Instituto Tecnológico y de Estudios Superiores de Monterrey

TC3006C: Inteligencia Artificial Avanzada para la Ciencia de Datos (Gpo 102)

Entregable 1: Selección de Dataset

Integrantes:
A01425341 - Carol Jatziry Rendon Guerrero
A00835733 - Rolando Ruiz Martínez
A01285367 - Natalia Olvera Ortiz
A01571463 - Ericka Sofía Rodríguez Sánchez
A00835576 - Marcos Renato Aquino Garcia

Profesor:
Dr. Alder López Cerda

Monterrey, N.L. a 17 de agosto de 2025

1. Motivación

Desde la aparición de las plataformas de streaming la música se ha transformado por completo, no solo en la forma que esta se produce, sino también en el cómo es escuchada. Spotify fue uno de los pioneros en este cambio, por lo que no solo ha sido testigo de estas transformaciones, sino que también fue capaz de acumular una gran cantidad de datos que reflejan la evolución de las tendencias musicales a nivel global. Esta información valiosa representa una gran oportunidad para explorar, analizar o incluso predecir, a través de técnicas de machine learning, diversos aspectos de la música en base a sus características.

2. Introducción

La revolución industrial musical, impulsada por plataformas como Spotify, ha brindado a los oyentes una mayor accesibilidad a la gran variedad de música que ofrecen artistas de todo el mundo. Ahora, a través de su consumo por medio de estas plataformas, podemos obtener datos acerca de las preferencias de los usuarios, abriéndonos la oportunidad de explorar qué factores influyen en la popularidad de las canciones. Y es que además de conocer aspectos básicos como el género o artista, ahora tenemos acceso a elementos fundamentales de las canciones como energía, tempo, duración, instrumentalidad, entre otros, que nos permiten explorar cuáles de ellos contribuyen al éxito de las canciones. Dentro de este contexto, el análisis de datos utilizando técnicas de machine learning tiene un gran potencial para encontrar patrones y hacer predicciones sobre la popularidad de las canciones, proporcionando insights sobre el impacto de sus características musicales.

3. Pregunta de investigación

Conociendo el contexto existen diversas rutas de investigación que podrían tomarse. En este caso, hemos decidido enfocar nuestro proyecto en la siguiente pregunta: *¿Pueden las características musicales de las canciones predecir su popularidad sin considerar el éxito del artista, o es este último un factor decisivo en el éxito de la canción?*

4. Variable objetivo

En este estudio, la variable objetivo es *popularity*, que es numérica y toma valores del 0 al 100. Esta variable se utiliza para medir el nivel de popularidad de las canciones en Spotify. Actualmente, al ser una variable continua, se puede abordar como un problema de regresión, en el cual se busque predecir el valor numérico de la popularidad. Sin embargo, también podrían establecerse niveles de popularidad en base a rangos, donde se trataría de un problema de clasificación.

5. Fuente

El conjunto de datos utilizado en este estudio proviene de Kaggle y fue publicado por Maharshi Pandya en 2022 (Pandya, 2022). Este dataset contiene una variedad de características de canciones de Spotify, como información sobre la popularidad, género, energía, entre otras métricas relevantes.

El dataset original cuenta con 114,000 registros y 21 columnas, de las cuales 1 es booleana, 9 son flotantes, 6 son enteros y 5 son objetos. En cuanto a la calidad del dataset, se encontraron 3 datos faltantes y 450 registros duplicados, lo que indica una calidad bastante aceptable de los datos.

El total de datos faltantes es de 3, por lo que, hablando sobre la calidad del dataset en cuanto a valores faltantes, se puede decir que es excelente.

En el 1 se puede mostrar un resumen estadístico de las variables numéricas en el dataset de Spotify.

Variable	**mean**	**std**	**min**	**50 %**	**max**
popularity	33.238535	22.305078	0.000	35.000000	100.000
duration_ms	228029.153114	107297.712645	0.000	212906.000000	5237295.000
danceability	0.566800	0.173542	0.000	0.580000	0.985
energy	0.641383	0.251529	0.000	0.685000	1.000
key	5.309140	3.559987	0.000	5.000000	11.000
loudness	-8.258960	5.029337	-49.531	-7.004000	4.532
mode	0.637553	0.480709	0.000	1.000000	1.000
speechiness	0.084652	0.105732	0.000	0.048900	0.965
acousticness	0.314910	0.332523	0.000	0.169000	0.996
instrumentalness	0.156050	0.309555	0.000	0.049000	1.000
liveness	0.213553	0.190378	0.000	0.132000	1.000
valence	0.474068	0.259261	0.000	0.464000	0.995
tempo	122.147837	29.978197	0.000	122.017000	243.372
time_signature	3.904035	0.432621	0.000	4.000000	5.000

Cuadro 1: Resumen estadístico (media, desviación estándar, mínimo, mediana y máximo) de las variables del dataset de Spotify.

Como panorama general, es posible apreciar que existe una variabilidad en cuanto al rango y la distribución de cada variable, donde algunas aparecen englobadas en los valores decimales entre 0 y 1 (energy, speechiness, acousticness, instrumentalness, liveness), mientras que otras aparecen como valores enteros en rangos distintos (key, mode, time_signature).

Así también, en la Figura 1 se observa cómo son las distintas distribuciones de las variables numéricas.

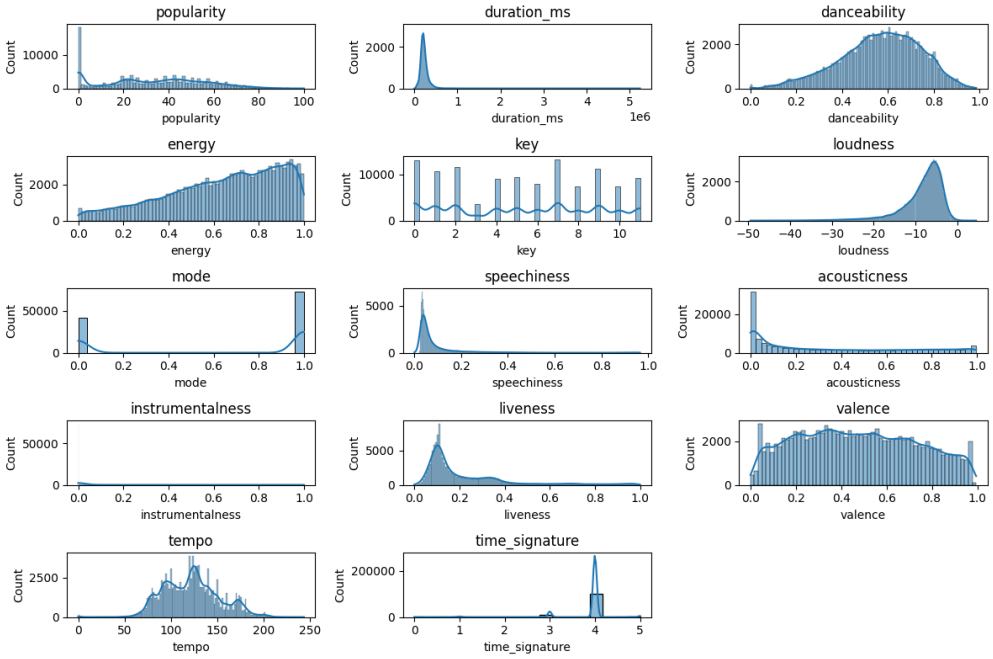

Figura 1: Histogramas de las variables numéricas.

Variables como danceability, tempo y valence se muestran con una curva con mayor dispersión, donde se mantienen valores frecuentes a lo largo del rango. Por el contrario, existen variables como duration_ms, acousticness y loudness, con picos más pronunciados y con sesgo, lo que indica que existen datos atípicos hacia algún lado de la distribución.

Por otro lado, en el Cuadro 2 aparecen los valores más frecuentes de algunas variables de tipo *object*.

Variable	Valor	Conteo
artists	The Beatles	279
artists	George Jones	271
artists	Stevie Wonder	236
artists	Linkin Park	224
artists	Ella Fitzgerald	222
album_name	Alternative Christmas 2022	195
album_name	Feliz Cumpleaños con Perreo	184
album_name	Metal	143
album_name	Halloween con perreito	123
album_name	Halloween Party 2022	115
track_name	Run Rudolph Run	151
track_name	Halloween	88
track_name	Frosty The Snowman	81
track_name	Little Saint Nick - 1991 Remix	76
track_name	Last Last	75
track_genre	acoustic	1000
track_genre	afrobeat	1000
track_genre	alt-rock	1000
track_genre	alternative	1000
track_genre	ambient	1000

Cuadro 2: Valores más frecuentes por variable en el dataset de Spotify.

Para la variable de artists, se muestran artistas como The Beatles y Linkin Park con una mayor frecuencia, de lo que se infiere que tienen un gran volumen de canciones en su discografía. Entre la variable track_name y album_name aparecen festividades del año como Christmas y Halloween, lo que indica que muchas canciones cubren la misma temática. Algo interesante que se aprecia en track_genre es que aparecen con 1000 de frecuencia por cada variable, y esto se repite de forma general para cada valor de género musical.

Debido a la gran cantidad de géneros, se hizo una agrupación de ellos para reducir la cardinalidad de la variable. En la Figura 2 se muestra cómo varía la popularidad según la agrupación de género musical.

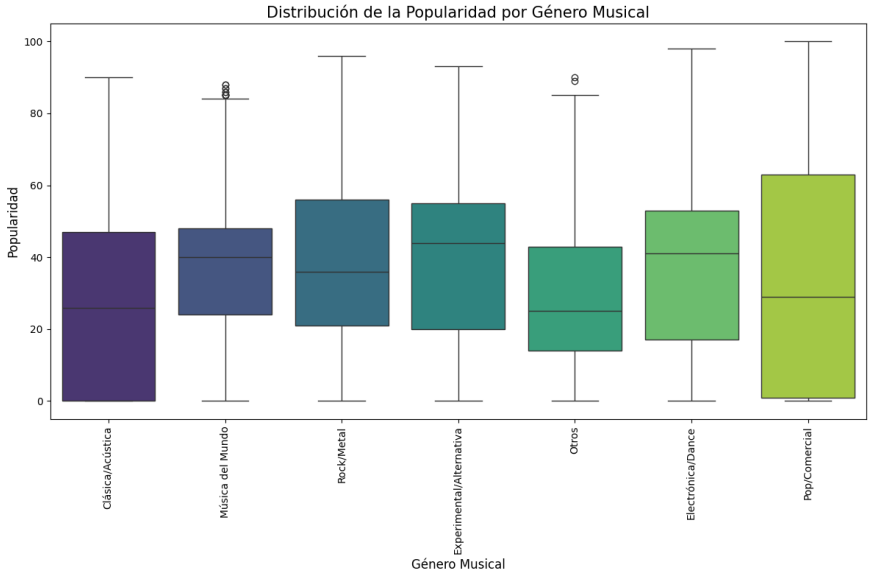

Figura 2: Boxplot de la poularidad según su género músical.

En esta se puede ver cómo los grupos Experimental/Alternativa y Electrónica/Dance son los más populares, mientras que la Clásica/Acústica y Otros se mantienen por debajo del rango.

Así también se estudió cómo es el perfil sonoro de la música basada en su grupo de género, mostrado en la Figura 3.

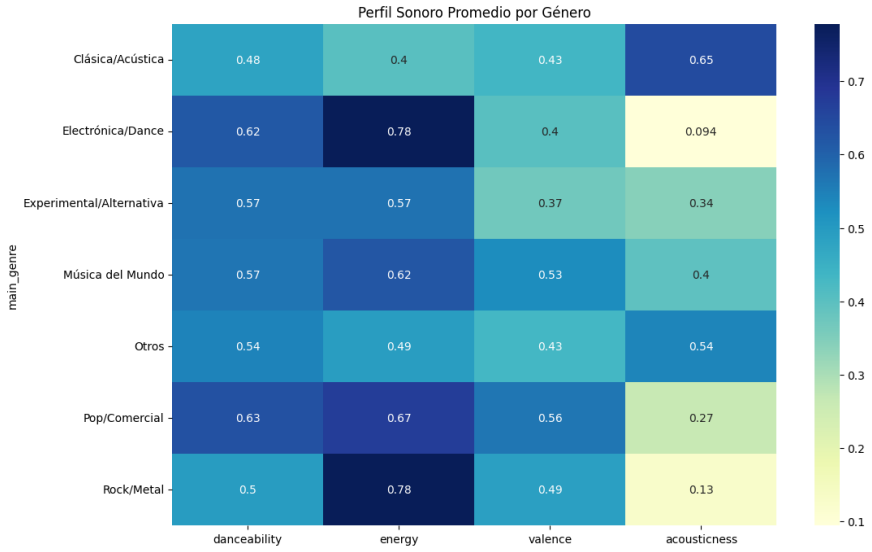

Figura 3: Diagrama de calor de características acústicas de música separada por género musical.

En cuanto a la variable de energía, los géneros de Rock/Metal y Electrónica/Dance sobresalen con un valor de 0.78, y al mismo tiempo tienen el valor menor en la categoría de acousticness,

siendo de 0.13 y 0.094, respectivamente. Por otro lado, el género de Pop/Comercial destaca en las variables de danceability, energy y valence. Por último, géneros agrupados en Otros y Clásica/Acústica cuentan con los valores más altos para acousticness.

Por último, se tiene la visualización de la correlación de las distintas variables númericas en la Figura 4, entre ellas las agrupaciones de géneros, que son realizadas como valores binarios.

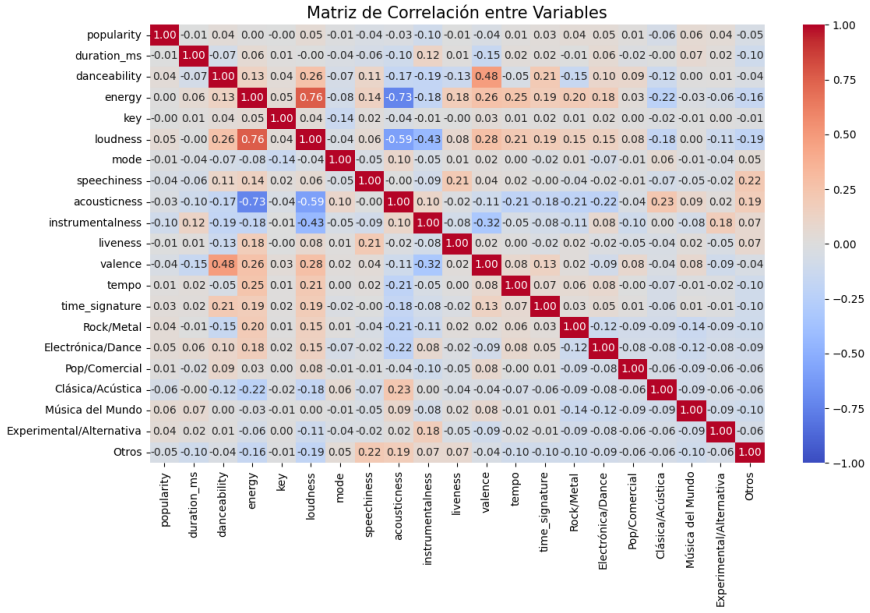

Figura 4: Diagrama de correlación de variables.

La variable de acousticness e instrumentalness tiene los valores más negativos de correlación con energy y loudness. En el lado positivo, variables como valence y danceability, así como loudness y energy, se mantienen altamente correlacionadas. Por otro lado, se aprecia cómo la variable de popularidad se mantiene constante con el valor cerca de 0 para una gran cantidad de variables, lo que indica poca o nula correlación lineal.

6. Diccionario de variables

- **popularity**: *(int64)* Nivel de popularidad de la canción de 0 a 100
- **loudness**: *(float64)* Nivel de volumen general dB
- **explicit**: *(bool)* 1 si la canción es explicita 0 si no
- **danceability**: *(float64)* Que tan bailable es la canción de 0 a 1
- **time_signature**: *(int64)* Métrica musical
- **tempo**: *(float64)* Velocidad de la canción en beats por minuto
- **energy**: *(float64)* Medida de intensidad de la pista de 0 a 1

- **key**: *(int64)* Tono musical codificado de 0 a 11
- **liveness**: *(float64)* Estima la presencia de público de 0 a 1
- **duration_ms**: *(int64)* Duración de la cancion en milisegundos
- **mode**: *(int64)* Modalidad de la pista 0 = menor, 1 = mayor
- **acousticness**: *(float64)* Confianza de que la pista sea acústica de 0 a 1
- **valence**: *(float64)* 0 = triste, 1 = feliz
- **speechiness**: *(float64)* Presencia de palabras habladas en la cancion de 0 a 1
- **instrumentalness**: *(float64)* Probabilidad de que la pista sea instrumental de 0 a 1
- **track_id**: *(object)* Id único de la canción
- **artists**: *(object)* Artista(s) que interpretan la canción
- **album_name**: *(object)* Álbum en el que aparece la canción
- **track_name**: *(object)* Nombre de la canción
- **track_genre**: *(object)* Género al que pertenece la canción

7. Riesgos/limitaciones

Uno de los posibles sesgos que tienen los datos, específicamente la variable de popularidad ("popularity"), se origina de la manera en que se calcula este dato. Esta se calcula con la cantidad de veces que se ha reproducido una canción y que tan recientes son estas reproducciones, lo que genera que los datos esten sesgados a lo que es exitoso al momento de haber extraido los datos. Respecto a la pregunta de investigación, el posible impacto que esto tendría, es que el modelo no podrá predecir la popularidad en general, si no la popularidad actual.

Este sesgo también se puede observar en la Figura 5, donde se observa que la moda en la medida de popularidad es de 0 a 5. Los datos están cesgados a la derecha sin embargo, esto no describe si la canción fue popular en algún momento en el tiempo.

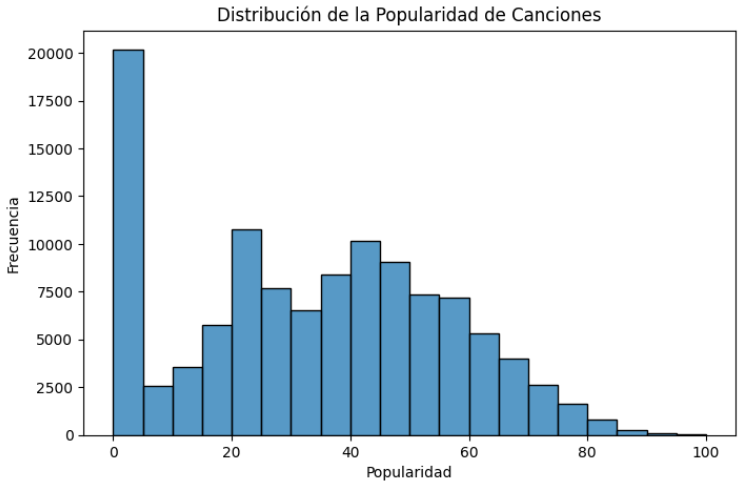

Figura 5: Histograma de la distribución de popularidad de canciones.

Una posible limitación puede ser el que no exista una variable de temporalidad, esto en el contexto de popularidad, nos limita mucho el scope que le podemos dar al proyecto, ya que si esta variable existiera, se podría analizar cómo evoluciona la popularidad de cierta canción con el tiempo o en sí visualizar tendencias estacionales como podría ser que en Navidad ciertas canciones se oigan más que en el resto del año, entre otras.

Referencias

Pandya, M. (2022). *Spotify tracks dataset.* Kaggle. Descargado de `https://www.kaggle.com/dsv/4372070` doi: 10.34740/KAGGLE/DSV/4372070

Anexos

Anexo 1: Link de GitHub: `https://github.com/natolvera/ML-Spotify-Equipo1/tree/main`