Aprendizaje No Supervisado - Trabajo final Maestría en Ciencia de Datos

Alumnos: Dominutti Nicolás, Suárez Gurruchaga Carlos Roque, y Telechea Hernán
Profesor: Fernández Piana Lucas

I. Introducción y objetivo

En el siguiente informe, buscamos encontrar estructuras o características latentes que logren discriminar en diferentes clusters a las canciones que componen la playlist. A partir de distintos métodos de clustering y utilizando técnicas de validación interna y externa, esperamos obtener grupos coherentes. Aplicamos 2 enfoques de trabajo. En el primero, aplicamos técnicas de clusterización ("k-means", DBSCAN y espectrales) y abordamos el problema a partir de UMAP y un método jerárquico divisivo. En el segundo enfoque, aplicamos PCA con kernel sigmoide para reducir la dimensionalidad e intentar observar una definición más clara de los clusters. En ambos casos, trabajamos con datos escalados. Estudiamos una base de datos que cuenta con 19 dimensiones, 5 de ellas categóricas y las restantes numéricas. Además, realizamos llamadas a la API de Spotify para conseguir otros datos, como el género de los artistas presentes en la base original. De la base resultante, escalamos las variables numéricas y decidimos eliminar la variable "key" porque no aportaba a nuestro análisis. No eliminamos observaciones al no encontrar pistas repetidas, aunque sí mantuvimos versiones diferentes de una misma canción (de estudio, en vivo o remasterizada), ya que difieren valores en algunas dimensiones.

II. Experimentación 1

Nos guiamos por 2 métricas de validación interna para determinar el número de clusters y comparar los resultados entre modelos: silhouette score y Calinski-Harabasz ("C-H"). También utilizamos UMAP para reducir la dimensionalidad y poder realizar una mejor visualización del problema.

Método	Silhouette	C-H
K-means	0.177248	54.313681
DBSCAN	0.343571	6.958293
Espectral	0.157718	49.989273
Divisivo	0.430192	304.527779

Figura 1. Tabla de métricas de los modelos corridos en la experimentación 1.

K-means: tanto la métrica de Silhouette como de Calinski-Harabasz alcanzan su máximo valor con 2 clusters. Una tercera métrica, la inercia, nos sugiere 3 clusters. Como indica la Tabla 1, el silhouette score obtenido es 0.18, lo que nos sugiere que, según las fuentes de validación interna, no son muy buenos los clusters formados, dado que un valor cercano a 0 indica solapamiento entre los grupos.
Espectrales: esté método obtuvo la peor métrica de Silhouette (0.16) para el mismo número de clusters (2). Nuevamente obtenemos clusters solapados.
DBSCAN: las métricas mencionadas nos sugieren dos valores de epsilon: 4.4 y 3.2. Observamos una mejor métrica de Silhouette, el doble que con "k-means", aunque sigue siendo pequeño. Esto refuerza la idea de que los clusters están solapados, ya que no es claro el criterio de división entre las canciones.
Método jerárquico divisivo: Se intentó encarar el problema desde otro punto de vista, aplicando UMAP para reducir la dimensionalidad y poder visualizar los datos. Si bien UMAP, busca preservar la estructura local de los datos y no da tanta importancia a la global, parecen observarse 2 clusters, separados por una zona menos densa de puntos. Luego se propuso aplicar un método jerárquico divisivo y evaluar las métricas mencionadas previamente. Este método obtuvo el mejor número de Silhouette, 0.43. Si bien sigue sin ser clara la distinción de clusters, parece que logramos encontrar nuestro mejor modelo (ver Gráfico 1).
Al observar los valores medios y de varianza de las features en cada cluster, concluimos lo siguiente:
- **Cluster 0:** este grupo de canciones parecen tener los niveles más altos de "energy", que se caracterizan por ser canciones que se perciben rápidas y ruidosas (por ejemplo, un rock o un cuarteto). También encontramos que los niveles de "acousticness" son de los más bajos, esto es, canciones que utilizan en mayor medida modificaciones del ritmo de forma electrónica (por ejemplo, "Palomitas de maíz" de Los Pekenikes y "Fuegos de Octubre" de Patricio Rey y sus Redonditos de Ricota). Un último patrón encontrado es el bajo nivel de "instrumentalness", que indica que las canciones tienen un gran contenido hablado.
- **Cluster 1:** en este grupo, encontramos canciones que en su mayoría no fueron grabadas en vivo. También observamos que su duración en promedio es menor al primer cluster.

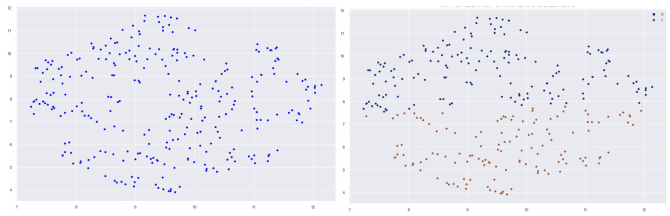

Figura 2. representación UMAP antes (izquierda) y después (derecha) de aplicar el método jerárquico divisivo.

III. Experimentación 2

Utilizamos técnicas de validación interna y externa. Obtuvimos los siguientes resultados de la interna:

método	Silhouette	C-H
K-means	0.220308	69.179589
GM	0.195276	61.471908
DBSCAN	0.454506	2.732768
Espectral	0.333903	5.816809

Figura 3. Tabla de métricas de los modelos corridos en la experimentación 2.

Como se observa en la Tabla 2, DBSCAN y espectrales presentan los mejores resultados de Silhouette. Sin embargo, estos modelos no logran identificar clusters claros cuando aplicamos procesos de validación externa (visualización del modelo ploteado con el kernel PCA, comparación con la distribución de las variables y el solapamiento entre los artistas y su género). Solo "k-means" logra identificar clusters interesantes, que además tiene el mayor valor en el índice C-H. Sumado al análisis de la inercia (ver Anexo: Imagen 1), seleccionamos el corte en 3 clusters.

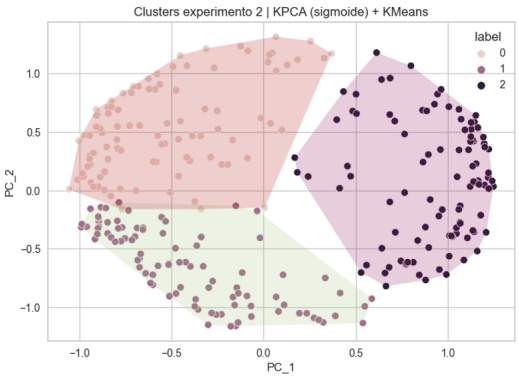

Figura 4. representación de los datos y clusters en los 1ros 2 componentes principales según el Kernel PCA.

Del proceso de validación externa, surgió que cada uno de los 3 clusters tiene sus características distintivas (surgidas del análisis de las variables y géneros, ver Anexo: Imágenes 2, 3 y 4):

Cluster 1

- menos acústico
- poco instrumental
- mayor componente de "liveness"
- mayor aceleración en "tempo"
- mayor popularidad de canción y artista
- mayor antigüedad desde el lanzamiento
- gran presencia de rock (local e internacional)

Cluster 2

- alta "danceability", energía, valencia y sonido
- mayor "speechiness"
- menor antigüedad desde el lanzamiento
- gran presencia de cumbia y artistas sin género (como "La bomba tucumana", Ricky Maravilla, La Banda Dominguera, El Jarra)

Cluster 3

- baja "danceability", energía, valencia y sonido
- más acústico e instrumental
- canciones de estudio más que en vivo
- menor aceleración en "tempo"
- menor popularidad de canción y artista
- gran presencia de rock y sin género (como Angel Mahler, Will Ferrell, Elena Roger)

Figura 5. descripción de los clusters obtenidos con la experimentación 2.

En cuanto al overlap de artistas en cada grupo (ver Anexo: Imagen 5), no surgieron valores que nos llamaran la atención, dado que la coincidencia máxima fue de solo 12

IV. Conclusiones

A partir de la idea que no existen clusters "reales", sino que su validez depende directamente del contexto y del objetivo perseguido (Henning, June 12, 2018), es que en el segundo enfoque, a pesar de que DBSCAN nos dio un silhoutte más elevado que "k-means", optamos por darle mayor importancia al criterio externo que a las métricas de validación interna. Bajo la misma idea de que no existe un cluster óptimo, podemos decir que ambos enfoques de trabajo no son excluyentes sino complementarios. Cada uno nos da información sobre lo que encontró, aunque no podemos concluir fuertemente sobre las variables.

V. Anexo:

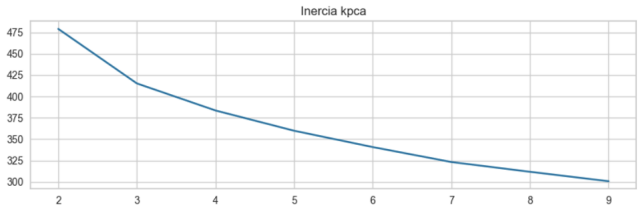

Figura 6. valor de la inercia según el n° de clusters utilizados en K-means.

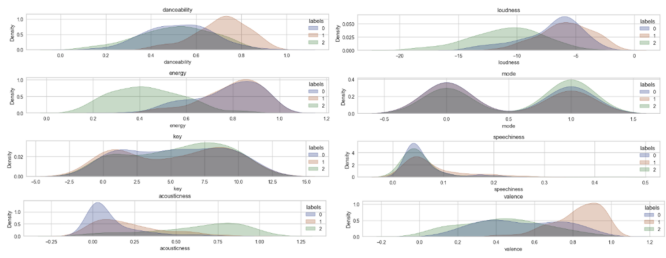

Figura 7. distribuciones de las variables del dataset según el cluster (parte 1).

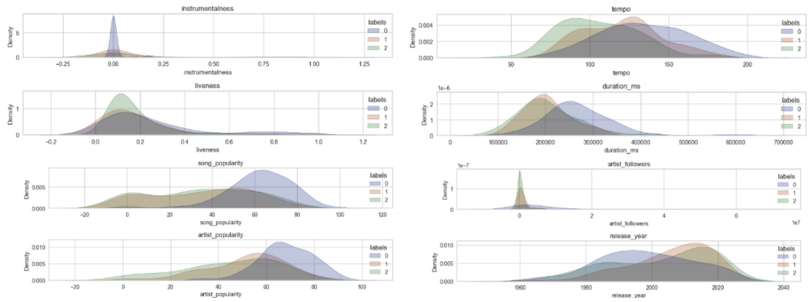

Figura 8. distribuciones de las variables del dataset según el cluster (parte 2).

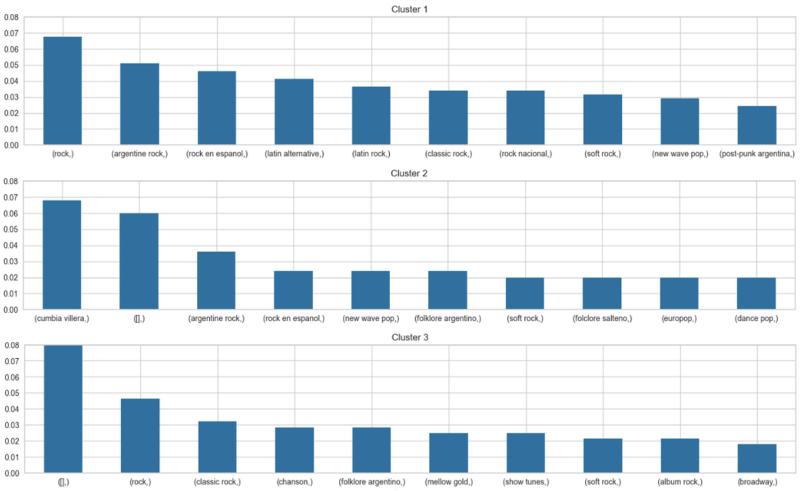

Figura 9. Top 10 géneros por cluster.

```
Overlap de artistas entre cluster 1 y 2
Artistas en el overlap: ['Rod Stewart', 'Damas Gratis', 'The Weather Girls', 'Onda Sabanera']
['Patricio Rey y sus Redonditos de Ricota', 'Madonna', 'Bizarrap', 'Billy Joel', 'Tina Turner']
% de overlap sobre cluster 1: 0.13
% de overlap sobre cluster 2: 0.12
--------------------------------------------------------------------------------
Overlap de artistas entre cluster 1 y 3
Artistas en el overlap: {'Billy Joel', 'Survivor', 'Scorpions'}
% de overlap sobre cluster 1: 0.04
% de overlap sobre cluster 3: 0.03
--------------------------------------------------------------------------------
Overlap de artistas entre cluster 2 y 3
Artistas en el overlap: ["Juan D'Arienzo", 'Victor Heredia', 'The Blues Brothers', 'Billy Joel']
['Dolly Parton', 'Modern Talking']
% de overlap sobre cluster 2: 0.08
% de overlap sobre cluster 3: 0.07
```

Figura 10. Overlap de artistas según cluster.