Análisis de la Popularidad de Artistas en Plataformas de Música y Redes Sociales

Universidad del Desarrollo
Facultad de Ingeniería
dataScience

Equipo de trabajo

Carlos Saquel

Ingeniero Civil Electrónico

Bruno San Martín

Juan Ugalde

Ingeniero en Biotecnología

Camilo Rivera

Ingeniero Civil Industrial

Introducción

Este estudio se centra en el análisis del dataset de Kaggle “Most Streamed Spotify Songs 2024”, que recopila información acerca de las reproducciones y likes de la canciones y artistas más populares de Spotify en el año 2024 y su éxito y popularidad en diversas plataformas de música y redes sociales, como Youtube, TikTok, Apple Music, entre otras.

Link del dataset:
https://www.kaggle.com/datasets/nelgiriyewithana/most-streamed-spotify-songs-2024

Objetivo

El propósito de este estudio es examinar cómo varían las tendencias de los artistas y canciones más reproducidas en distintas plataformas de música y redes sociales, a través de algoritmos y transformación de datos, con el fin de identificar posibles diferencias significativas en el éxito y la popularidad de los artistas en cada plataforma.

Descripción y limpieza del dataset

El dataset contiene **4.600 registros y 29 variables**, que representan las canciones más reproducidas en Spotify durante 2024. Incluye información sobre los **artistas, el ranking máximo alcanzado, la fecha de lanzamiento, y las métricas de reproducciones y likes** en diversas plataformas, tales como YouTube, TikTok, Apple Music, Deezer, Amazon, Pandora, SoundCloud y Shazam. Además, el dataset cuenta con una variable categórica binaria que indica si la canción es explícita.

Los tipos de variables presentes en el dataset son:

- 22 variables de tipo object (categóricas o de texto)
- 6 variables de tipo float64 (numéricas continuas)
- 1 variable de tipo int64 (numérica discreta)

	Track	Album Name	Artist	Release Date	ISRC	All Time Rank	Track Score	Spotify Streams	Spotify Playlist Count	Spotify Playlist Reach	...	SiriusXM Spins	Deezer Playlist Count	Deezer Playlist Reach	Amazon Playlist Count	Pandora Streams	Pandora Track Stations	Soundcloud Streams	Shazam Counts	TIDAL Popularity	Explicit Track
0	MILLION DOLLAR BABY	Million Dollar Baby - Single	Tommy Richman	4/26/2024	QM24S2402528	1	725.4	390,470,936	30,716	196,631,588	...	684	62.0	17,598,718	114.0	18,004,655	22,931	4,818,457	2,669,262	NaN	0
1	Not Like Us	Not Like Us	Kendrick Lamar	5/4/2024	USUG12400910	2	545.9	323,703,884	28,113	174,597,137	...	3	67.0	10,422,430	111.0	7,780,028	28,444	6,623,075	1,118,279	NaN	1
2	i like the way you kiss me	I like the way you kiss me	Artemas	3/19/2024	QZJ842400387	3	538.4	601,309,283	54,331	211,607,669	...	536	136.0	36,321,847	172.0	5,022,621	5,639	7,208,651	5,285,340	NaN	0
3	Flowers	Flowers - Single	Miley Cyrus	1/12/2023	USSM12209777	4	444.9	2,031,280,633	269,802	136,569,078	...	2,182	264.0	24,684,248	210.0	190,260,277	203,384	NaN	11,822,942	NaN	0
4	Houdini	Houdini	Eminem	5/31/2024	USUG12403398	5	423.3	107,034,922	7,223	151,469,874	...	1	82.0	17,660,624	105.0	4,493,884	7,006	207,179	457,017	NaN	1

5 rows × 29 columns

Descripción y limpieza del dataset

Con el objetivo de garantizar la calidad y confiabilidad de los resultados del análisis, se implementaron los siguientes pasos de limpieza y transformación del dataset:

- Eliminación de la variable '*TIDAL Popularity*' debido a la presencia de valores nulos, que podrían distorsionar las métricas.
- Eliminación de duplicados basándose en el identificador único de la canción '*ISRC*', para asegurar que cada registro sea único y relevante.
- Eliminación de registros con valores nulos en la columna de artistas, ya que la falta de esta información clave afectaría las conclusiones del análisis.
- Conversión de la columna '*Release Date*' de formato texto a formato de fecha, y creación de variables auxiliares como '*Año*' y '*Mes*' para facilitar el análisis temporal.
- Transformación de variables categóricas (tipo object) a formato numérico (float64) para habilitar el análisis cuantitativo.

Al finalizar el proceso de limpieza, el dataset quedó optimizado con 4,593 registros y 30 columnas, lo que permite un análisis más preciso y robusto, reduciendo el riesgo de sesgos o errores en los resultados.

```
<class 'pandas.core.frame.DataFrame'>
Index: 4593 entries, 0 to 4599
Data columns (total 30 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Track                        4593 non-null   object
 1   Album Name                   4593 non-null   object
 2   Artist                       4593 non-null   object
 3   Release Date                 4593 non-null   datetime64[ns]
 4   ISRC                         4593 non-null   object
 5   All Time Rank                4593 non-null   float64
 6   Track Score                  4593 non-null   float64
 7   Spotify Streams              4485 non-null   float64
 8   Spotify Playlist Count       4528 non-null   float64
 9   Spotify Playlist Reach       4526 non-null   float64
 10  Spotify Popularity           3794 non-null   float64
 11  YouTube Views                4290 non-null   float64
 12  YouTube Likes                4283 non-null   float64
 13  TikTok Posts                 3425 non-null   float64
 14  TikTok Likes                 3618 non-null   float64
 15  TikTok Views                 3617 non-null   float64
 16  YouTube Playlist Reach       3589 non-null   float64
 17  Apple Music Playlist Count   4037 non-null   float64
 18  AirPlay Spins                4100 non-null   float64
 19  SiriusXM Spins               2475 non-null   float64
 20  Deezer Playlist Count        3677 non-null   float64
 21  Deezer Playlist Reach        3670 non-null   float64
 22  Amazon Playlist Count        3543 non-null   float64
 23  Pandora Streams              3492 non-null   float64
 24  Pandora Track Stations       3330 non-null   float64
 25  Soundcloud Streams           1266 non-null   float64
 26  Shazam Counts                4017 non-null   float64
 27  Explicit Track               4593 non-null   int64
 28  Release Date Year            4593 non-null   int32
 29  Release Date Month           4593 non-null   object
dtypes: datetime64[ns](1), float64(22), int32(1), int64(1), object(5)
```

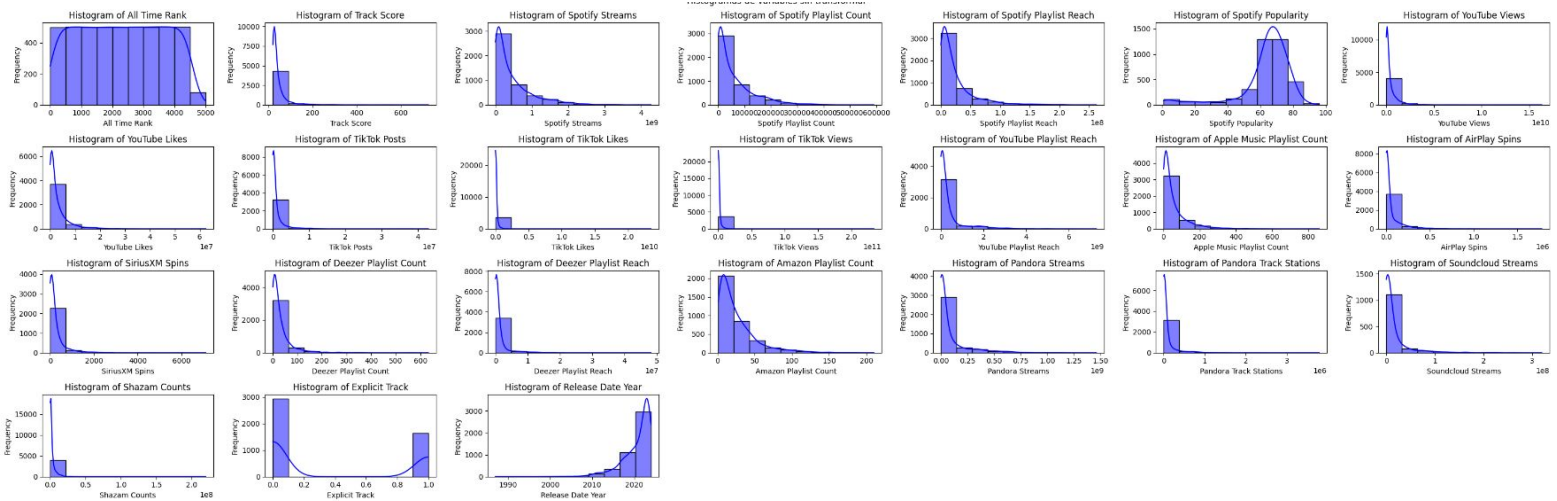
Análisis exploratorio de datos

Se comienza analizando las variables en su forma original, generando histogramas y posteriormente una matriz de correlación para detectar patrones.

- Los histogramas de algunas variables numéricas muestran sesgo positivo con colas más largas hacia la derecha. Para mejorar la distribución y facilitar el análisis, sería recomendable aplicar una transformación logarítmica.

- Adicionalmente, se detecta una cantidad significativa de valores nulos en varias variables, por lo que sería conveniente imputar los datos faltantes para optimizar el análisis.

Histogramas de variables sin transformar

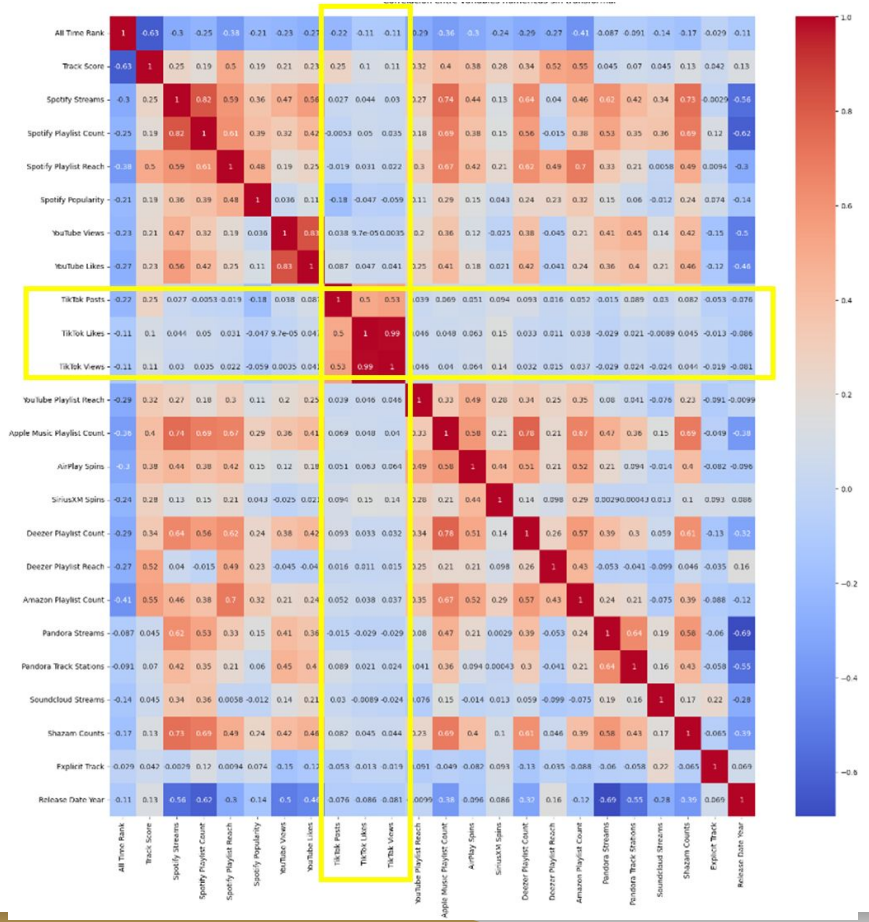


Matriz de correlaciones entre variables sin transformar

Examinar la matriz de correlación nos permite examinar las relaciones entre variables y detectar patrones que pueden influir en el análisis. En este caso, la matriz de correlación revela las siguientes observaciones:

- Existe una relación positiva entre las reproducciones y los likes en todas las plataformas, lo que sugiere que un aumento en el número de reproducciones suele estar asociado con un incremento en los likes.

- Sin embargo, en TikTok, la relación entre reproducciones y otras plataformas es negativa, lo que sugiere que el público objetivo en esta plataforma podría ser diferente.



Transformaciones y tratamiento de datos atípicos

Para optimizar el análisis y mejorar la calidad de los datos, se realizaron las siguientes transformaciones:

- **Transformación Logarítmica:** Se aplicó una transformación logarítmica a las variables con sesgo positivo para normalizar su distribución y facilitar un análisis más equilibrado y preciso.
- **Imputación de Valores Faltantes:** Se utilizaron técnicas de imputación para reemplazar los valores faltantes con la mediana, utilizando la función SimpleImputer de la librería sklearn. Esta estrategia ayuda a mitigar el impacto de las colas en la distribución de los datos.
- **Estandarización:** Se llevó a cabo una estandarización basada en la mediana y el Intervalo Intercuartil (IQR) para reducir la influencia de los valores atípicos y asegurar una distribución más uniforme de los datos.

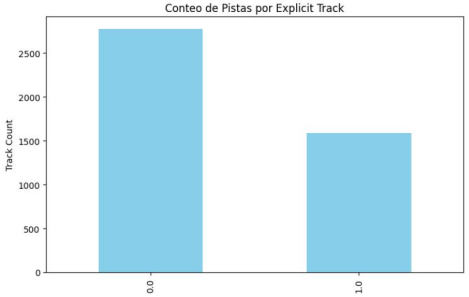
Estas transformaciones aseguran que los datos sean más homogéneos y fiables, facilitando un análisis más efectivo y una interpretación más precisa de los resultados.

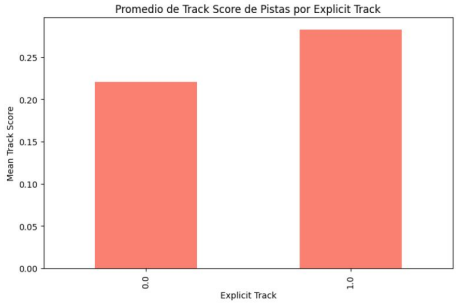
```
# imputar por mediana para evitar problemas por las colas
columns_imputed = df.select_dtypes(include='number').columns
df_imputado = df_transf_log.copy()

imputer = SimpleImputer(strategy='median')
imputer.fit(df_imputado[columns_imputed])

joblib.dump(imputer, 'imputer_mediana.pkl')

imputer_loaded = joblib.load('imputer_mediana.pkl')

df_imputado[columns_imputed] = imputer_loaded.transform(df_imputado[columns_imputed])
```

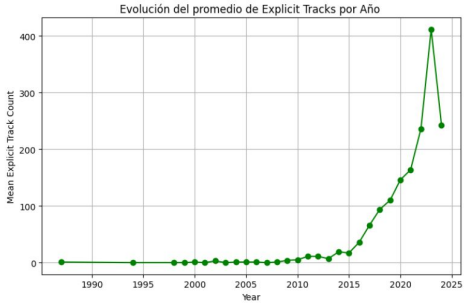
```
# En primera instancia vamos a standarizar con la mediana y el intervalo intercuartil
# para disminuir el efecto de las colas
columns_normalized = df.select_dtypes(include='number').columns
df_normalizado = df_imputado.copy()

scaler = RobustScaler()
scaler.fit(df_normalizado[columns_normalized])

joblib.dump(scaler, 'scaler_mediana.pkl')

scaler_loaded = joblib.load('scaler_mediana.pkl')

df_normalizado[columns_normalized] = scaler_loaded.transform(df_normalizado[columns_normalized])
```

Principales hallazgos

- **Música Explícita**: Comprender el impacto del contenido explícito en el streaming puede proporcionar información sobre cómo diferentes tipos de contenido resuenan con las audiencias. También puede ayudar a los artistas y sellos discográficos a tomar decisiones informadas sobre el contenido de su música y cómo podría afectar su alcance y popularidad.



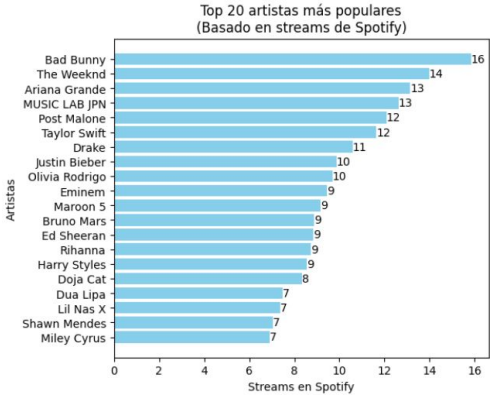


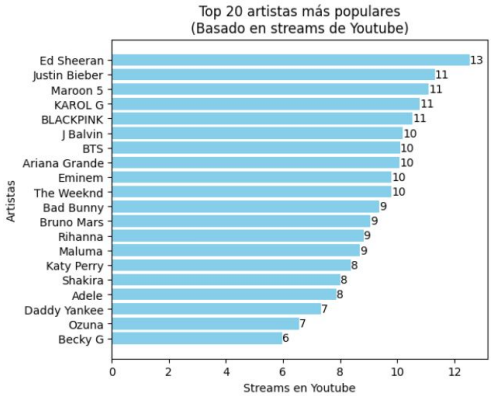


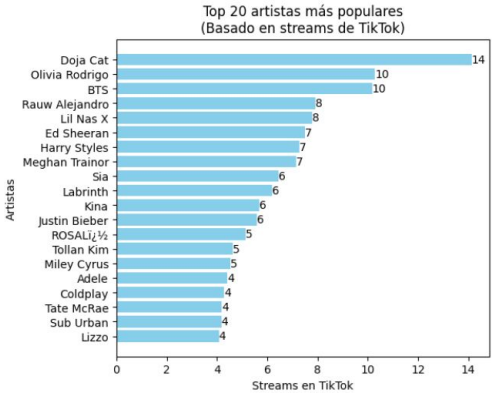
El análisis del contenido explícito en la música muestra su importancia en la escena actual. Lejos de reducir el alcance de una canción, este tipo de contenido puede hacerla más atractiva, sobre todo para los jóvenes. Además, la curva de popularidad indica que la música explícita ha cuadruplicado su demanda en los últimos 10 años. A medida que las actitudes sociales cambian, el impacto del contenido explícito seguirá siendo un tema de interés para artistas y sellos discográficos.

Principales hallazgos

- **Artistas más populares por plataforma**: Comprender el impacto de los artistas más escuchados en una plataforma es útil para identificar tendencias de consumo y preferencias del público. Este análisis ayuda a los artistas y a las empresas a ajustar sus estrategias de marketing, promoción y producción para alinearse con las demandas actuales del mercado.







El uso de canciones varía entre plataformas de redes sociales. Spotify y YouTube tienden a destacar canciones que son muy populares y taquilleras. En cambio, TikTok al parecer elige canciones principalmente por su capacidad para mejorar el contenido de los videos, independientemente de su popularidad general. Conocer estas diferencias es importante para entender las tendencias en cada plataforma y adaptar las estrategias de marketing y contenido de manera efectiva.

Conclusión

En conclusión, la música explícita puede ayudar a aumentar la popularidad de una canción entre audiencias jóvenes, mientras que las preferencias de los artistas varían por plataforma, con Spotify y YouTube destacando éxitos globales y TikTok enfocándose en canciones que mejoran el contenido de video.

El análisis y limpieza de datos son fundamentales para entender estas dinámicas, permitiendo a artistas y empresas adaptar sus estrategias de manera efectiva y aprovechar las oportunidades de mercado. Estos procesos aseguran una interpretación precisa de las tendencias y la optimización de las campañas para alcanzar el máximo impacto