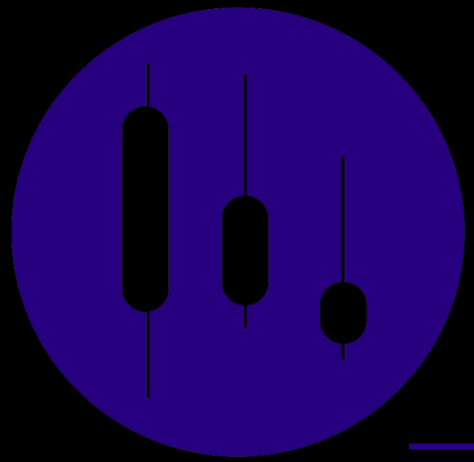
CODERHOUSE

Spotify



SpotiSci

Ciencia de datos aplicada a tu música

Por

Franco Falco

SpotiSci

SECCIONES

1. CONTEXTO COMERCIAL
2. PROBLEMA COMERCIAL
3. AUDIENCIA
4. PREGUNTA DE INVESTIGACIÓN E HIPÓTESIS
5. FUENTES
6. ANÁLISIS EXPLORATORIO Y MODELADO
7. CONCLUSIÓN

CONTEXTO COMERCIAL

- La industria de la música generó, a nivel global, **59.480 millones de dólares** en 2022.
- Solo el **9%** de la música se comercializa en formato **físico**.
- El ritmo de incremento en las suscripciones está en torno a **18% anual** y el número de usuarios registrados en 2022 es de **616 millones** en todo el mundo.
- Mercado caracterizado por una alta concentración del capital: **Universal Music Group** es en la actualidad la mayor compañía a nivel global, acaparando el **32,1% de los ingresos**.
- **Spotify** se constituye como actor dominante reteniendo alrededor del **31%** del total de suscriptores a servicios de ***streaming***.

PROBLEMA COMERCIAL

SpotiSci

- Para los distintos actores de la industria musical contemporánea es difícil predecir cómo va a comportarse un tema después de su lanzamiento. Es decir, si va a seguir una trayectoria "**exitosa**" o, más bien, va a ser un "**fracaso**" desde el punto de vista comercial.
- Es conocido que ciertas ***características socioeconómicas*** (por ejemplo, la difusión a través de distintos medios, el renombre del artista, etc.) influyen significativamente en el nivel de reproducciones y/o ventas.
- Sin embargo, una canción con características muy distantes a lo requerido para ser ***mainstream***, aún con buena publicidad, podría nunca alcanzar el éxito esperado.

AUDIENCIA

SpotiSci

Productores discográficos o artistas independientes que tengan la intención de adaptar las características de su música a los gustos y exigencias del mercado **con el objetivo de lograr una mejor *performance* de la misma** en términos de reproducciones y demás formas de interacción con los usuarios **en las plataformas de *streaming*.**

Pregunta de investigación

SpotiSci

¿Cuáles son las **principales características** musicales que se relacionan con la **popularidad** de un lanzamiento y que permitirían predecir un alto número de reproducciones / interacciones por parte de los usuarios?

HIPÓTESIS

SpotiSci

La ***popularidad*** de una determinada canción publicada en un servidor de *streaming* sigue una evolución temporal relativamente **predecible** a partir de ciertas características intrínsecamente musicales de la misma.

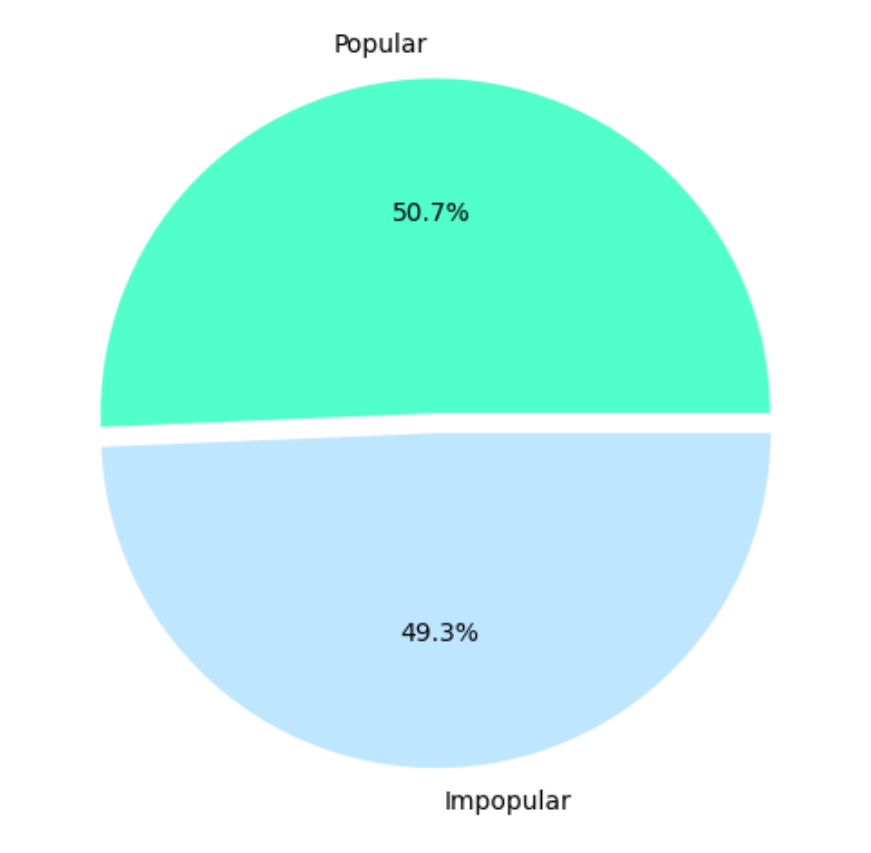
FUENTES

- ***Spotify popularity dataset:*** Es una base de datos, subida por un usuario al sitio *Kaggle*, que recoge información sobre *scores* de *popularidad* elaborados por Spotify para más de 40.000 canciones. En el conjunto de datos, la popularidad se representa como un parámetro binario asignando un valor de 1 para una canción popular y de 0 para una impopular. Una canción se considera popular si figura en el Billboard Hot 100, apareció en el mercado estadounidense y pertenece a un género convencional. Cuenta, entre sus ventajas, con un **amplio número de registros** y con información relevante sobre **variables intrínsecamente musicales**. Por otra parte, una desventaja es que no contiene información sobre el número de reproducciones de cada canción.

ANÁLISIS EXPLORATORIO

SpotiSci

POPULARIDAD

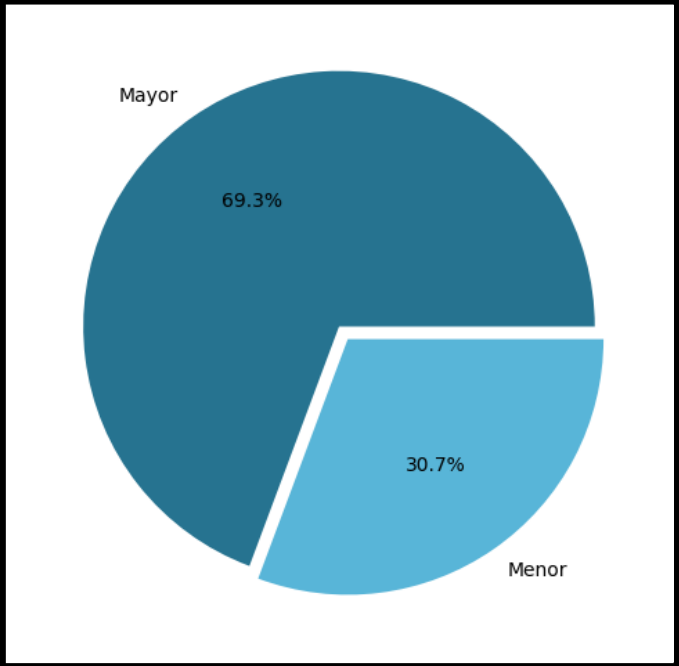


Aproximadamente el **51%** de los temas contenidos en este *dataset* fueron categorizados como **populares** y el **49%** restante como **impopulares**, lo cual da cuenta de una muestra **bien balanceada**, dato de interés a la hora de elaborar un modelo de *machine learning*.

ANÁLISIS EXPLORATORIO

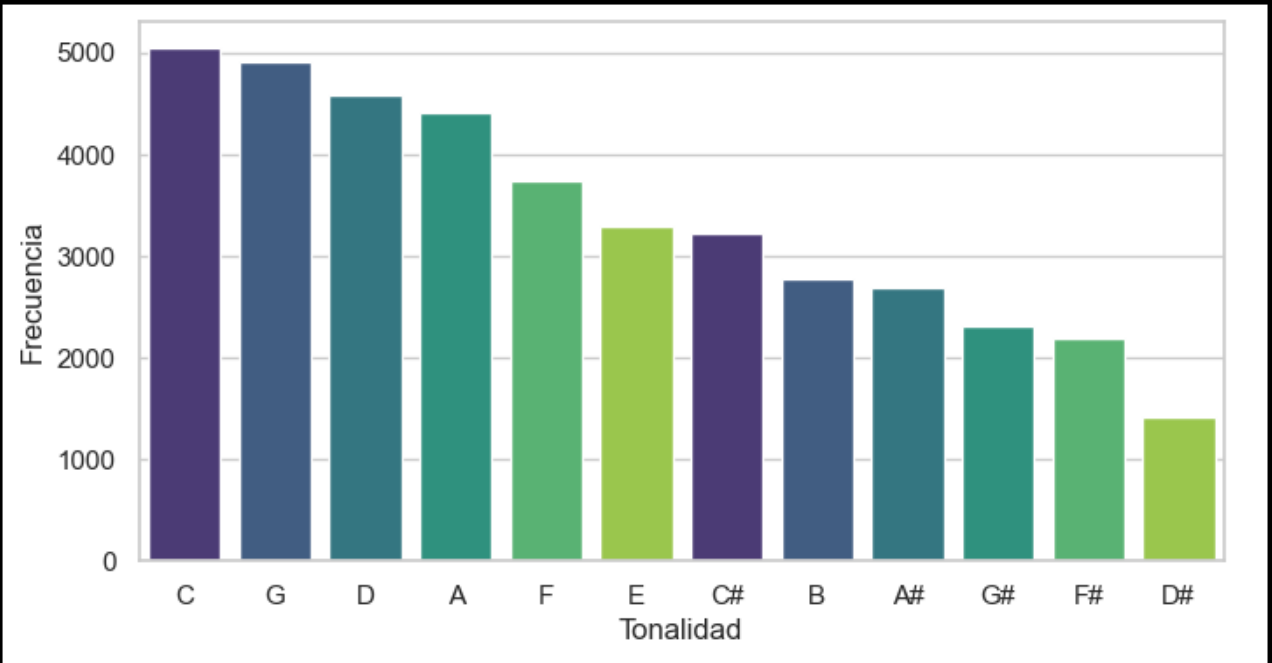
SpotiSci

MODO



El modo predominante es **mayor**.

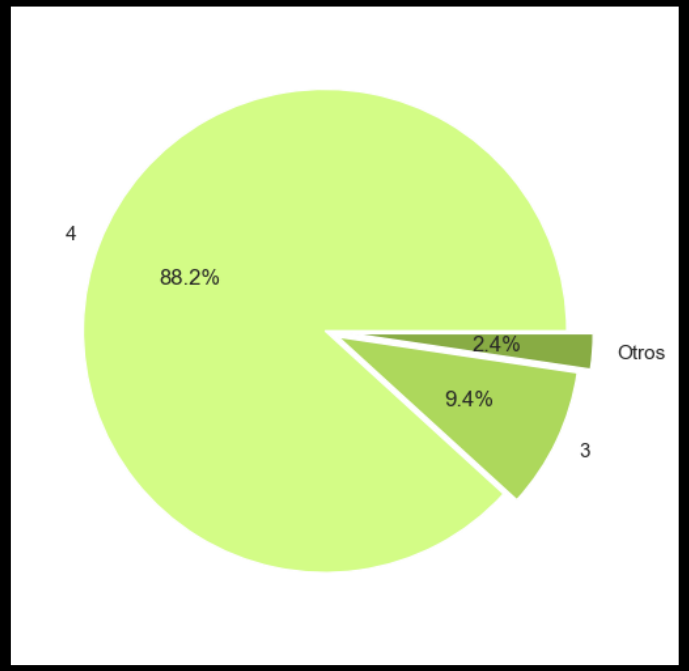
TONALIDAD



Alrededor de **la mitad** de los temas (46,7%) utilizan tonalidades de **C**, **G**, **D** o **A**, dejando a las restantes 8 tonalidades la otra mitad del dataset.

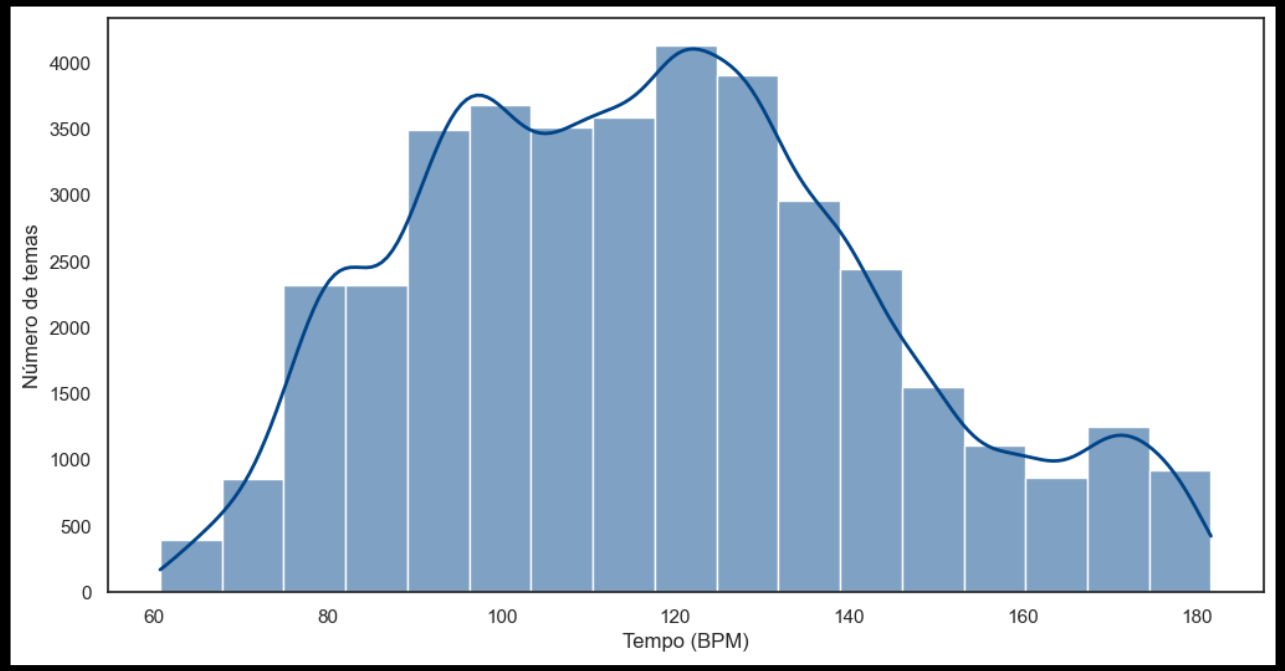
Análisis exploratorio

Compás



El compás de **4/4** es predominante.

Tempo

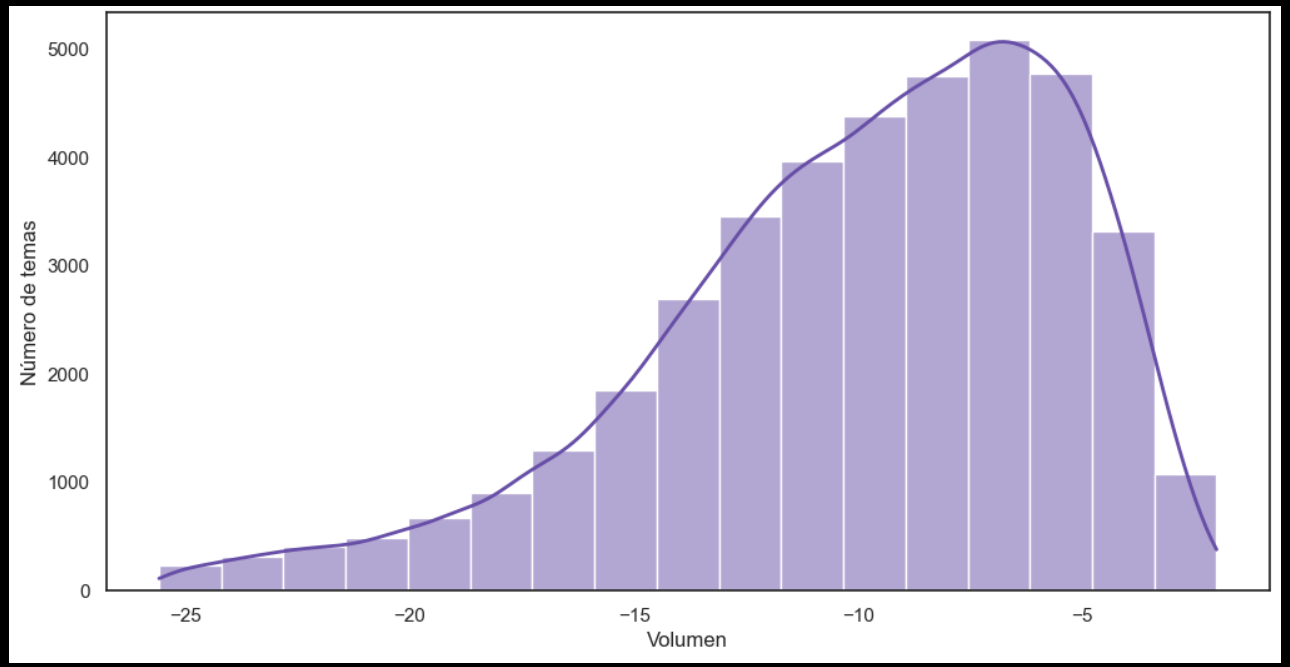


Cierta tendencia a una **bimodalidad** en la distribución, con valores que se concentran en torno a los **90** y a los **120 BPM**, si bien este último pico es más prominente.

ANÁLISIS EXPLORATORIO

SpotiSci

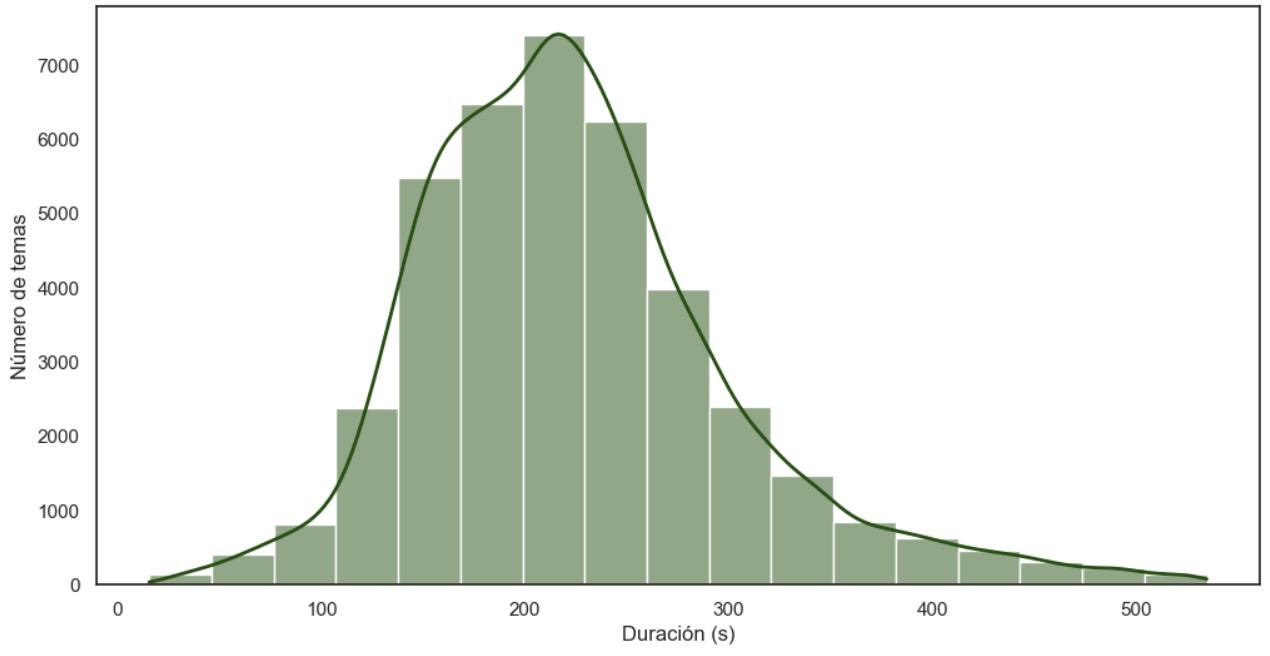
VOLUMEN



Solo alrededor de un **16%** de las canciones se encuentran **en torno a -14 LUFS** (rango de normalización sugerido por la norma *ITU 1770*), con un **70,4%** de las canciones **superando ese valor**, en un intento de aproximación a 0 dB que configura un claro *sesgo negativo*.

ANÁLISIS EXPLORATORIO

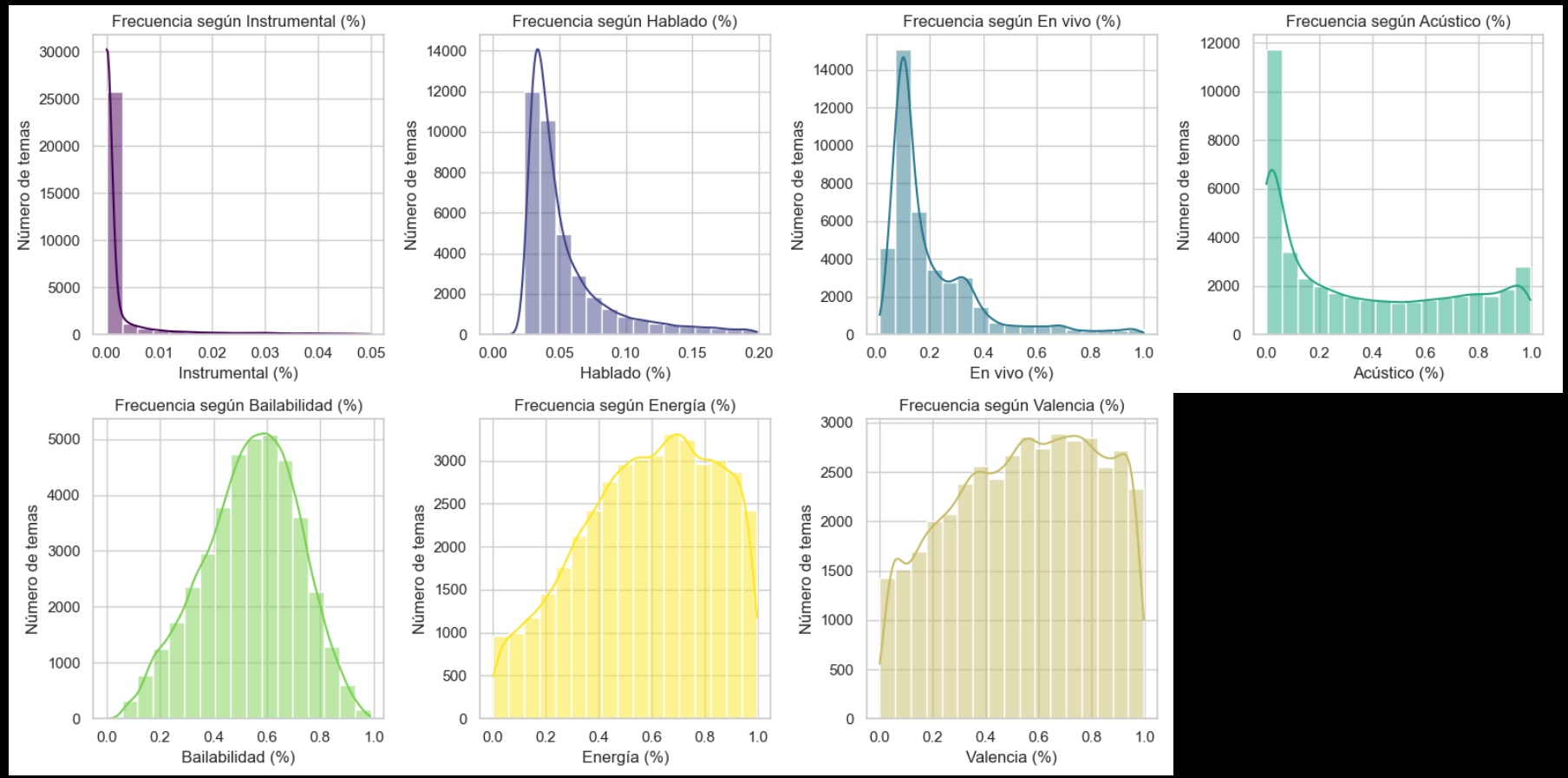
DURACIÓN



El grueso de los temas se concentra **entre los 160 y los 260 segundos**, aproximadamente. La distribución se acerca a la normalidad, aunque demuestra un ligero *sesgo a derecha*, con una predominancia de temas más breves que el promedio (**234,7 s**).

Análisis exploratorio

Variables subjetivas



Análisis exploratorio

SpotiSci

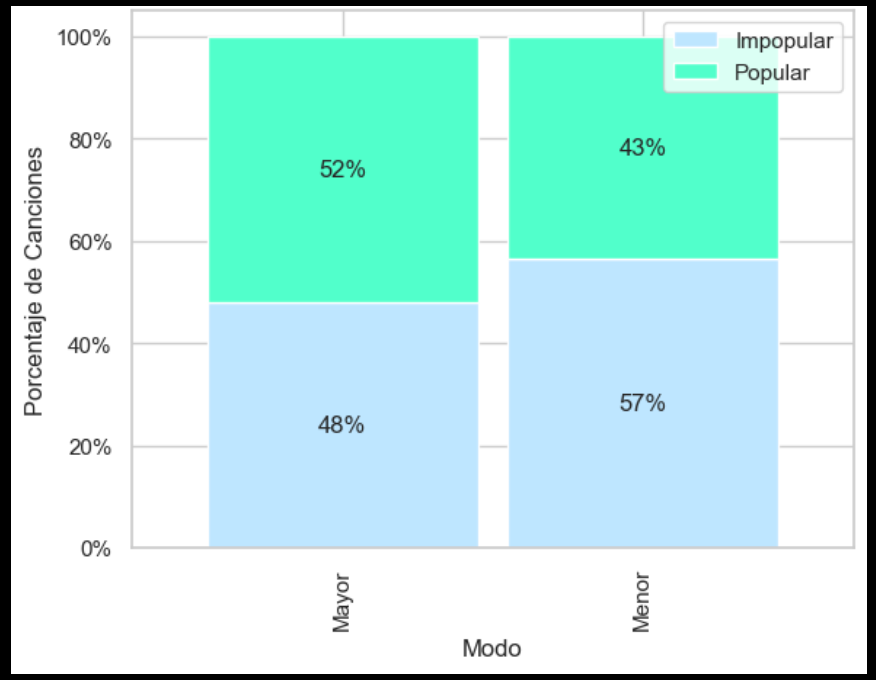
Variables subjetivas

1. Prácticamente, no hay temas **instrumentales** o la *instrumentalidad* de la muestra es muy baja.
2. El componente **hablado** es mínimo o nulo, con una mediana de 4,3%.
3. El número de temas **en vivo** y **acústicos** es ligeramente mayor, aunque aún con una distribución que presenta fuerte *sesgo positivo*. Las respectivas medianas son de 13,2% y 26,1%, respectivamente.
4. ***Bailabilidad***, **energía** y **valencia** se comportan de forma similar, con ligero *sesgo negativo* y una concentración de la distribución en torno al 55-60%. ***Bailabilidad*** es la variable con una distribución más homogénea y más *normal* (promedio de 53,9%, mediana de 55,1%).

SpotiSci

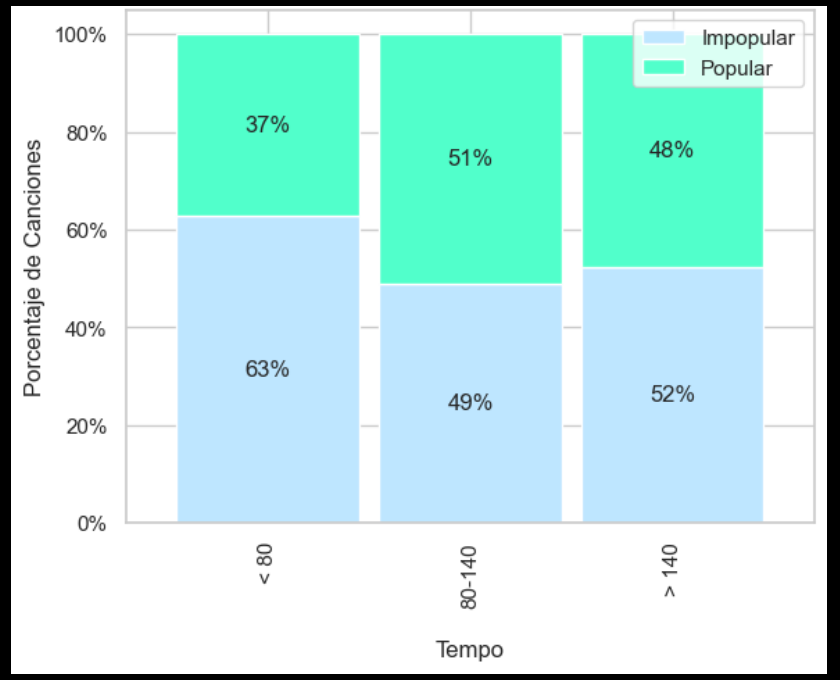
ANÁLISIS EXPLORATORIO

MODO VS. POPULARIDAD



Ligera tendencia hacia una **mayor popularidad** por parte de las canciones en **modo mayor**.

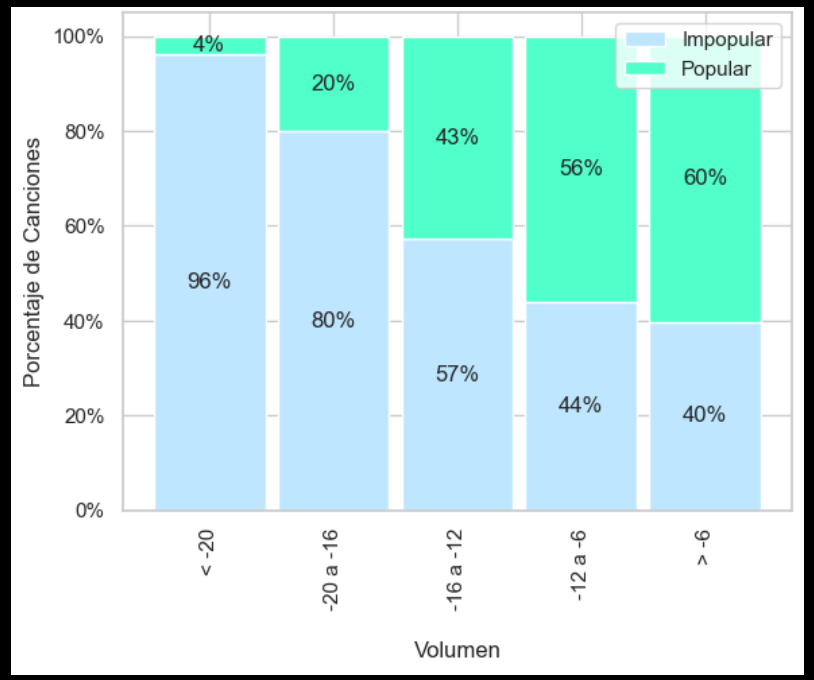
TEMPO VS. POPULARIDAD



Se verifica una tendencia hacia una **menor popularidad** en las canciones con **tempos por debajo de 80 BPM**.

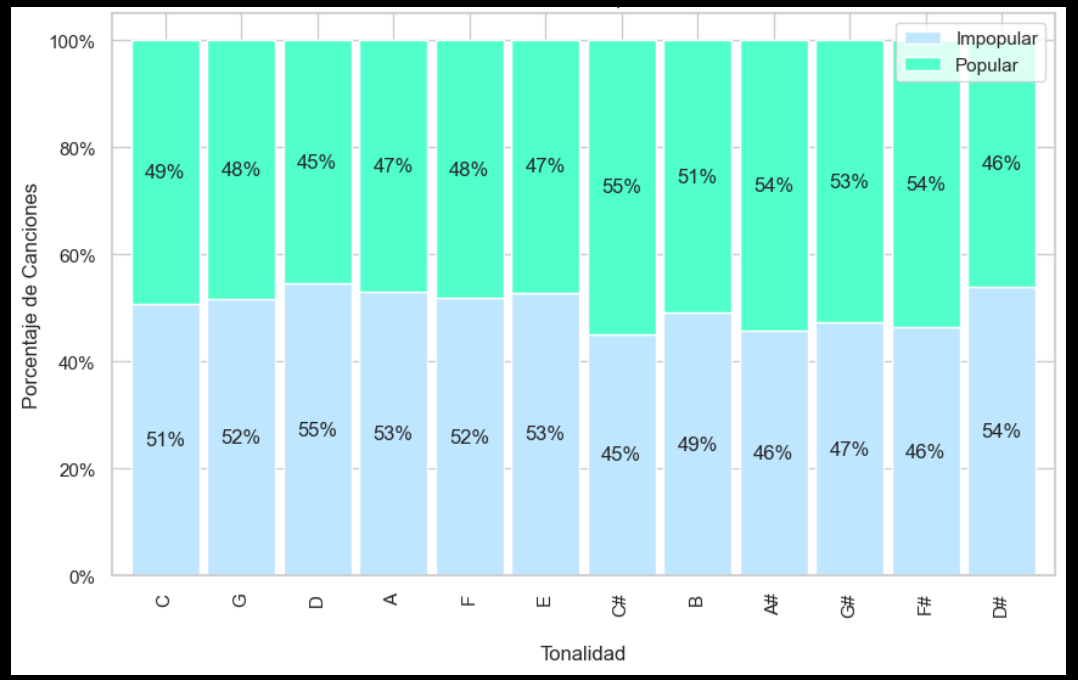
Análisis Exploratorio

Volumen vs. Popularidad



Clara tendencia hacia la **popularidad** a **mayor volumen**, lo cual explica mejor la curva de distribución.

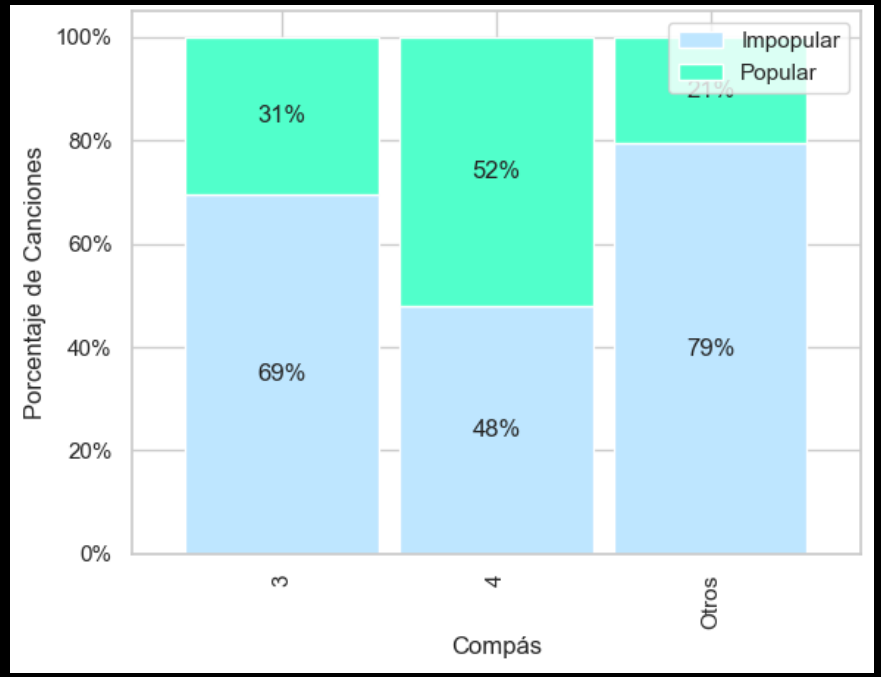
Tonalidad vs. Popularidad



No se objetivan tendencias sustanciales.

ANÁLISIS EXPLORATORIO

COMPÁS VS. POPULARIDAD

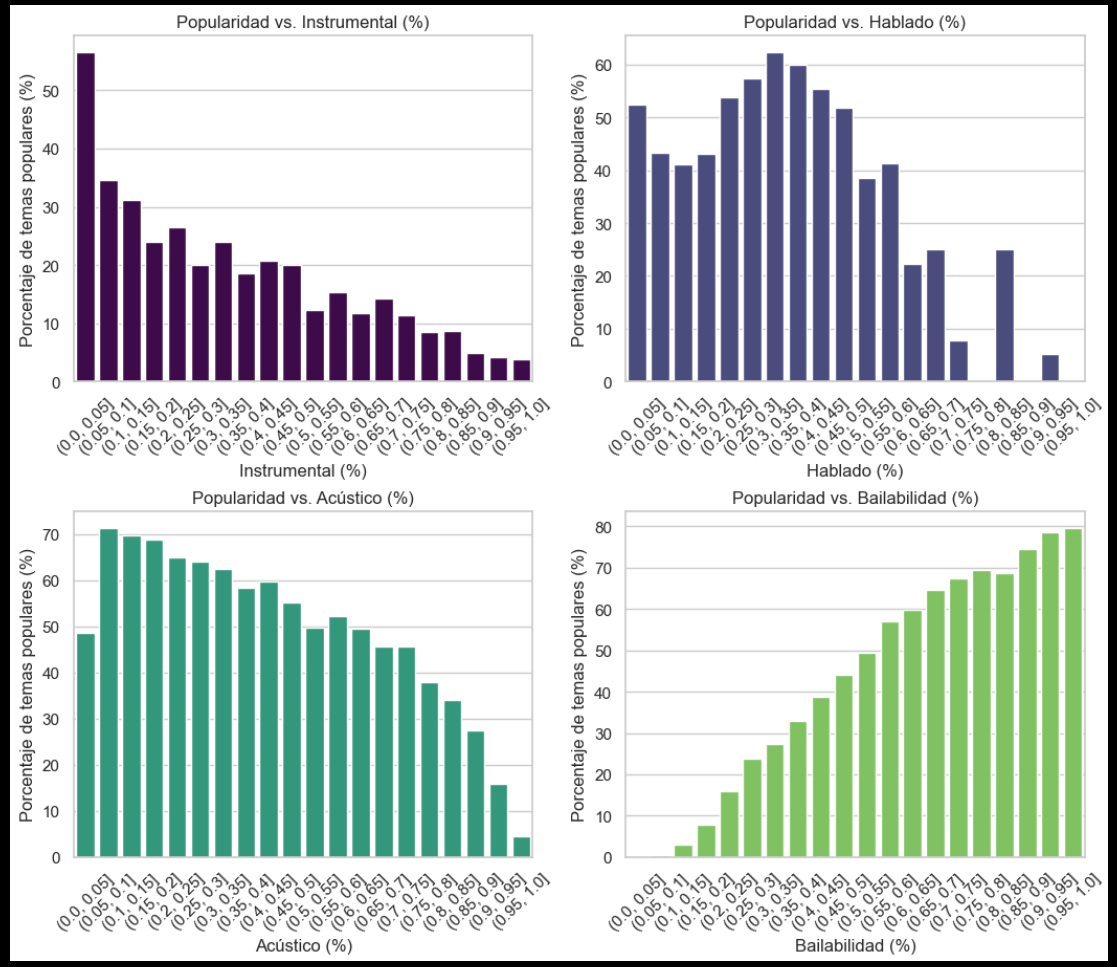


Los **compases de 3 tiempos** u **otras variantes** (por ejemplo: 5/4, 6/8, etc.) son mayoritariamente **impopulares**.

En el caso del **compás de 4/4** la proporción entre canciones populares e impopulares se encuentra *balanceada*.

Análisis Exploratorio

Variables Subjetivas vs. Popularidad



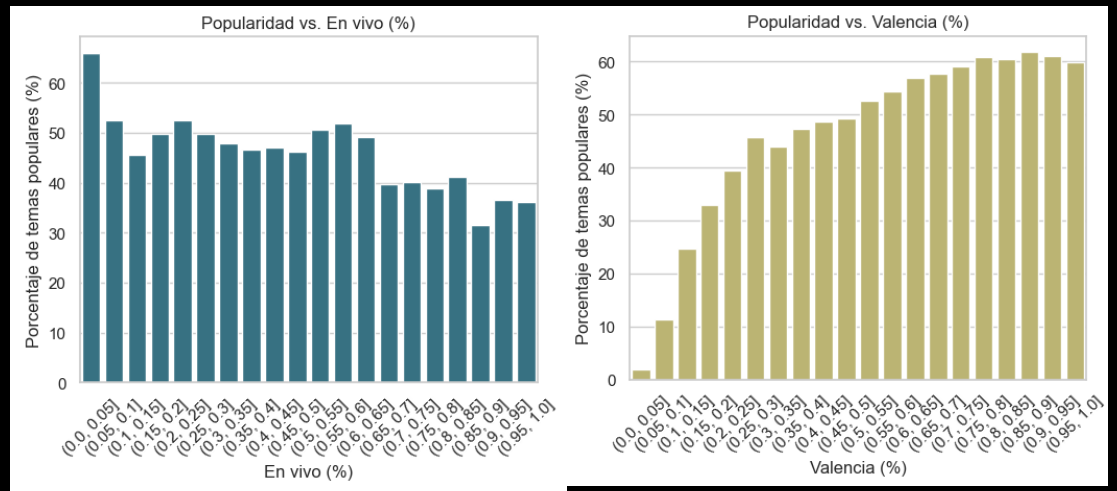
Las variables subjetivas **instrumental**, **hablado**, **en vivo** y **acústico** parecen relacionarse con el grado de *popularidad* mediante una ***proporción inversa***.

Lo contrario, una ***proporción directa***, ocurre con **bailabilidad**, **energía** y **valencia**.

La tendencia cobra mayor fuerza en **instrumental** y **bailabilidad**.

Análisis exploratorio

Variables subjetivas vs. popularidad



Las variables subjetivas **instrumental**, **hablado**, **en vivo** y **acústico** parecen relacionarse con el grado de *popularidad* mediante una ***proporción inversa***.

Lo contrario, una ***proporción directa***, ocurre con **bailabilidad**, **energía** y **valencia**.

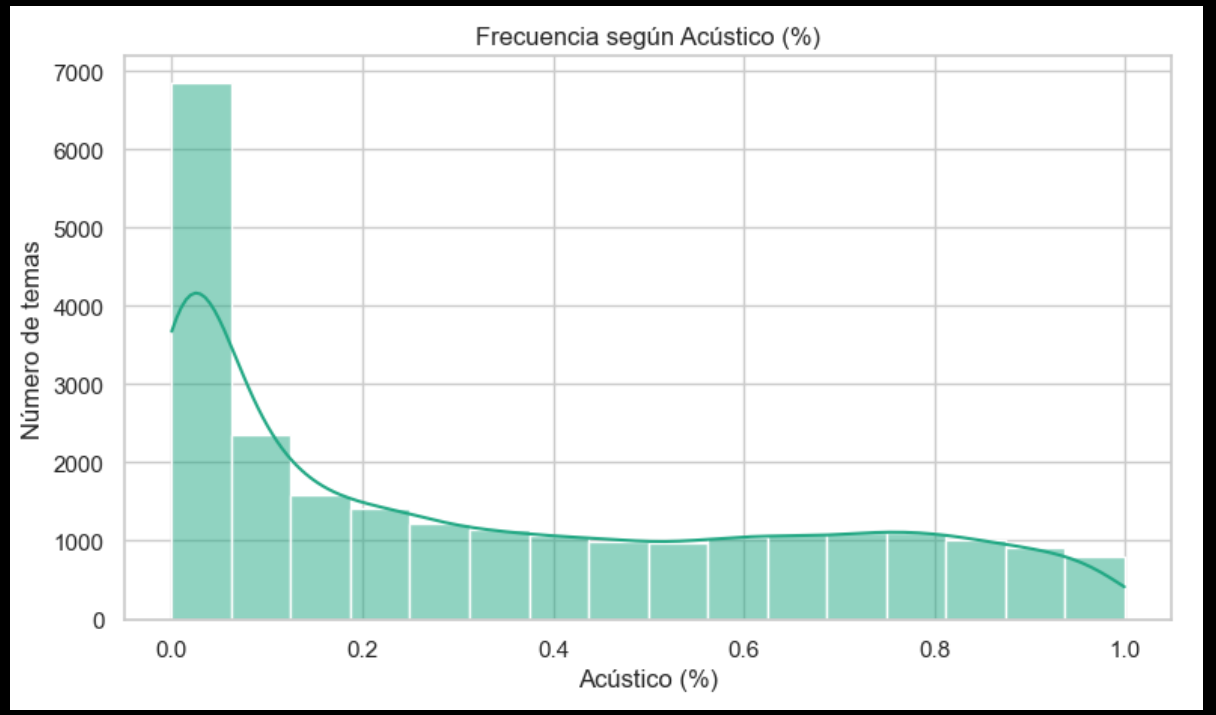
La tendencia cobra mayor fuerza en **instrumental** y **bailabilidad**.

MODELADO

- La variable categórica **popularidad**, de naturaleza *dicotómica*, definida como 0 (impopular) o 1 (popular) en base a criterios ya expresados, fue elegida como variable *target*.
- Se realizó limpieza de *outliers* a través del método de **rango intercuartílico (IQR),** dada la distribución *no normal* de todas las variables independientes no categóricas. El *dataset* libre de *outliers* se redujo a alrededor de 25.000 filas.
- A través de un método de *feature selection* conocido como **Sequential Forward Selection (SFS)** se definieron las **seis** variables mejor correlacionadas con la columna **'Popularidad'**, a saber: 'Modo', 'Volumen', 'Bailabilidad (%)', 'Hablado (%)', 'Acústico (%)', 'Instrumental (%)'. Las mismas se utilizaron para entrenar a los modelos de *machine learning*. El número óptimo de variables a introducir fue establecido tras pruebas sucesivas.

MODELADO

- Por último, se realizó *re-escalado* entre 0 y 1 de las seis variables independientes seleccionadas mediante el método **Robust Scaler**.



La gráfica de **'Acústico (%)'** impresiona haberse modificado ligeramente tras la *ingeniería de variables*, dado que presentaba previamente altas concentraciones de casos en ambos extremos de la distribución, mientras que ahora solo los conserva en el extremo inferior, configurando una curva *sesgada a derecha*.

MODELADO

- Se configuró un modelo de *aprendizaje supervisado* de **Árbol de Decisiones Clasificador (Decision Tree Classifier)**, a fin de ser empleado como modelo *benchmark,* es decir, de referencia para los modelos restantes.
- Los *hiperparámetros* se ajustaron automáticamente a través de un proceso conocido como *grid search*, usando *accuracy* como métrica *target*.
- Se aplicó una *estratificación* al momento de la separación de los subconjuntos de entrenamiento (80%) y de prueba (20%) para garantizar mayor homogeneidad entre ambos.
- A su vez, se utilizó *cross validation* para dividir el conjunto de datos en múltiples subconjuntos, entrenar el modelo en diferentes combinaciones de estos subconjuntos y evaluar su desempeño con los datos restantes.

MODELADO

SpotiSci

ÁRBOL DE DECISIONES: MÉTRICAS DEL MODELO

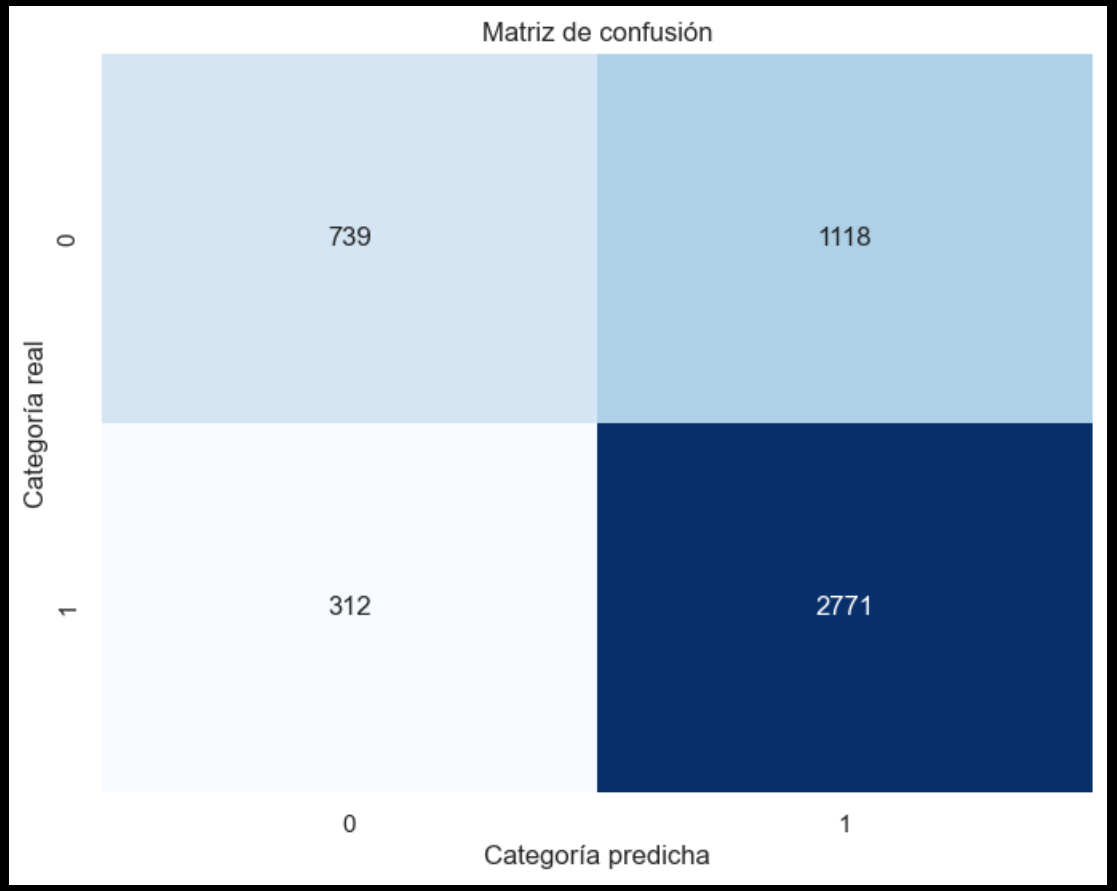
```
Training Accuracy: 0.7236282648309375
Testing Accuracy: 0.7105263157894737

Training Precision: 0.724762839385018
Testing Precision: 0.7125224993571613

Training Recall: 0.8983861811694104
Testing Recall: 0.898799870256244

Training F1 Score: 0.8022885283893395
Testing F1 Score: 0.7948938611589214

Training AUC: 0.7495679308151939
Testing AUC: 0.7327682108933402
```



MODELADO

- Además, se configuró otro modelo básico, uno de **Regresión Logística**. Obteniéndose, mediante el mismo proceso, un rendimiento ligeramente **inferior**.
- Un modelo más complejo, de tipo ***Random Forest Classifier***, consistente en una combinación de árboles de decisión, se confeccionó a continuación.
- En cuarto lugar, se aplicó un ***Stacking Model*** que combinaba los dos primeros con la intención de mejorar la *performance* general y se ajustaron sus hiperparámetros por *grid search* como en todos los anteriores.
- Por último, se instanció un modelo de **Redes Neuronales (*Deep Learning*)**, consistente en tres capas de *perceptrones*: una de entrada (de 64 unidades), una oculta (de 32) y una de salida (de neurona única). En este caso, por limitaciones de recursos, no fue posible realizar *tuning* de hiperaparámetros ni validación cruzada.

MODELADO

SpotiSci

COMPARACIÓN DE MODELOS

	Algoritmo	CV score promedio	Accuracy	Precision	Recall	F1 Score	AUC
0	Árbol de Decisiones	0.701532	0.710526	0.712522	0.898800	0.794894	0.732768
1	Regresión Logística	0.675618	0.688664	0.706055	0.858579	0.774883	0.717146
2	Random Forest Classifier	0.711372	0.726721	0.726063	0.902692	0.804800	0.760231
3	Stacking Model	0.688414	0.718219	0.721625	0.892961	0.798202	0.736402
4	Red Neuronal (Deep Learning)	NaN	0.723887	0.731984	0.879663	0.799057	0.748802

Random Forest Classifier es el modelo de aprendizaje automático que **mejor** se adapta a la tarea de predecir la *popularidad* de los temas en base a la serie restringida de características intrínsecamente musicales con que se alimenta. Supera al resto de los modelos en prácticamente todas las métricas, a excepción de *precision* (precisión), en que es superado por Red Neuronal (*Deep Learning*) en alrededor de 1%.

MODELADO

RANDOM FOREST CLASSIFIER: MÉTRICAS DEL MODELO

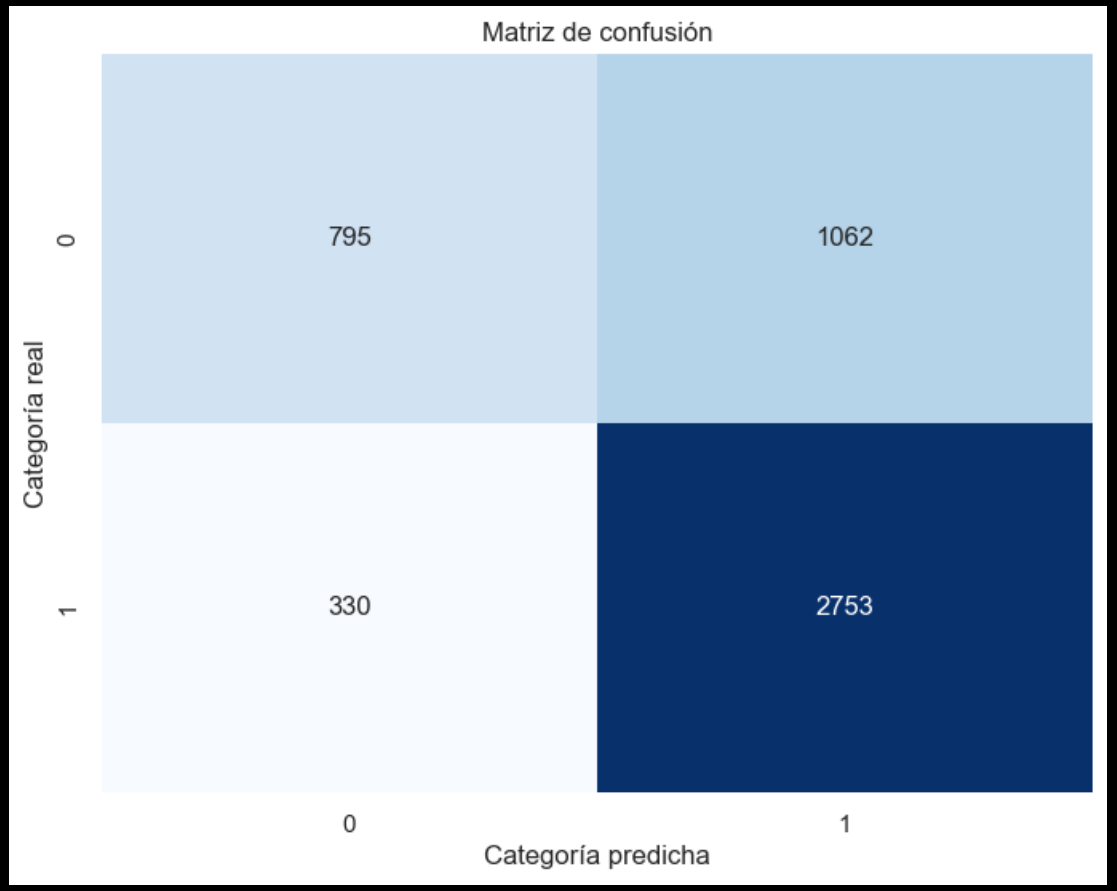
```
Training Accuracy: 0.7325875683336708
Testing Accuracy: 0.7182186234817813

Training Precision: 0.7315374507227332
Testing Precision: 0.7216251638269987

Training Recall: 0.9029275808936826
Testing Recall: 0.8929614012325657

Training F1 Score: 0.8082465246270553
Testing F1 Score: 0.7982023775007249

Training AUC: 0.7618159537144209
Testing AUC: 0.7364015775359551
```



MODELADO

RANDOM FOREST CLASSIFIER: MÉTRICAS DEL MODELO

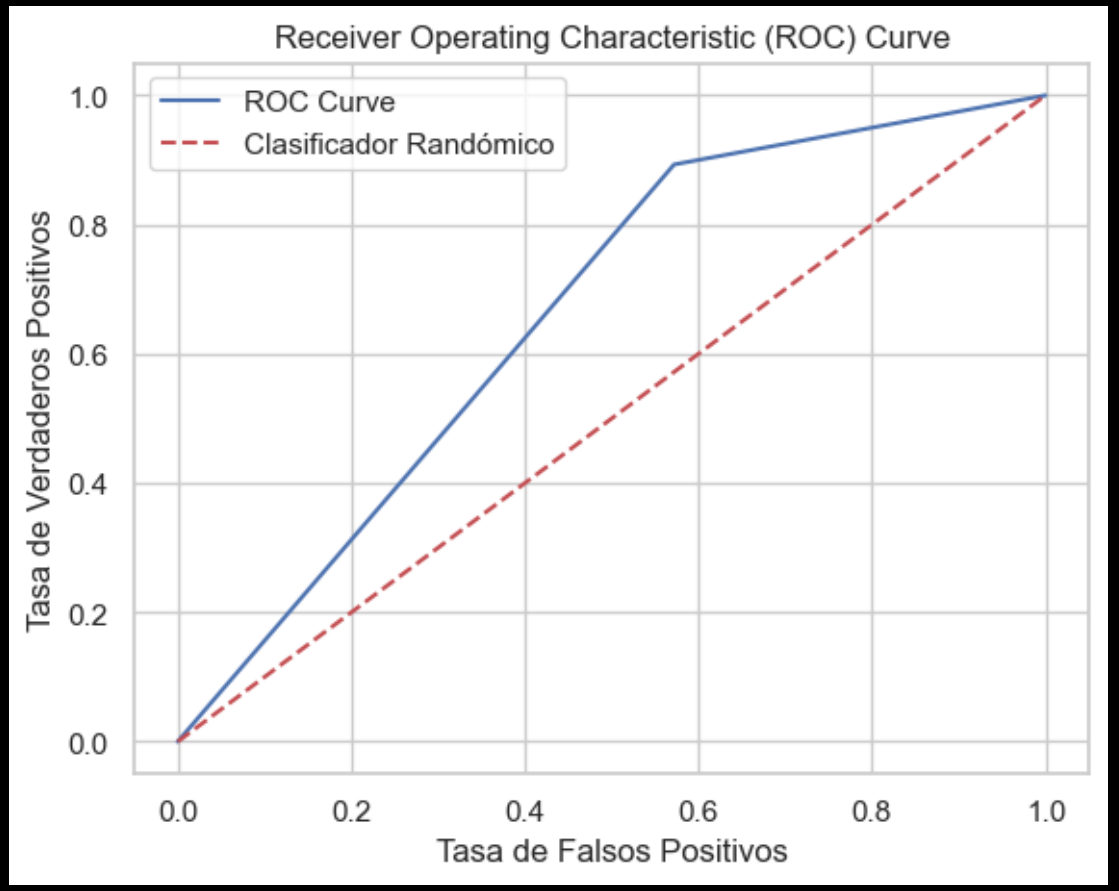
```
Training Accuracy: 0.7325875683336708
Testing Accuracy: 0.7182186234817813

Training Precision: 0.7315374507227332
Testing Precision: 0.7216251638269987

Training Recall: 0.9029275808936826
Testing Recall: 0.8929614012325657

Training F1 Score: 0.8082465246270553
Testing F1 Score: 0.7982023775007249

Training AUC: 0.7618159537144209
Testing AUC: 0.7364015775359551
```



CONCLUSIÓN

- Si bien son numerosos los factores ***extra-musicales*** que influyen en la *popularidad* de una canción publicada en un servidor de *streaming*, el presente trabajo demuestra que ciertas características ***intrínsecamente musicales*** (como por ejemplo el volumen o la percepción de "bailabilidad") presentan una relevancia significativa desde el punto de vista estadístico en relación con esa popularidad, y podrían ser utilizados para **predecir** si un tema tiene **potencial** para devenir un éxito comercial o no.
- No obstante, también debe destacarse que, ya sea por los motivos previamente expuestos, por limitaciones cuantitativas o cualitativas del *dataset* o por limitaciones en el modelado, **ninguno** de los modelos de aprendizaje automático desarrollados consiguió arrojar predicciones de una exactitud **satisfactoria**, conservando todos ellos *accuracies* **por debajo de 0,75**.
- Quizás, en futuros trabajos, la incorporación de nuevas variables que representen factores ***extra-musicales*** (por ejemplo: sello discográfico, inversión en publicidad, etc.) podría mejorar sensiblemente las métricas y conducirnos al desarrollo de modelos predicitivos más robustos.

SpotiSci

MUCHAS GRACIAS