaron variables clave en tres niveles (bajo, medio, alto) y se generó una matriz binaria de transacciones (playlists × tracks) para la minería de asociaciones.

* **Extracción de Reglas de Asociación:**

Con la matriz de transacciones se ejecutó FP-Growth, un algoritmo eficiente para datos voluminosos. Se fijó un soporte mínimo de 1 %, confianza superior a 0,2 y lift mayor de 1,5 para filtrar las reglas más relevantes. Entre los hallazgos destacan patrones como la co-ocurrencia de temas *pop* con pistas *danceable* (confianza ~65 %) y la fuerte asociación *acoustic*→*relaxed* (confianza > 70 %), lo que refleja afinidades estilísticas y emocionales claras dentro de las playlists.

* **Agrupamiento (Clustering):**

Para descubrir arquetipos musicales, se redujo la dimensionalidad de las 83 variables mediante PCA (reteniendo > 90 % de la varianza) y UMAP para una proyección 2D interpretativa. La curva de codo y el coeficiente de silhouette indicaron k = 5 como número óptimo de clusters en K-Means; comparaciones con MiniBatchKMeans, DBSCAN y Agglomerative confirmaron su estabilidad y claridad interpretativa. Los cinco grupos identificados fueron: High-Energy, Chill, Urban, Jazz & Classics y Mixed Popular, cada uno con perfiles acústicos y de género distintivos.

IV. RESULTADOS KDD

El análisis exploratorio reveló que la longitud media de las playlists es de 57 pistas, con outliers que superan las 100 canciones. En cuanto a géneros, destacan *rhy* (R&B) y *hip* (hip-hop), mientras que los moods *happy* y *party* dominan aproximadamente el 30 % de los casos. Las reglas de asociación confirmaron preferencias de co-ocurrencia robustas, y el clustering expuso cinco arquetipos bien diferenciados, lo que sugiere la existencia de segmentos de usuarios con estilos y estados de ánimo claramente definidos.

V. (A EDITAR) TRABAJO FUTURO

Se demuestra que un pipeline KDD completo—desde la integración de datos hasta la minería y el clustering—puede desvelar relaciones profundas en el MPD y enriquecer sistemas de recomendación más allá de la similitud de audio, incorporando afinidad emocional y estilística. Como pasos siguientes, se planea etiquetar playlists según los clusters obtenidos, construir features derivadas de reglas y embeddings de audio, y entrenar clasificadores supervisados (Logistic Regression, SVM, Random Forest, XGBoost, MLP). Se contemplan validación externa, optimización de hiperparámetros, análisis de interpretabilidad (SHAP/LIME), despliegue vía API REST y un ciclo de feedback con usuarios para refinar el modelo.