Escuela de Ingeniería y Ciencias

Campus Monterrey

ML-1 - Proyecto Machine Learning
Selección de Dataset
Unidad de formación TC3006C.102
Grupo 102
Equipo 5

Nadia Salgado Alvarez	**A01174509**
Regina Reyes Juárez	**A01275790**
Gilberto Angel Camacho Lara	**A01613895**
Santiago Miguel Lozano Cedillo	**A01198114**

17 de agosto de 2025

Indice

1. Motivación (¿Por qué Spotify?)

La música siempre ha sido muy importante en la vida de las personas; sin embargo, la forma de consumirla ha cambiado significativamente en los últimos años debido a tendencias sociales. Por ejemplo, en los años 90, cuando se hablaba de pop, los referentes eran artistas como Michael Jackson u otros muy conocidos, lo que permitía deducir qué género predominaba en ese momento.

Por esta razón, resulta interesante analizar las nuevas tendencias sociales, ya que el consumo musical actual va más allá de escuchar una canción. La popularidad de cada tema está fuertemente ligada a la tendencia social, los estilos musicales, los retos virales y otros fenómenos culturales.

En comparación con otros datasets, como los de Netflix o TMDB, que contienen muchos datos de tipo objeto y requieren transformaciones complejas para poder modelarlos, el dataset de Spotify cuenta con más datos numéricos, lo que permite formular hipótesis más interesantes y explorar la relación entre música y tendencias sociales sin que la transformación de los datos afecte de manera negativa al modelo.

2. Preguntas de investigación

- ¿Se puede inferir comportamientos o actitudes culturales de la sociedad a partir de la popularidad de ciertos géneros o artistas musicales?
- ¿Se puede predecir el género de una canción a partir de las características de audio?
- ¿Se puede predecir la popularidad de una canción a partir de las características de audio?

3. Variable objetivo

Para este proyecto nuestra variable objetivo será **popularity**, por lo cual a través de las potenciales variables para tomar en cuenta en la predicción: **danceability, energy y tempo**, buscaremos predecir el puntaje de popularidad de una canción que puede ir desde 0 hasta 100.

4. Fuente

El dataset proviene de Kaggle y contiene información recopilada de la plataforma Spotify. Incluye aproximadamente 114,000 canciones, acompañadas de datos relevantes como el artista, nombre del álbum, título de la canción, género y diversas características musicales. Una ventaja importante de este dataset es que es público y ha sido ampliamente utilizado en proyectos de análisis musical, lo que lo convierte en una fuente confiable y relevante para esta investigación.

5. Diccionario de variables

Variable	**Tipo**	**Descripción**
track_id	Texto	Identificador único asignado a cada canción en el dataset.
artists	Texto	Artista o grupo de artistas que interpretan la canción.
album_name	Texto	Nombre del álbum al cual pertenece la canción.
track_name	Texto	Título de la canción.
popularity	Numérica (0-100)	Medida que indica qué tan popular es la canción, basada en reproducciones recientes.
duration_ms	Numérica	Duración de la canción expresada en milisegundos.
explicit	Booleana	Indica si la canción contiene contenido explícito (1 = Sí, 0 = No).
danceability	Numérica [0-1]	Qué tan adecuada es la canción para bailar, considerando ritmo y estabilidad.
energy	Numérica [0-1]	Intensidad percibida; canciones rápidas y ruidosas tienen valores altos.
key	Entera (0-11)	Tono musical de la canción.
loudness	Numérica (dB)	Nivel promedio de volumen de la canción; valores negativos representan menor volumen.
mode	Binaria	Modalidad de la canción: 1 = mayor (alegre), 0 = menor (triste).
speechiness	Numérica [0-1]	Proporción de palabras habladas en la pista. Valores altos = rap, podcasts.
acousticness	Numérica [0-1]	Estima la probabilidad de que una canción sea acústica.
instrumentalness	Numérica [0-1]	Probabilidad de que una canción sea instrumental.
liveness	Numérica [0-1]	Indica presencia de público en la grabación; valores altos sugieren que es en vivo.
valence	Numérica [0-1]	Nivel de positividad o alegría percibida en la canción.
tempo	Numérica (BPM)	Velocidad de la canción medida en beats por minuto.

Variable	Tipo	Descripción
time_signature	Entera	Compás estimado de la canción (ej. 3 = 3/4, 4 = 4/4).
track_genre	Categórica	Género musical asignado a la canción (ej. pop, rock, hip-hop, jazz).

6. Riesgos y limitaciones

- **Correlaciones bajas con la variable objetivo**: La popularidad no muestra relaciones lineales fuertes con la mayoría de las variables numéricas, lo que complica la predicción directa usando técnicas simples de regresión.
- **Dependencia de factores externos**: La variable *popularity* está influenciada por fenómenos externos a las características de audio (marketing, tendencias en redes sociales o inclusión en playlists oficiales), lo que introduce ruido que puede reducir la capacidad predictiva de los modelos.
- **Sesgo hacia géneros dominantes**: El dataset está fuertemente inclinado hacia géneros mainstream como pop y hip-hop, lo que puede generar modelos sesgados que no generalicen bien para géneros con menor representación.
- **Datos faltantes e inconsistencias**: Algunas columnas como *track_genre* presentan valores faltantes o inconsistentes, requiriendo técnicas de limpieza e imputación que podrían afectar la calidad final del modelo.
- **Necesidad de normalización**: Variables como *duration_ms*, *loudness* y *tempo* presentan escalas muy distintas y valores atípicos, lo que obliga a aplicar normalización o estandarización para evitar que ciertas variables dominen el entrenamiento.
- **Limitaciones en la definición de popularidad**: El cálculo de la popularidad por Spotify depende de la frecuencia de reproducciones y de la recencia", lo que implica que la métrica puede no reflejar adecuadamente la calidad musical ni la percepción cultural a largo plazo.

7. Conclusión inicial

En esta primera entrega se logró justificar la elección del dataset de Spotify frente a otras alternativas. La música resulta un ámbito cercano y relevante para el análisis, ya que refleja cambios culturales y sociales que influyen directamente en la popularidad de los artistas y sus canciones. Además, al tratarse de un conjunto de datos con gran cantidad de variables numéricas, se facilita el planteamiento de modelos de predicción y se reduce la complejidad en el procesamiento inicial de los datos, a diferencia de otros datasets como los de Netflix o TMDB. Durante el análisis también se identificaron algunos retos importantes. La variable de popularidad no presenta correlaciones fuertes con las demás características, lo que anticipa la necesidad de probar modelos más complejos que puedan captar patrones no lineales. Igualmente, factores externos como las tendencias en redes sociales o el marketing influyen en la popularidad y representan una limitación que debe considerarse. El dataset de Spotify ofrece una base sólida para el desarrollo del proyecto, con suficiente información para formular hipótesis interesantes y al mismo tiempo con limitaciones que harán necesario aplicar un proceso riguroso de limpieza, normalización y selección de técnicas de modelado en las siguientes etapas del trabajo.

Además, este primer acercamiento permitió identificar los pasos que se deberán seguir en las siguientes fases del proyecto. La limpieza y preparación de los datos será fundamental para corregir valores faltantes, tratar outliers y aplicar técnicas de normalización que permitan trabajar con variables en escalas distintas. Estos procesos no sólo garantizarán la calidad del dataset, sino que también facilitarán la implementación de modelos predictivos más robustos. De esta forma, el trabajo realizado en esta semana establece una base sólida para continuar con el ciclo completo de aprendizaje automático, avanzando hacia el modelado, la optimización y la evaluación en las próximas fases.

8. Fuentes

- **Dataset**:
  `https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset`
- **Material complementario (videos)**:
  `https://drive.google.com/drive/folders/1427rtiko5B4zHfT6uiTp9AOHILpQmusV?usp=sharing`
- **Repositorio del proyecto**:
  `https://github.com/santiagolc02/Proyecto-ML`