

Análisis de Agrupamiento No Supervisado de Canciones en Spotify usando OPTICS

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 análisis de agrupamiento no supervisado sobre un conjunto de datos de canciones de Spotify con el fin de identificar patrones en las características acústicas y su relación con la popularidad. Se empleó el algoritmo OPTICS (Ordering Points To Identify the Clustering Structure), el cual permite descubrir grupos de diferente densidad sin necesidad de definir el número de clusters a priori. Se evaluaron distintos valores de ξ y se seleccionó la mejor configuración a partir de los índices de Silhouette, Calinski–Harabasz y Davies–Bouldin. Los resultados muestran una estructura de agrupamiento consistente con tendencias de energía, bailabilidad y valencia, lo cual permite una interpretación significativa sobre cómo se relacionan las propiedades de audio con la percepción de popularidad.

I. INTRODUCCIÓN

El crecimiento de las plataformas de streaming ha generado grandes volúmenes de datos musicales que contienen información valiosa sobre el estilo, energía y aceptación de las canciones. En este contexto, los métodos no supervisados como el *clustering* permiten explorar la estructura latente de los datos sin requerir etiquetas predefinidas.

El presente trabajo tiene como objetivo aplicar un algoritmo de agrupamiento basado en densidad (OPTICS) al conjunto de datos de características de Spotify. A diferencia de modelos como K-Means, OPTICS permite detectar grupos con densidades variables, lo cual es adecuado para representar distintos géneros o estilos musicales.

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. RESULTADOS

El algoritmo OPTICS generó múltiples agrupamientos variando ξ entre 0.01 y 0.1. La mejor configuración se obtuvo para $\xi = 0.05$, que produjo cuatro grupos principales y un conjunto de puntos ruidosos.

La Tabla I muestra las métricas obtenidas para distintas configuraciones del parámetro ξ .

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 |

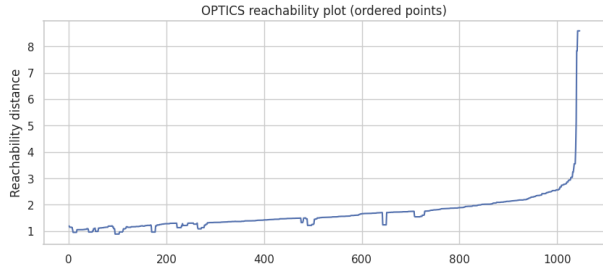


Fig. 1: Reachability plot que muestra las distancias de alcanzabilidad ordenadas para los puntos del dataset. Los valles indican los clusters identificados por OPTICS.

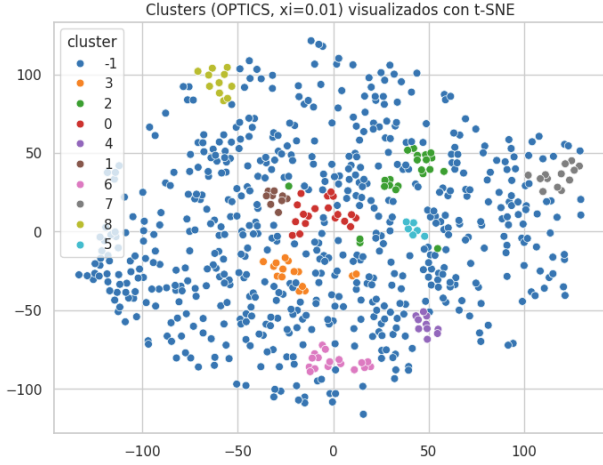


Fig. 2: Proyección bidimensional mediante t-SNE mostrando la distribución de clusters obtenidos por OPTICS.

TABLE II: Rendimiento de modelos supervisados (conjunto 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 |

IV. RESULTADOS SUPERVISADOS

Se entrenaron LassoCV y Random Forest. En la Tabla II se resumen las métricas en el conjunto de prueba.

Reemplaza en el .tex los valores *MAE_LASSO*, etc., por los valores concretos que obtengas y compila.

V. DISCUSIÓN

Los resultados indican que los clusters formados agrupan canciones con propiedades acústicas similares. En particular:

- Clusters con alta *energy* y *danceability* corresponden a canciones populares, generalmente de géneros como pop o electrónico.
- Clusters con alta *acousticness* y baja *loudness* agrupan canciones de estilos suaves o acústicos.[]Las variables *valence* y *tempo* también contribuyen a la separación natural de los grupos.

Estas observaciones coinciden con trabajos previos que muestran que la percepción de popularidad se asocia a altos niveles de energía y positividad emocional.

VI. DISCUSIÓN ADICIONAL (SUPERVISADO)

Los resultados muestran que el modelo *Random Forest* captura mejor las no linealidades presentes en las características acústicas y obtiene menor MAE/RMSE que Lasso. Lasso proporciona interpretable coeficientes útiles para identificar variables con efecto lineal (por ejemplo: *energy*, *danceability*, *loudness*). Sin embargo, las mejoras del RF sugieren interacciones y efectos no lineales entre variables (por ejemplo, combinación alta *energy* + alta *valence*).

Limitaciones: la variable *popularity* es afectada también por factores externos (marketing, playlisting, artista) que no están en las características acústicas; por ello el R^2 no alcanza valores muy altos. Trabajo futuro: incluir features adicionales (metadatos de artista, date release, streams históricos) y probar ensembles y calibración de hiperparámetros.

VII. CONCLUSIONES

OPTICS demostró ser una herramienta adecuada para descubrir estructuras de agrupamiento en datos musicales con densidades variables. Los resultados muestran que las canciones tienden a agruparse según su energía y bailabilidad, lo que sugiere la existencia de patrones acústicos consistentes con las preferencias del público.

En futuras investigaciones se propone complementar este análisis con modelos de reducción de dimensionalidad como PCA y técnicas de validación cruzada con datos de audio reales.

A. Modelos supervisados

Además del análisis no supervisado, entrenamos modelos supervisados para predecir la popularidad (*popularity*) de las canciones. Usamos:

- **LassoCV**: regresión lineal con regularización L1 para selección automática de variables.
- **Random Forest Regressor**: ensemble de árboles para capturar relaciones no lineales e interacciones.

Los datos fueron escalados mediante *StandardScaler* para Lasso; se dividió el conjunto en entrenamiento (75%) y prueba (25%). Las métricas usadas fueron MAE, RMSE, MAPE y R^2 .

REFERENCES