

¿Qué características hacen que los tracks de Spotify sean éxitos?

by Laverne Evelyn & Mastrodonato Aldana

Agenda

1. Contexto y problema
2. Hipótesis/Preguntas de Interés
3. Metadata
4. Análisis Exploratorio
5. Insights y Recomendaciones
6. Aplicación de algoritmos de ML
7. Modelo elegido
8. Conclusiones finales



Contexto y problema

Contexto comercial

- La cantidad de tracks en Spotify aumento a lo largo de los años.
- No todos son éxitos o poseen gran popularidad entre la gente.
- Se quiere saber si ciertas características influyen en la popularidad de los mismos.
- Se busca identificar patrones en los tracks más populares.

Problema comercial

¿Existen patrones particulares en los tracks que puedan ser indicativos de éxitos?

Contexto analítico

- Obtuvimos un dataframe con tracks y sus principales características, a través de APIs públicas y de Kaggle.
- Realizamos un análisis exploratorio y la limpieza del mismo.
- Analizamos cómo evoluciona la popularidad a lo largo de los años
- Analizamos cómo afectan ciertas características a la popularidad.

Preguntas de interés

Pregunta principal

¿Existen patrones particulares en los tracks que puedan ser indicativos de éxitos?

Preguntas secundarias

- ¿Qué características de los tracks afectan la popularidad?
- ¿A lo largo de los años aumento la popularidad?
- ¿Cómo afecta la duración de un track en su popularidad?
- ¿Cómo afecta el instrumentalness (si el track contiene palabras habladas o no) en la popularidad?
- ¿Cómo afecta la danceability (que tan adecuado es el track para bailar) en la popularidad?
- ¿Las canciones explícitas son más populares que las canciones no explícitas?
- ¿Existe algún patrón en las canciones más populares en cuanto a las variables acousticness, danceability, instrumentalness, energy y valence?



Metadata



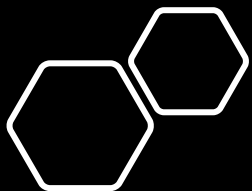
Variables

- Id
- Name
- Artista
- Duration_min
- Explicit (contenido considerado ofensivo o no apto para niños)
- Year
- Popularity (variable objetivo)
- Danceability (qué tan apto es un track para bailar)
- Energy
- Key (tono de un tema)
- Loudness (volumen general de un track en decibeles)
- Mode (tipo de escala en que la melodía se reproduce)
- Speechiness (detecta la presencia de palabras habladas en un track)
- Acousticness
- Instrumentalness (si un track es no vocal)
- Liveness (detecta la presencia de una audiencia en la grabación del track)
- Valence (describe la positividad musical de un track)
- Tempo (velocidad o el ritmo de un track).



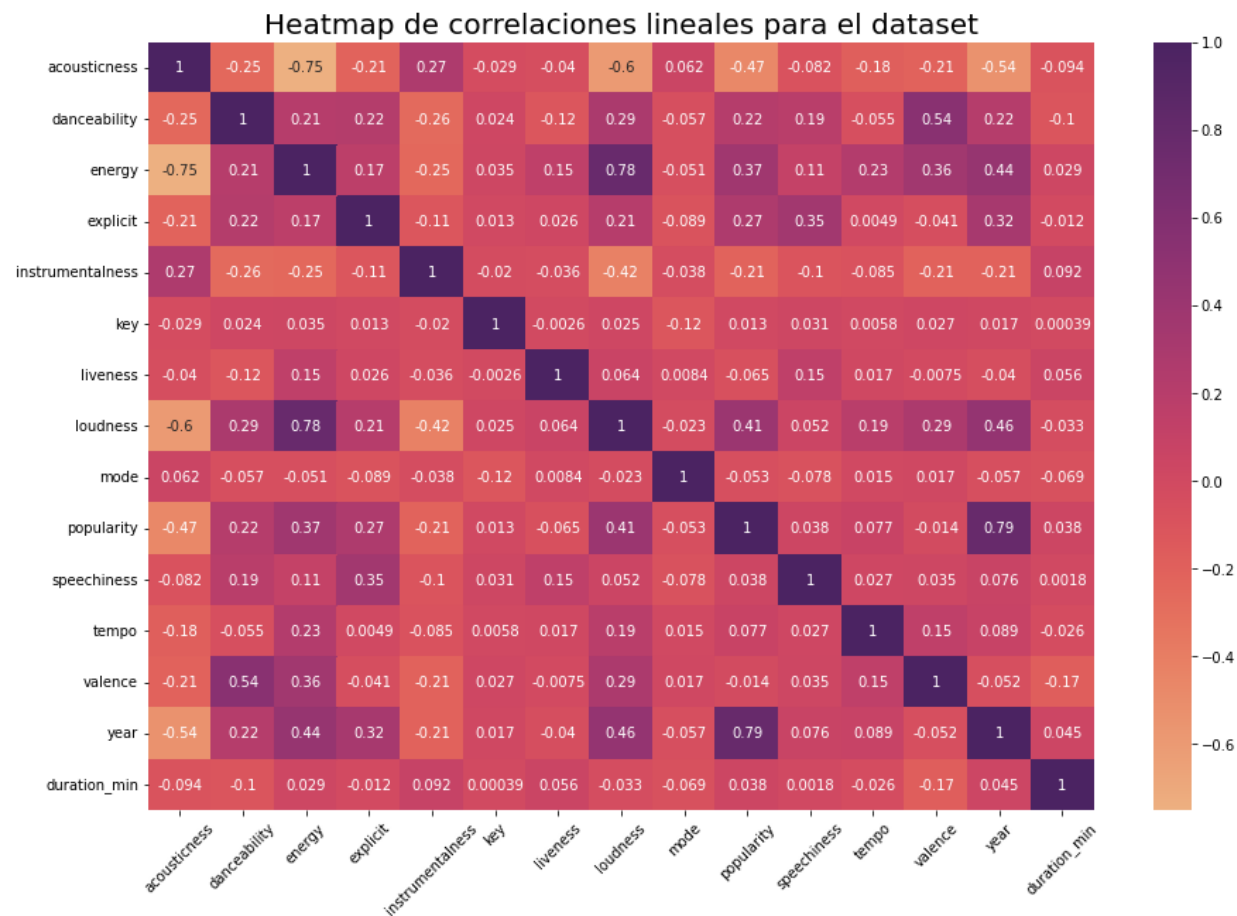
Análisis exploratorio

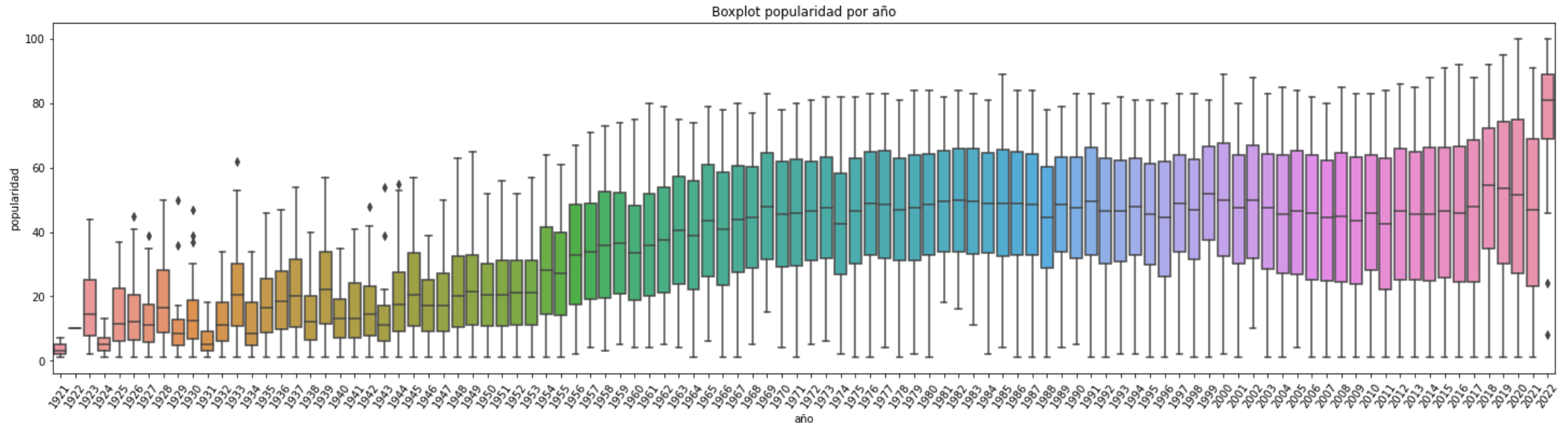




¿Qué características de los tracks afectan la popularidad?

- Correlación positiva con popularity: Danceability, explicit, loudness y energy (a medida que aumentan en valor, aumenta la popularidad).
- Correlación negativa con popularity: Acousticness e instrumentalness.



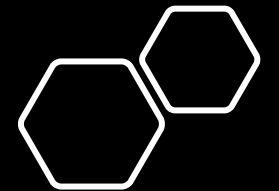


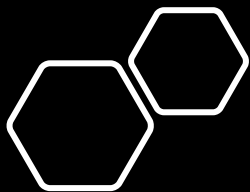
¿A lo largo de los años aumento la popularidad?

- A lo largo de los años la popularidad se ha ido incrementando.
- La popularidad se mantiene consistente a lo largo de las décadas
- Los años 2021 y 2022 presentan un comportamiento sustancialmente diferente a los años anteriores por poseer datos insuficientes, por lo que resulta atípico.

→ Agrupamos los tracks por décadas.

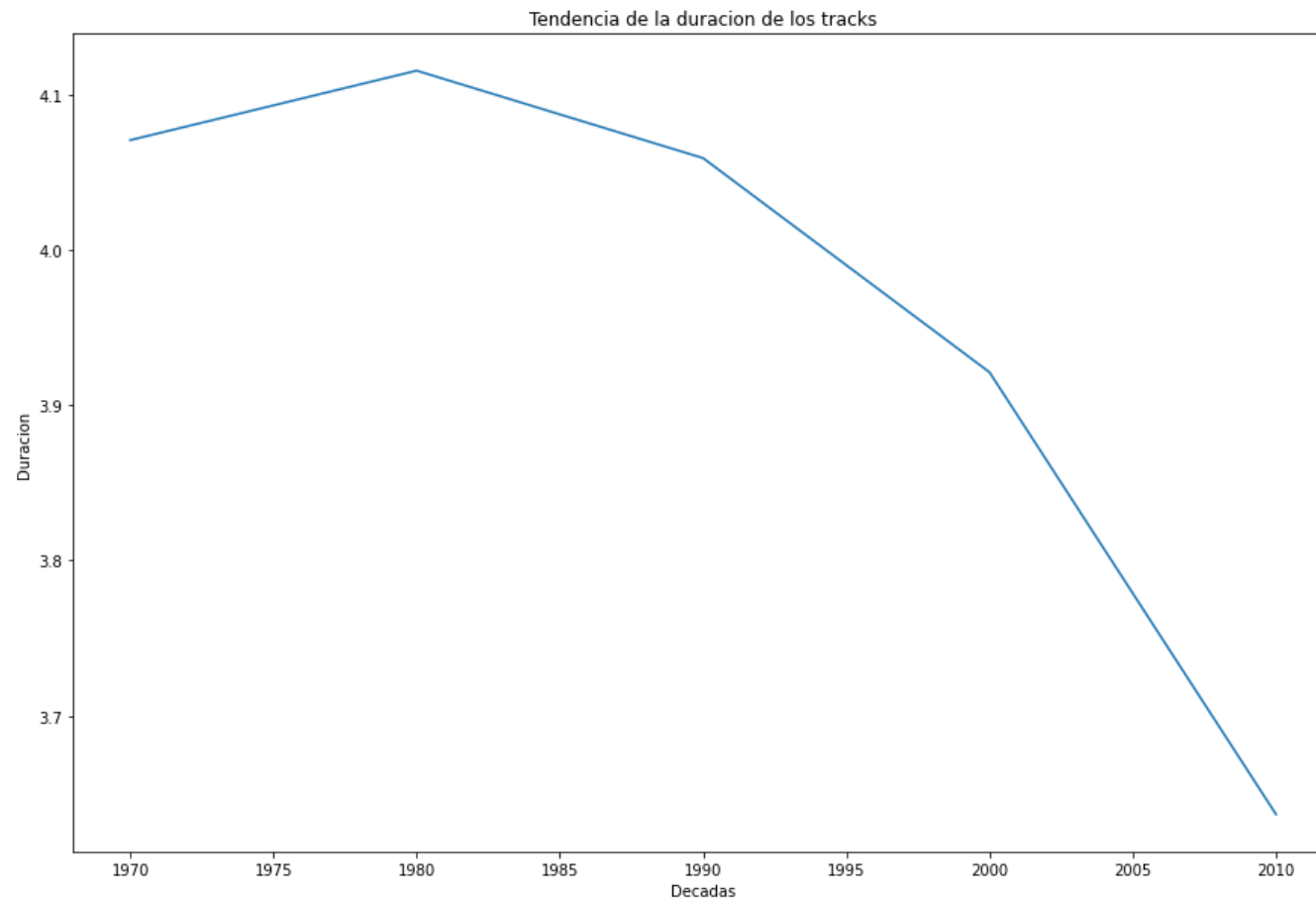
→ Excluimos los años anteriores a 1970, ya que no son significativos, y los años 2021 y 2022 por falta de información.

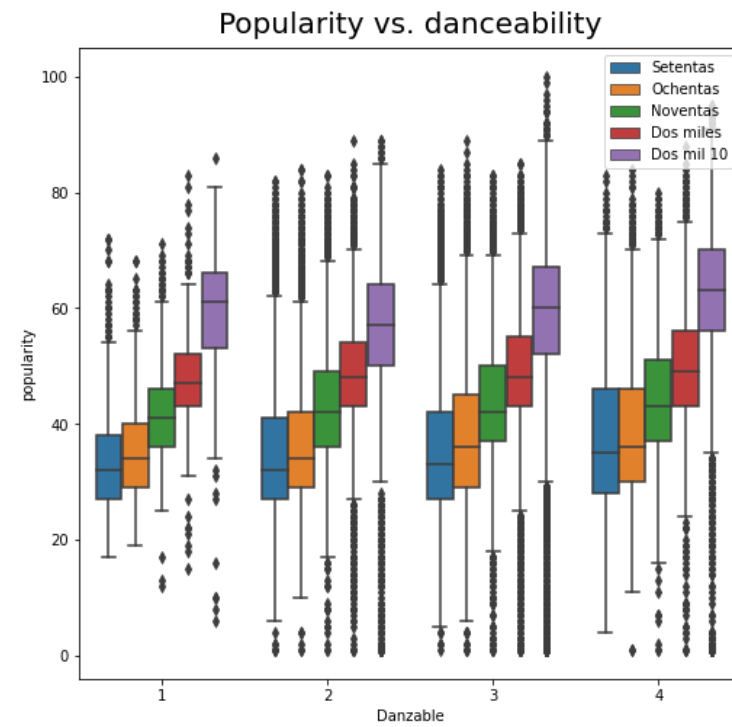
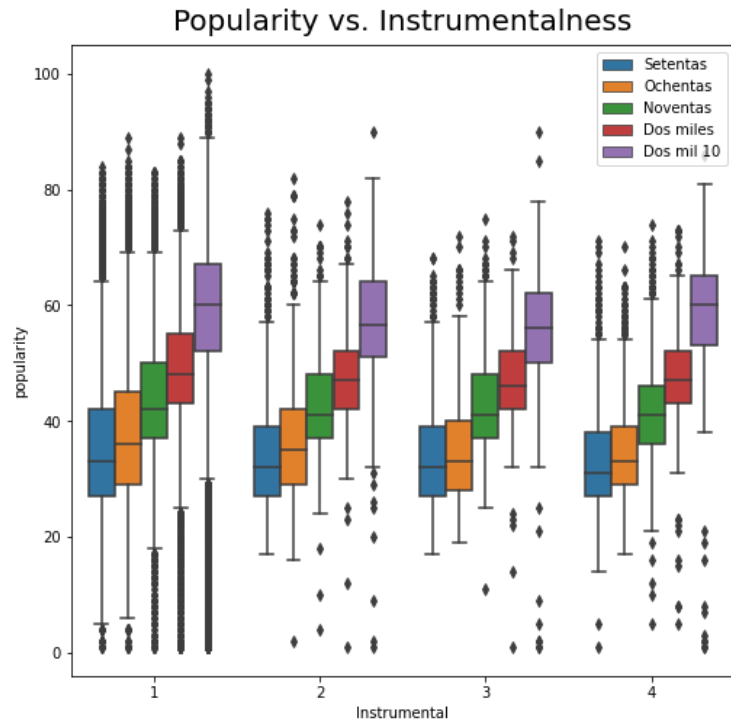




¿Cómo afecta la duración de un track en su popularidad?

- Valores atípicos para la duración en minutos: entre 10 y 30 minutos, se excluyeron del análisis.
- Tendencia a duraciones más cortas: entre 0 y 5 min.
- El promedio de los tracks va disminuyendo a lo largo de las décadas.
- No se puede concluir que la duración de los tracks afecte significativamente a la popularidad.



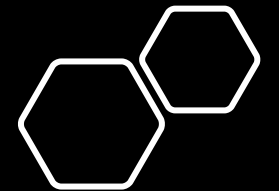


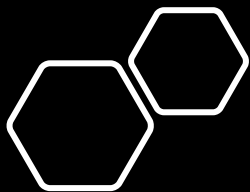
¿Cómo afecta el instrumentalness y la danceability en la popularidad?

- Tracks más populares → niveles de danzabilidad altos (mayores a 0.5 en general).
- Tracks más populares → instrumentalness menor a 0.25.
- Parecería que hay presencia de outliers pero los valores se encuentran en los rangos normales que pueden tomar las variables.
- Los valores que parecen atípicos, en realidad son los tracks más populares, de los que se busca conocer las características.

Instrumentalness → si el track contiene palabras habladas o no

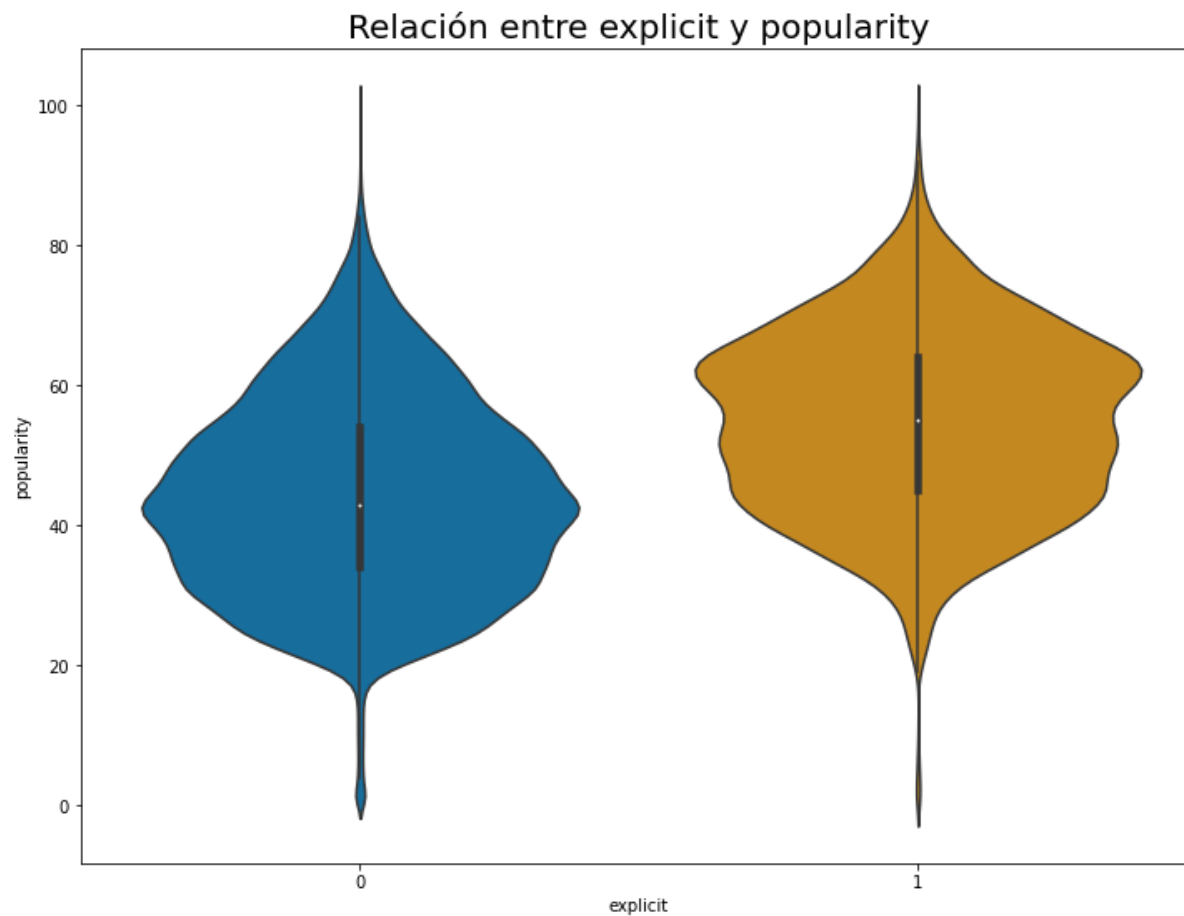
Danceability → que tan adecuado es el track para bailar

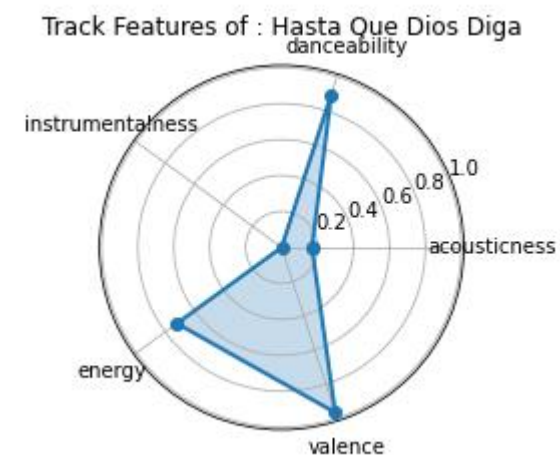
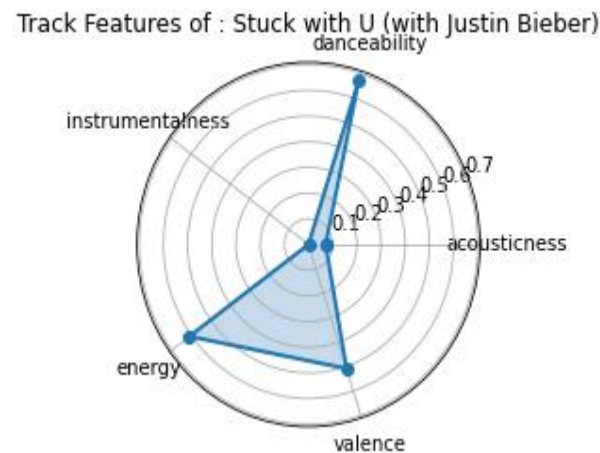
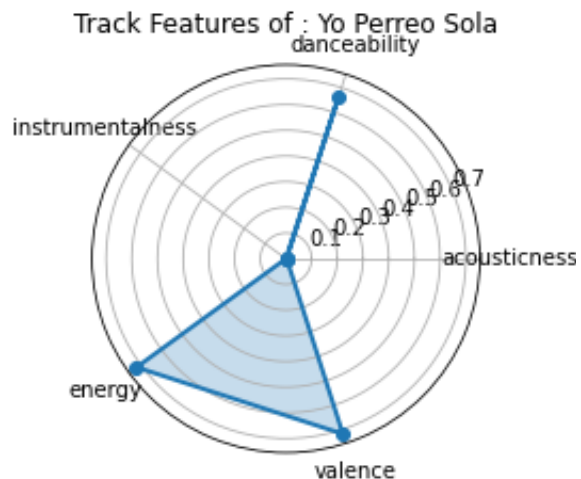
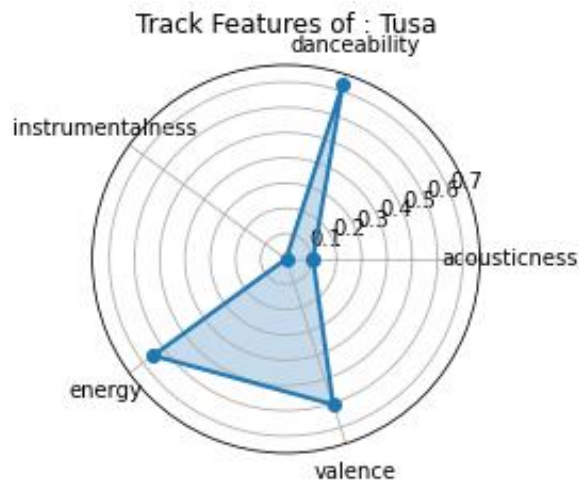




¿Las canciones explícitas son más populares que las canciones no explícitas?

- Tracks explícitos → Mayor concentración en rangos de popularidad altos (mediana 60/80)
- Tracks no explícitos → Mayor concentración en rangos medios. (Mediana 40/60)





¿Existe algún patrón en las canciones más populares y las menos populares en cuanto a las variables acousticness, danceability, instrumentalness, energy y valence?

Elegimos al azar algunas de las canciones más populares para graficar sus principales características (acousticness, danceability, instrumentalness, energy y valence):

- Tusa
- Yo Perreo Sola
- Stuck with U (with Justin Bieber)
- Hasta Que Dios Diga

De los gráficos de radar, se puede observar que los tracks más populares poseen características similares, se puede observar un patrón.

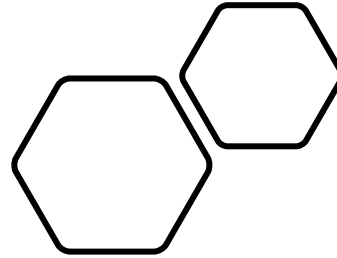




Insights y Recomendaciones



Observaciones y recomendaciones



- A lo largo de los años la popularidad se ha ido incrementando.


→ Agrupamos los tracks por décadas (comportamiento consistente)

→ Excluimos los años anteriores a 1970, ya que no son significativos, y los años 2021 y 2022, por tener poca información/incompleta.

→ Se excluyen los tracks con popularidad 0 (La variable objetivo no puede ser 0).

- Duración → Tendencia a tracks más cortos (entre 0 y 5 min).
- No se puede concluir que la duración de los tracks afecte significativamente a la popularidad.
- Tracks más populares → niveles de danzabilidad altos (mayores a 0.5 en general) e instrumentality menor a 0.25.
- Tracks explícitos → Mayor concentración en rangos de popularidad altos (mediana 60/80)
- Tracks no explícitos → Mayor concentración en rangos medios (mediana 40/60).
- Tracks más populares → Poseen características similares, se puede observar un patrón.

Buscamos realizar un modelo que logre predecir un éxito (nivel de popularidad alto) a partir de las características más relevantes encontradas.



Aplicación de algoritmos de ML



Aplicación de modelos de ML – Resumen de Resultados

A continuación, se presenta un resumen de los modelos que se probaron, con los diferentes parámetros y los % de train/test empleados, junto con los resultados obtenidos

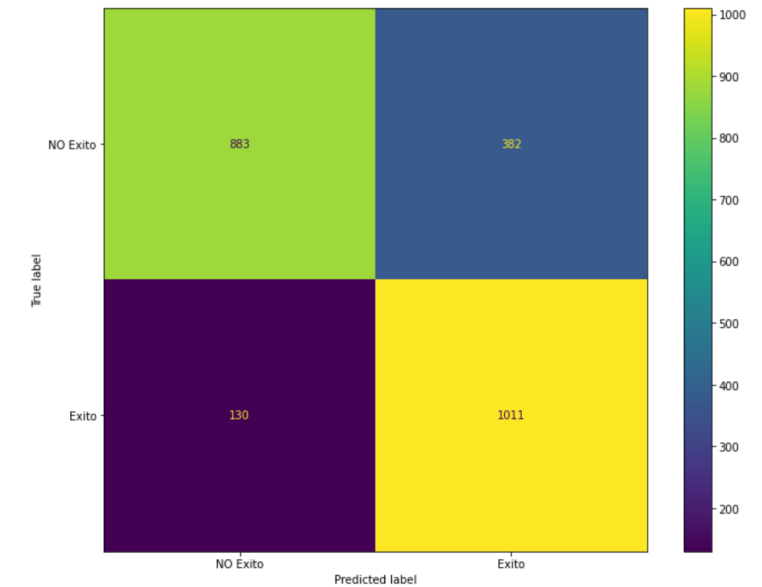
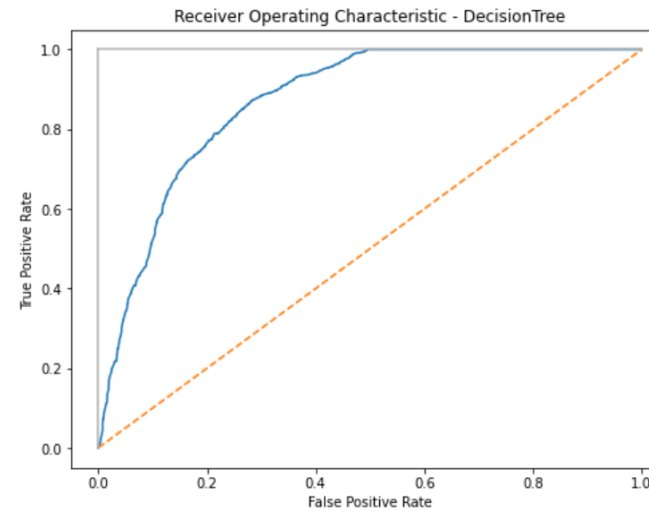
ID	Modelo	Profundidad árboles	%Train/%Test	# Variables	Accuracy	roc_auc_score	Clase	Precision	Recall	F1 Score
1	Random forest	Default (random_state=1)	70/30	12	0,6646	0,7157	0 (no éxito) 1 (éxito)	0,6700 0,6561	0,7100 0,6083	0,6900 0,6313
2	Random forest	Default (random_state=1)	70/30	5 principales (PCA)	0,6135	0,6523	0 (no éxito) 1 (éxito)	0,6300 0,5969	0,6600 0,5578	0,6400 0,5767
3	Random forest	Default (random_state=1)	80/20	12	0,6739	0,7241	0 (no éxito) 1 (éxito)	0,6800 0,6662	0,7300 0,6128	0,7000 0,6384
4	Random forest	Default (random_state=1)	80/20	5 principales (PCA)	0,6273	0,6650	0 (no éxito) 1 (éxito)	0,6400 0,6065	0,6600 0,5885	0,6500 0,5974
5	Random forest	Default (random_state=1)	90/10	12	0,6899	0,7457	0 (no éxito) 1 (éxito)	0,6900 0,6868	0,7400 0,6363	0,7100 0,6606
6	Random forest	Default (random_state=1)	90/10	5 principales (PCA)	0,6318	0,6741	0 (no éxito) 1 (éxito)	0,6400 0,6175	0,6700 0,5872	0,6600 0,6020
7	Random forest	Ajuste de parámetros	90/10	12	0,6712	0,7300	0 (no éxito) 1 (éxito)	0,6600 0,6815	0,7600 0,5758	0,7100 0,6242
8	SVM	Default (random_state=1)	70/30	12	0,5567	0,6316	0 (no éxito) 1 (éxito)	0,5500 0,7117	0,9600 0,0978	0,7000 0,1720
9	SVM	Default (random_state=1)	80/20	12	0,5349	0,6302	0 (no éxito) 1 (éxito)	0,5300 0,6340	0,9800 0,0300	0,6900 0,0573
10	SVM	Default (random_state=1)	90/10	12	0,5354	0,6291	0 (no éxito) 1 (éxito)	0,5300 0,6900	0,9900 0,0298	0,6900 0,0572
11	XGBClassifier	random_state=7	70/30	12	0,6416	0,6933	0 (no éxito) 1 (éxito)	0,6500 0,6318	0,7000 0,5715	0,6800 0,6001
12	XGBClassifier	random_state=7	90/10	12	0,6324	0,6788	0 (no éxito) 1 (éxito)	0,6300 0,6305	0,7200 0,5357	0,6700 0,5793
13	GradientBoostin	random_state=7	70/30	12	0,6422	0,6921	0 (no éxito) 1 (éxito)	0,6500 0,6330	0,7100 0,5710	0,6800 0,6004



Modelo elegido

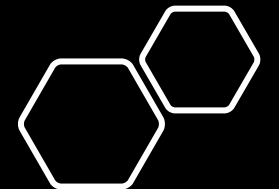


	precision	recall	f1-score	support
0	0.87	0.70	0.78	1265
1	0.73	0.89	0.80	1141
accuracy			0.79	2406
macro avg	0.80	0.79	0.79	2406
weighted avg	0.80	0.79	0.79	2406



Modelo elegido

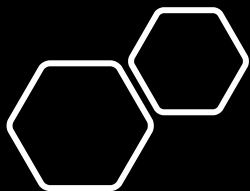
- De todos los modelos de clasificación, el que obtuvo las mejores métricas fue el random forest, con los parámetros en default (random_state=1), entrenando 90%/10% y con la totalidad de las 12 variables.
- Luego, creamos una variable sintética que cuente la cantidad de hits que tienen los artistas para poder mejorar el rendimiento del modelo. Entendemos que ciertos artistas tienen más popularidad que otros y esto podría ser una variable explicativa.
- A continuación, se presentan los resultados del modelo elegido.





Conclusiones finales





Conclusiones finales

- Utilizamos la última década (años 2010 a 2020) por su significatividad junto con las variables más relevantes obtenidas del EDA.
- Entrenamos varios modelos de ML → Random Forest, SVM, XGBClassifier y GradientBoost. A su vez, diferentes parámetros y % de train/test.
- Mejores métricas → Random forest, parámetros en default (random_state=1), 90%/10% y 12 variables.
- Creamos nueva variable → Cantidad de hits por artistas para poder mejorar el rendimiento del modelo.
- Modelo final → Random forest con la nueva variable mejoró significativamente.
- Accuracy: 78,72%, ROC: 86,98%, Precisión: 73% (éxito) y 87% (no éxito), f1 score: 80% (éxito) y 78% (no éxito)
- La variable 'cant de hits' explica gran parte del modelo como habíamos supuesto.
- PCA con 5 primeros componentes → Accuracy 64,46%, más bajo de lo que obtuvimos anteriormente.
- Métodos de Cross Validation → Máximo accuracy 76,85%, más bajo que el obtenido en el random forest realizado anteriormente.
- Mejores parámetros del Cross Validation y quitando las variables categóricas → Accuracy de 76,47%, ROC de 85,71% y tanto la precisión como el f1 score para éxitos disminuyen respecto al modelo anterior.
- **Mejor modelo** → Random forest con random_state=1, parámetros en default y 13 variables.