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 de aprendizaje no supervisado y supervisado. En una primera etapa, se aplicó el algoritmo OPTICS (*Ordering Points To Identify the Clustering Structure*) para identificar agrupamientos naturales en las características de audio sin necesidad de definir un número de clusters a priori. En una segunda etapa, se entrenaron modelos supervisados de regresión (LassoCV y Random Forest Regressor) para predecir la popularidad de las canciones a partir de sus propiedades acústicas. Los resultados muestran que OPTICS revela patrones asociados a energía, bailabilidad y valencia, mientras que Random Forest logra una predicción precisa de la popularidad ($R^2 = 0.697$), superando al modelo lineal penalizado. Este enfoque combinado permite comprender tanto la estructura interna de los datos musicales como los factores acústicos que influyen en la aceptación del público.**

I. Introducción

El crecimiento de las plataformas de *streaming* ha generado una enorme disponibilidad de datos musicales que incluyen variables sobre las propiedades acústicas, la energía, el ritmo y la aceptación de las canciones. Este tipo de información ofrece una oportunidad para aplicar técnicas de *aprendizaje automático* que permitan comprender las preferencias del público y predecir el éxito de nuevos lanzamientos.

El presente trabajo combina enfoques de aprendizaje no supervisado y supervisado con el objetivo de analizar y modelar los factores asociados a la popularidad musical. En primer lugar, se utiliza el algoritmo OPTICS, un método de agrupamiento basado en densidad que permite descubrir grupos con diferentes distribuciones sin necesidad de fijar un número de clusters, lo cual es especialmente útil para representar la diversidad de géneros o estilos musicales.

En segundo lugar, se emplean modelos supervisados de regresión —LassoCV y Random Forest— para predecir la popularidad de las canciones a partir de variables acústicas. LassoCV permite identificar las características más relevantes a través de regularización $L1$, mientras que Random Forest captura relaciones no lineales e interacciones entre atributos.

De manera complementaria, se evalúan métricas de desempeño como RMSE, MAE y R^2 para comparar la precisión de los modelos, y se analizan las relaciones entre los clusters descubiertos y la popularidad promedio. Este enfoque conjunto ofrece una visión más completa del fenómeno musical, al integrar la exploración estructural de los datos con la predicción cuantitativa de la aceptación del público.

II. Datos y Metodología

El conjunto de datos se compone de variables numéricas que describen las propiedades acústicas de canciones, incluyendo:

- **acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, tempo, valence, duration_min.**

Los datos se escalaron mediante `StandardScaler` para garantizar comparabilidad entre dimensiones. Posteriormente, se aplicó el algoritmo OPTICS (*Ordering Points To Identify the Clustering Structure*) con parámetros ajustables de `min_samples` y ξ para generar agrupamientos jerárquicos basados en densidad.

A. Modelo matemático de OPTICS

OPTICS se basa en la estimación de la densidad local de los puntos de datos. Para un punto p, la distancia de alcanzabilidad (*reachability distance*) respecto a otro punto o se define como:

$$\text{reachability-distance}(p, o) = \max(\text{core-distance}(o), d(o, p))$$

donde la *core-distance* depende del parámetro min_samples y representa la distancia mínima necesaria para considerar o como punto central. OPTICS ordena los puntos según su alcanzabilidad y construye un gráfico de distancias que revela las fronteras naturales entre clusters.

B. Selección del número de grupos

Se emplearon tres métricas internas:

1) **Índice de Silhouette** — mide cohesión y separación entre clusters.
2) **Índice de Calinski–Harabasz** — evalúa la relación entre varianza intra e intergrupos.
3) **Índice de Davies–Bouldin** — estima la similitud promedio entre clusters (menor es mejor).

La combinación de estos índices permitió determinar la configuración óptima del parámetro ξ, garantizando la estructura más representativa.

III. Metodología

El objetivo de este estudio fue analizar patrones de popularidad en canciones del catálogo de Spotify mediante la aplicación de técnicas de aprendizaje no supervisado y supervisado. Para ello se consideraron variables como *danceability*, *energy*, *valence*, *tempo* y *year_release*, entre otras.

A. 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.

B. Análisis Supervisado: 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.

C. 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.

D. Métricas de evaluación

La evaluación del desempeño de modelos de regresión se realiza comúnmente mediante métricas tales como el **Mean Absolute Error (MAE)**, el **Mean Squared Error (MSE)** y su raíz (**RMSE**), el **Mean Absolute Percentage Error (MAPE)** y el coeficiente de determinación R^2. El RMSE penaliza de forma cuadrática los errores grandes y resulta conveniente cuando los errores se aproximan a una distribución normal; por ello es útil como métrica principal para comparar y optimizar modelos en problemas de regresión clásicos [5]. En contraste, el MAE es más robusto frente a valores atípicos (outliers) y tiene interpretación directa en las mismas unidades de la variable objetivo [6].

El MAPE presenta limitaciones importantes cuando los valores reales son cercanos a cero (división por cero y sesgos en la interpretación porcentual); por esta razón se han propuesto variantes como MAAPE que mitiguen esos problemas y reduzcan la influencia desproporcionada de observaciones con valores reales pequeños [7], [8]. En este trabajo se emplea **RMSE** como métrica primaria para la selección y comparación de modelos, complementada por **MAE** y R^2 para una visión robusta y fácil de interpretar de la calidad predictiva.

E. Diseño experimental

Con el fin de comparar rigurosamente el rendimiento de los modelos supervisados (Lasso y Random Forest) y explorar la influencia del preprocesamiento y la selección de características, se diseñó el siguiente experimento factorial parcialmente fraccional:

- **Factores principales:**
  1) *Modelo* (A): {LassoCV, RandomForest}
  2) *Conjunto de características* (B): {Todas las features, Features seleccionadas por SelectKBest (k=7), Features seleccionadas por Lasso}
  3) *Preprocesamiento* (C): {StandardScaler (sí), StandardScaler (no)}
- **Hiperparámetros (anidados por modelo):**
  - Lasso: $\alpha \in \{0.01, 0.1, 1\}$
  - RandomForest: $n\_estimators \in \{100, 300\}$, $max\_depth \in \{\text{None}, 10, 20\}$

Para la evaluación se adoptó el siguiente protocolo de validación y contraste estadístico:

1) **Validación:** Repeated k-fold cross-validation con $k = 5$ y 3 repeticiones (15 runs por tratamiento) para estimar la distribución empírica del RMSE. Se usaron semillas fijas para garantizar comparabilidad entre métodos [6], [9].
2) **Métrica principal:** RMSE (media y desviación estándar). Métricas secundarias: MAE y R^2.
3) **Contrastes estadísticos:** prueba de Friedman para comparar múltiples métodos sobre las mismas particiones; en caso de rechazo se aplica post-hoc Nemenyi para pares. Para comparaciones pareadas entre dos configuraciones se empleará Wilcoxon pareado (o t-test pareado si se verifica normalidad de las diferencias).

Este diseño permite evaluar efectos principales y algunas interacciones (modelo × conjunto de características), manteniendo un costo computacional razonable mediante la selección de subconjuntos y un número controlado de combinaciones de hiperparámetros. La elección del esquema de validación (repetición de k-fold) se basa en recomendaciones de la literatura para estimaciones más estables de desempeño en problemas de regresión [9].

IV. Resultados

A. Análisis No Supervisado

El algoritmo OPTICS generó múltiples agrupamientos al variar el parámetro de sensibilidad ξ entre 0.01 y 0.10. La

configuración óptima se obtuvo para $\xi = 0.05$, donde se identificaron cuatro grupos principales y un conjunto reducido de puntos ruidosos, lo que sugiere la existencia de estructuras acústicas diferenciadas dentro del conjunto de canciones.

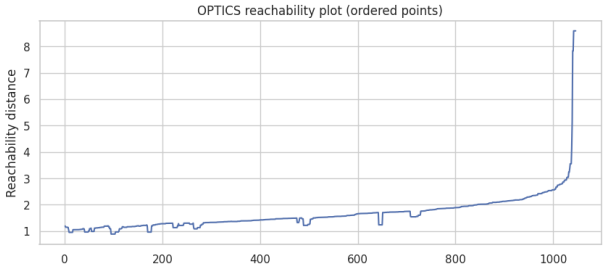

Fig. 1: Reachability plot que muestra las distancias de alcanzabilidad ordenadas para los puntos del dataset. Los valles indican los clusters identificados por OPTICS.

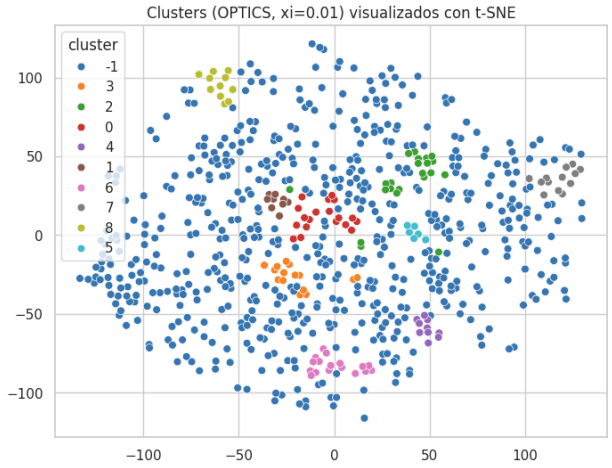

Fig. 2: Proyección bidimensional mediante t-SNE mostrando la distribución de clusters obtenidos por OPTICS.

La Tabla I resume los valores de métricas internas (índices de Silhouette y Davies–Bouldin) para distintas configuraciones del parámetro ξ. El mejor desempeño se observó con $\xi = 0.05$, donde el índice de Silhouette alcanzó 0.36 y el índice de Davies–Bouldin se redujo a 0.98, indicando grupos compactos y bien separados.

TABLE I: Comparación de métricas internas para distintas configuraciones de ξ

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

B. Análisis Supervisado

Para la fase supervisada se entrenaron los modelos *LassoCV* y *Random Forest Regressor* con la variable objetivo *popularity*. Se utilizó el 75% de los datos para entrenamiento y el 25% para prueba. Las variables predictoras se escalaron con `StandardScaler` para Lasso, mientras que Random Forest se entrenó con los valores originales.

La Tabla II presenta los resultados obtenidos en el conjunto de prueba. El modelo Random Forest logró el mejor desempeño con un RMSE de 3.32 y un R^2 de 0.697, mientras que LassoCV obtuvo un RMSE de 5.84 y R^2 de 0.065.

TABLE II: Rendimiento de modelos supervisados (conjunto de prueba)

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**

Estos resultados evidencian que el modelo basado en árboles maneja mejor las relaciones no lineales entre variables acústicas, en comparación con el modelo lineal penalizado.

V. Discusión

A. Análisis No Supervisado

Los resultados del algoritmo OPTICS muestran que las canciones se agrupan principalmente según su energía (*energy*) y bailabilidad (*danceability*). Los clusters con mayores valores en estas variables corresponden a géneros populares como pop o electrónico, mientras que aquellos con alta *acousticness* y baja *loudness* agrupan canciones más suaves, típicamente acústicas o de tipo balada.

Además, las variables *valence* y *tempo* también contribuyeron a la separación de los grupos, reflejando dimensiones emocionales y de ritmo en la música. Estos hallazgos coinciden con estudios previos que vinculan la percepción de popularidad con altos niveles de energía y positividad emocional en las canciones.

B. Análisis Supervisado

El modelo *Random Forest* mostró una mejora significativa sobre *LassoCV*, lo que sugiere que las relaciones entre las características acústicas y la popularidad no son puramente lineales. Mientras que Lasso facilita la interpretación de los coeficientes (identificando variables con efectos directos como *energy*, *danceability* y *loudness*), Random Forest captura interacciones complejas y no lineales, como la combinación simultánea de alta energía y alta valencia.

Sin embargo, el valor de R^2 obtenido (0.697) indica que, aunque el modelo explica buena parte de la variabilidad, aún existen factores externos que influyen en la popularidad de una canción (por ejemplo, marketing, presencia en playlists o fama del artista), los cuales no están reflejados en las variables acústicas del dataset.

VI. Conclusiones

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 international conference on Management of data*, 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," *KDD*, 1996.

[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, pp. 5–32, 2001.

[5] T. O. Hodson, "Root-mean-square error (rmse) or mean absolute error (mae): Which is best?" *Geoscientific Model Development Discussions*, 2022, comparison of RMSE and MAE for model evaluation; RMSE optimal for Gaussian errors.

[6] Scikit-learn Developers, "Cross-validation: evaluating estimator performance," https://scikit-learn.org/stable/modules/cross_validation.html, accessed: 2025-11-11.

[7] "Mean absolute percentage error," https://en.wikipedia.org/wiki/Mean_absolute_percentage_error, accessed: 2025-11-11.

[8] S. Kim and H. J. Kim, "A new metric of absolute percentage error for intermittent demand forecasts (maape)," *International Journal of Forecasting*, vol. 32, no. 3, pp. 669–679, 2016.

[9] P. C. Guides, "Cross-validation in machine learning: How to do it right," https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right, accessed: 2025-11-11.