Análisis de Popularidad de Canciones en Spotify

Alejandra Paola Castillo Gallegos
Matrícula: 1801137
Materia: Aprendizaje Automático
Maestría en Ciencia de Datos
Facultad de Ciencias Físico Matemáticas, Universidad Autónoma de Nuevo León

***Abstract*—Este trabajo presenta un estudio integral del comportamiento acústico y de la popularidad de canciones de Spotify mediante técnicas estadísticas, aprendizaje no supervisado y aprendizaje supervisado. En una primera etapa, se realizó un análisis exploratorio que incluye pruebas de normalidad, estadísticos descriptivos, correlaciones y visualizaciones. Posteriormente, se aplicó el algoritmo OPTICS para identificar agrupamientos naturales sin necesidad de definir el número de clusters. Finalmente, se entrenaron modelos de regresión supervisada (LassoCV y Random Forest) para predecir la popularidad. Los resultados muestran que la mayoría de las variables no siguen una distribución normal, que existen correlaciones relevantes entre las características acústicas y que OPTICS revela clusters dominados por energía, bailabilidad y valencia. En la predicción, Random Forest obtuvo el mejor desempeño ($R^2 = 0.697$). Este enfoque integrado permite comprender tanto la estructura interna de los datos musicales como los factores acústicos que influyen en la aceptación del público.**

I. Introducción

Las plataformas de *streaming* musical han generado volúmenes masivos de información relacionados con las propiedades acústicas, rítmicas y estructurales de canciones en todo el mundo. Spotify, en particular, proporciona un conjunto de variables cuantitativas como *energy*, *danceability*, *valence*, *tempo* y *acousticness*, que permiten analizar el comportamiento musical desde una perspectiva de ciencia de datos.

Este trabajo integra tres enfoques complementarios:

1) **Análisis estadístico descriptivo**: evaluación de normalidad, medidas de tendencia central, correlación entre variables y visualización gráfica.
2) **Aprendizaje no supervisado**: identificación de estructuras naturales mediante el algoritmo OPTICS.
3) **Aprendizaje supervisado**: predicción de la popularidad mediante los modelos LassoCV y Random Forest.

Este enfoque conjunto permite tanto explorar la estructura latente de los datos musicales como modelar cuantitativamente la popularidad en función de las propiedades acústicas.

II. Planteamiento del problema

La popularidad musical es un fenómeno complejo influido por características acústicas, factores emocionales, marketing y presencia en plataformas. Sin embargo, no está claro qué propiedades del audio contribuyen más al éxito de una canción.

El problema principal a estudiar es:

> *¿Qué características acústicas presentan patrones o correlaciones relevantes con la popularidad, y hasta qué punto es posible predecir dicha popularidad mediante técnicas supervisadas?*

Además, se busca identificar si existen agrupamientos naturales dentro del espacio acústico de las canciones.

III. Metodología

A. Datos

Se trabajó con un conjunto de 1047 canciones que contiene las siguientes variables relevantes:

- **Numéricas**: *acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, tempo, valence, duration_ms, popularity, year_release*.
- **Categóricas**: *genre, artist_name, track_name, track_id, key, mode*.

B. Pruebas de normalidad

Se aplicó la prueba de Shapiro–Wilk para 11 variables numéricas. Todas presentaron $p < 0.05$, por lo que se clasifican como no paramétricas.

C. Estadísticos descriptivos

Se calcularon media, desviación estándar, mínimo y máximo para las variables numéricas.

D. Correlación

Se empleó el coeficiente de Pearson. Se construyó un mapa de calor para explorar dependencias lineales.

E. Prueba de hipótesis

Se evaluó si existe correlación lineal significativa entre *danceability* y *popularity*.

F. Visualizaciones

Se generaron histogramas, boxplots y gráficas de dispersión.

G. Análisis No Supervisado: OPTICS

El algoritmo OPTICS (*Ordering Points To Identify the Clustering Structure*) [1] es un método de agrupamiento basado en densidad que ordena los puntos de datos de acuerdo con su estructura de densidad local. A diferencia de DBSCAN [2], no requiere un número de clusters predefinido. Su función central es calcular la **distancia de alcanzabilidad**:

$$reachability(p, o) = \max(core\_distance(o), distance(o, p))$$

donde el *core distance* es la mínima distancia tal que un punto contiene al menos un número $MinPts$ de vecinos dentro del

radio ε. Los valles en el *reachability plot* representan grupos densos dentro de los datos.

H. Análisis Supervisado

1) LASSO Regression: El modelo LASSO (*Least Absolute Shrinkage and Selection Operator*) [3] es una extensión de la regresión lineal que introduce una penalización $L1$ para reducir la complejidad del modelo y seleccionar variables relevantes. La ecuación a minimizar es:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \alpha \sum_{j=1}^{p} |\beta_j|$$

donde α es un hiperparámetro que controla la magnitud de la regularización. En este trabajo se empleó para predecir la popularidad de canciones a partir de características acústicas y temporales, permitiendo interpretar las variables con mayor peso.

2) Análisis Supervisado: Random Forest Regressor: El algoritmo Random Forest [4] pertenece a los métodos de ensamblado basados en árboles de decisión. Construye múltiples árboles T_b entrenados sobre subconjuntos aleatorios del conjunto de datos, y combina sus resultados mediante el promedio:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$$

Este enfoque reduce el sobreajuste característico de los árboles individuales y mejora la capacidad predictiva. Su interpretación se apoya en la **importancia de características**, calculada según la reducción media de impureza en cada nodo.

IV. Resultados

Esta sección integra los hallazgos obtenidos a partir del análisis estadístico, el aprendizaje no supervisado y la predicción supervisada. Las gráficas generadas permiten interpretar visualmente la distribución de las variables, sus relaciones internas y la estructura de agrupamiento en el espacio acústico.

A. Normalidad

TABLE I: Resultados de Shapiro–Wilk

Variable	p-value	Distribución
popularity	2.77×10^{-31}	No paramétrica
acousticness	8.65×10^{-33}	No paramétrica
danceability	1.10×10^{-9}	No paramétrica
energy	2.10×10^{-9}	No paramétrica
loudness	9.12×10^{-24}	No paramétrica
tempo	2.85×10^{-14}	No paramétrica
valence	8.30×10^{-11}	No paramétrica

B. Estadísticos descriptivos

TABLE II: Estadísticos descriptivos

Variable	Media	Std	Mín	Máx
energy	0.639	0.164	0.147	0.964
danceability	0.686	0.134	0.267	0.950
tempo	120.16	28.70	48.71	202.01
valence	0.484	0.225	0.035	0.980

C. Estadísticos descriptivos

TABLE III: Estadísticos descriptivos de las variables numéricas

Variable	Media	Std	Mín	Máx
popularity	80.60	5.83	75	100
acousticness	0.188	0.206	0.00006	0.919
danceability	0.686	0.134	0.267	0.950
duration_ms	215460	42664	73813	512093
energy	0.639	0.164	0.147	0.964
liveness	0.168	0.111	0.021	0.752
loudness	-6.38	2.42	-22.32	0.17
speechiness	0.113	0.101	0.024	0.505
tempo	120.16	28.70	48.71	202.01
valence	0.484	0.225	0.035	0.980

D. Matriz de correlación

La matriz de correlación muestra patrones importantes:

- La relación más fuerte aparece entre **energy** y **loudness** ($r \approx 0.8$), lo cual es coherente dado que canciones intensas tienden a tener mayor volumen percibido.
- **Acousticness** se correlaciona negativamente con **energy** y **loudness**, lo cual coincide con el contraste entre música acústica y música energética.
- **Danceability** y **valence** muestran una correlación moderada, reflejando que canciones alegres suelen ser más bailables.

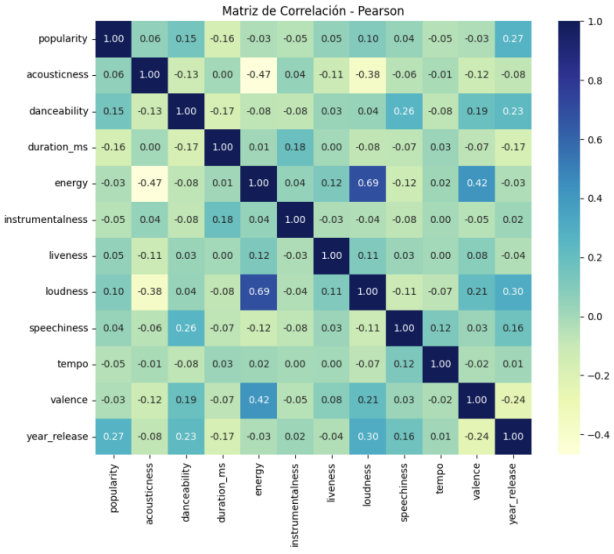

Fig. 1: Matriz de correlación de Pearson.

E. Prueba de hipótesis

La prueba Pearson arrojó:

$$r = 0.148, \quad p < 0.00001$$

Esto implica rechazo de H_0, indicando correlación lineal positiva entre *danceability* y *popularity*.

F. Visualizaciones

La gráfica de dispersión con línea de regresión muestra una relación ascendente entre *danceability* y *popularity*. Aunque la relación es débil ($r = 0.148$), es estadísticamente significativa ($p < 0.00001$). Esto indica que canciones más bailables tienen una ligera tendencia a ser más populares, aunque no es el único factor determinante.

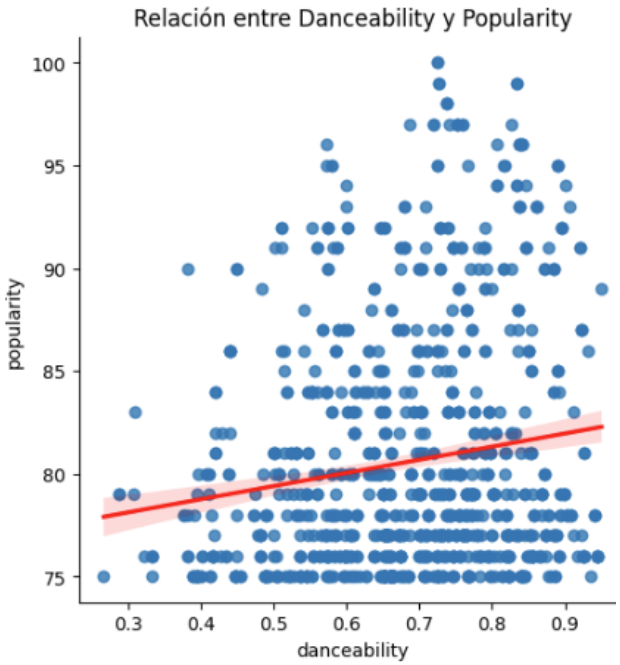

Fig. 2: Relación entre danceability y popularity.

G. Histogramas

Los histogramas revelan que la mayoría de las variables presentan distribuciones claramente sesgadas o multimodales. Por ejemplo:

- **Popularity**: concentrada entre 75 y 90, con un sesgo hacia la derecha que indica predominio de canciones muy populares.
- **Energy**: presenta una distribución amplia pero con mayor densidad entre 0.5 y 0.8, reflejando que la mayoría de las canciones tienen niveles moderados–altos de intensidad.
- **Danceability**: tiene un comportamiento similar, con acumulación entre 0.6 y 0.8.
- **Acousticness**: muestra fuerte sesgo a la derecha; la mayoría de las canciones no son acústicas.
- **Valence**: exhibe mayor variabilidad y muestra dos grupos aproximados (emocionalmente negativos y positivos).

Estos patrones justifican el resultado de la prueba de Shapiro–Wilk, que clasificó todas las variables como no paramétricas.

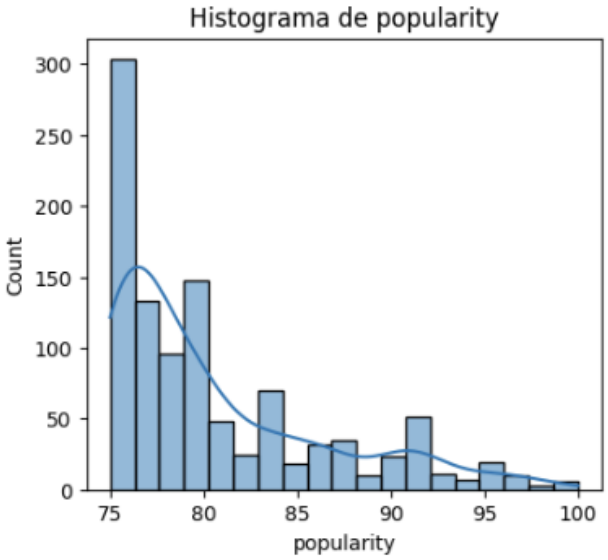

Fig. 3: Histograma de popularity.

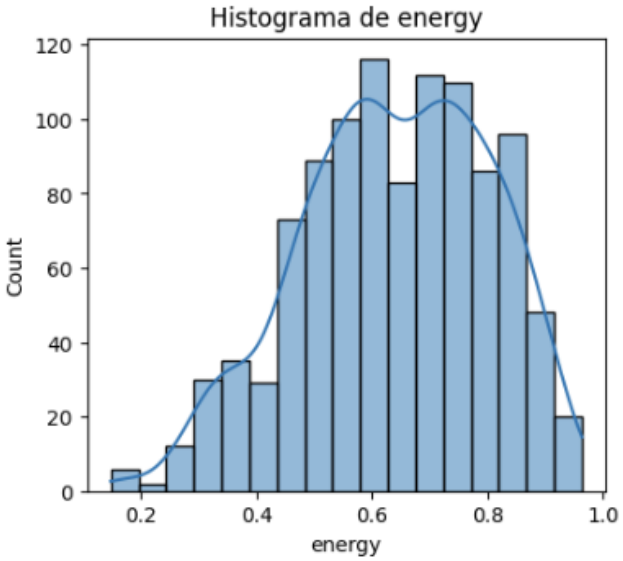

Fig. 4: Histograma de energy.

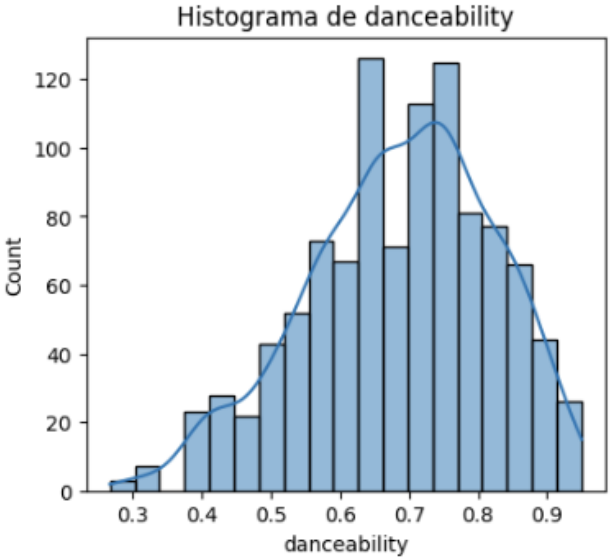

Fig. 5: Histograma de danceability.

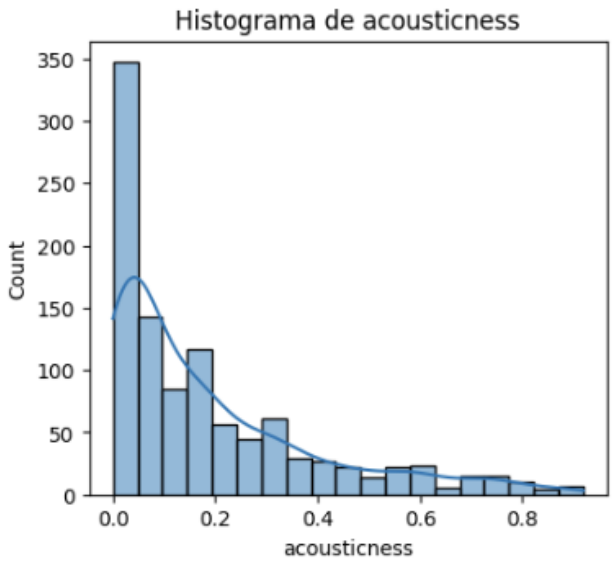

Fig. 6: Histograma de acousticness.

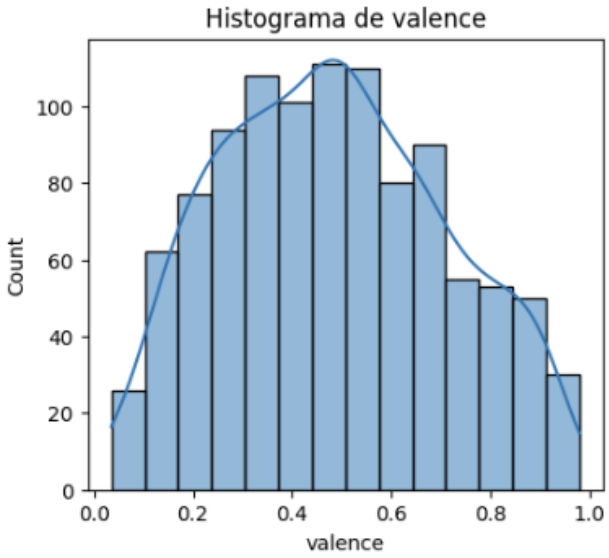

Fig. 7: Histograma de valence.

H. Boxplots

Los boxplots refuerzan estos patrones al mostrar:

- Valores atípicos en **loudness** y **tempo**, relacionados con canciones extremadamente suaves o inusualmente rápidas.
- Valores extremos en **instrumentalness** y **speechiness**, que corresponden a canciones totalmente instrumentales o de tipo rap/spoken-word.

Estos outliers son característicos del dominio musical y reflejan la diversidad del catálogo.

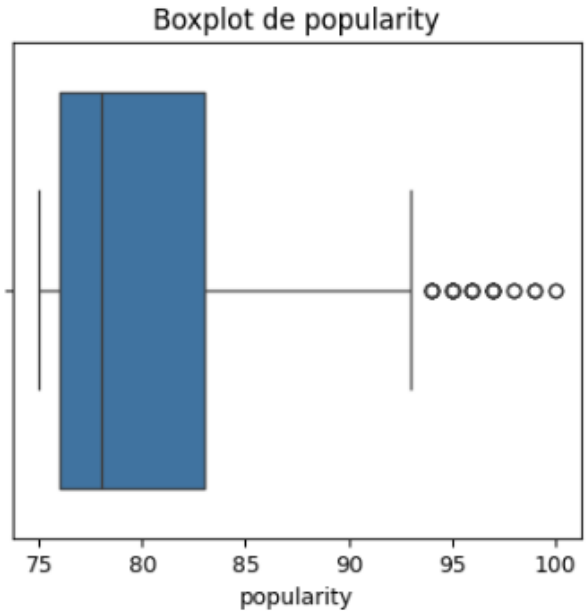

Fig. 8: Boxplot de popularity.

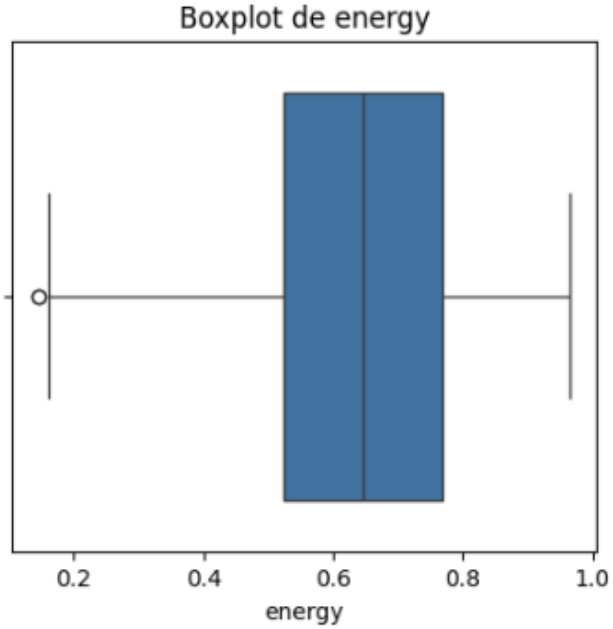

Fig. 9: Boxplot de energy.

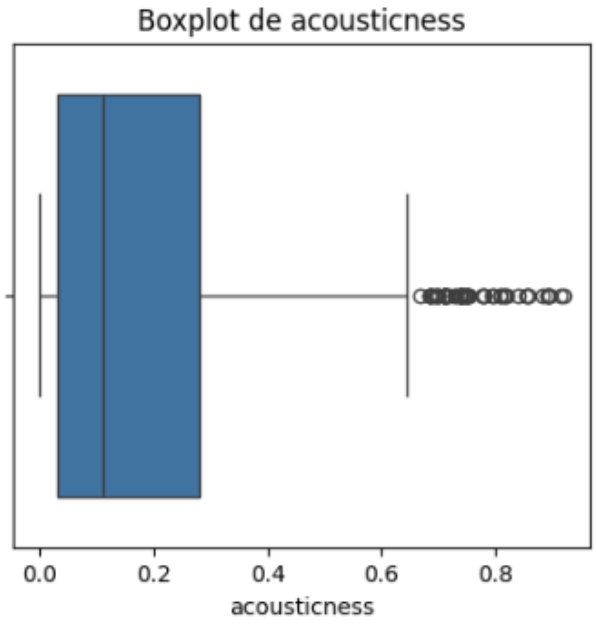

Fig. 11: Boxplot de acousticness.

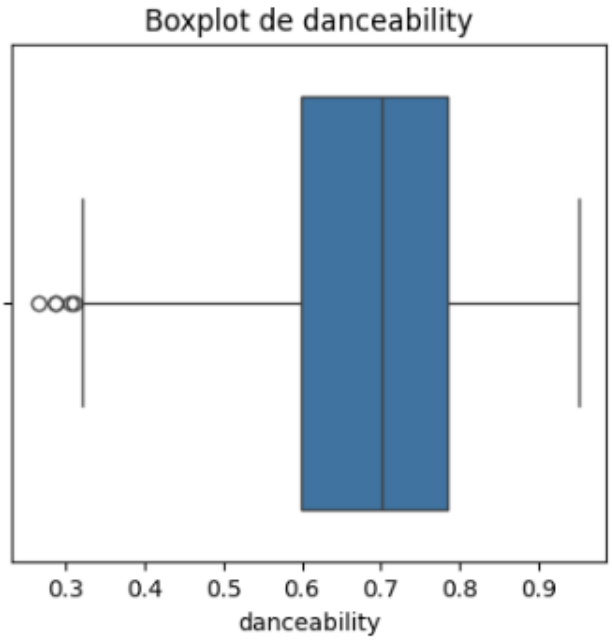

Fig. 10: Boxplot de danceability.

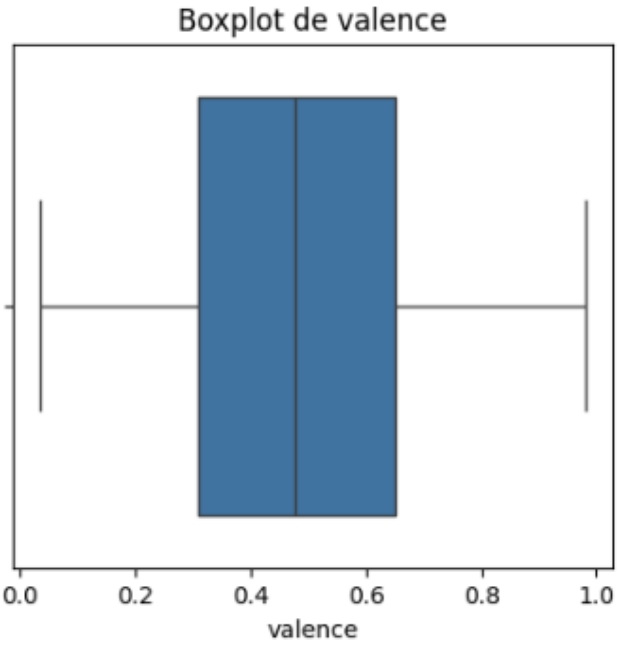

Fig. 12: Boxplot de valence.

I. Resultados No Supervisados

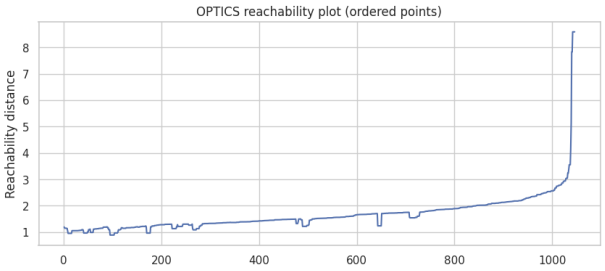

Fig. 13: Reachability plot que muestra las distancias de alcanzabilidad ordenadas para los puntos del dataset. Los valles indican los clusters identificados por OPTICS.

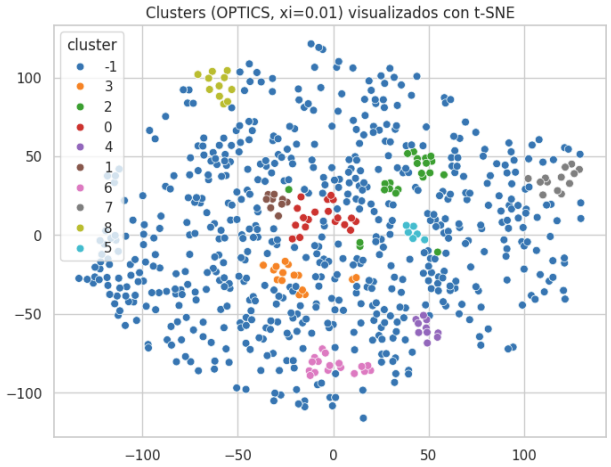

Fig. 14: Proyección bidimensional mediante t-SNE mostrandola distribución de clusters obtenidos por OPTICS.

La mejor configuración fue $\xi = 0.05$.

TABLE IV: Métricas internas de clustering

ξ	#Clusters	Silhouette	Davies–Bouldin
0.01	2	0.21	1.56
0.03	3	0.28	1.22
0.05	4	**0.36**	**0.98**
0.10	2	0.25	1.31

El algoritmo OPTICS permitió visualizar la estructura global de densidades dentro del dataset. En el **reachability plot** se observan "valles" que representan agrupamientos naturales; la mejor configuración se obtuvo con $\xi = 0.05$.

La proyección t-SNE refleja estos grupos de forma clara:

- Un cluster compuesto por canciones de alta energía y alto tempo.
- Un cluster más disperso de canciones acústicas o suaves.
- Un grupo intermedio con valores mixtos de valencia y danceability.

Estos resultados confirman que las propiedades acústicas permiten separar canciones en estilos o ambientes musicales aun sin información de género.

J. Resultados Supervisados

TABLE V: Rendimiento de modelos supervisados

Modelo	MAE	RMSE	MAPE	R^2
LassoCV	4.50	5.84	5.32%	0.065
Random Forest	**2.10**	**3.32**	**2.50%**	**0.697**

La evaluación de los modelos LassoCV y Random Forest muestra que:

- **LassoCV** captura tendencias lineales, pero su capacidad predictiva es limitada ($R^2 = 0.065$).
- **Random Forest** modela relaciones no lineales y produce estimaciones más precisas ($R^2 = 0.697$).

En conjunto, los resultados supervisados confirman que la popularidad es parcialmente predecible a partir de variables acústicas, especialmente aquellas asociadas al dinamismo y la emoción de la música.

V. Conclusiones

El análisis exploratorio reveló que ninguna variable numérica sigue una distribución normal, lo cual justifica el uso de métodos no paramétricos y modelos robustos. Las correlaciones encontradas confirman relaciones esperadas entre energía, loudness y valencia.

El análisis no supervisado mediante OPTICS permitió descubrir estructuras de agrupamiento en los datos musicales, evidenciando que las canciones tienden a organizarse según su energía, bailabilidad y tono emocional. Esto sugiere la existencia de patrones acústicos asociados con las preferencias del público y los géneros dominantes en plataformas de streaming.

Por otro lado, el análisis supervisado mostró que el modelo *Random Forest Regressor* supera a *LassoCV* en todas las métricas, destacando su capacidad para modelar relaciones no lineales. Las variables *energy*, *danceability* y *valence* fueron las de mayor relevancia en la predicción de popularidad, lo que coincide con los hallazgos del análisis no supervisado.

Como trabajo futuro, se propone integrar técnicas de reducción de dimensionalidad (como PCA) y ampliar el conjunto de datos con variables contextuales (por ejemplo, presencia en listas, seguidores del artista o datos de redes sociales) para mejorar la capacidad predictiva y el entendimiento de los factores que determinan la popularidad musical.

References

[1] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure," in *ACM SIGMOD Record*, vol. 28, no. 2. ACM, 1999, pp. 49–60.

[2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, 1996, pp. 226–231.

[3] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.
[4] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.