**Introducción**

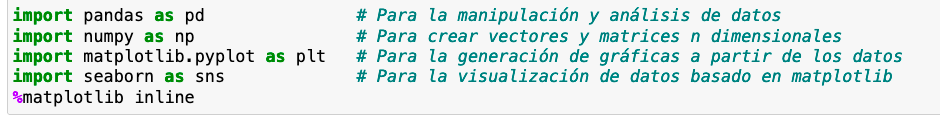
Para esta práctica se generarán dos modelos de clustering, uno particional y uno jerárquico, a partir de los datos que se propongan. A partir de cada uno de los modelos se generarán interpretaciones de las agrupaciones generadas y se analizarán los resultados que se generaron. Con esta práctica se reforzará el conocimiento del clustering y se podrá hacer uso de datos que proponga el alumno.

**Objetivos**

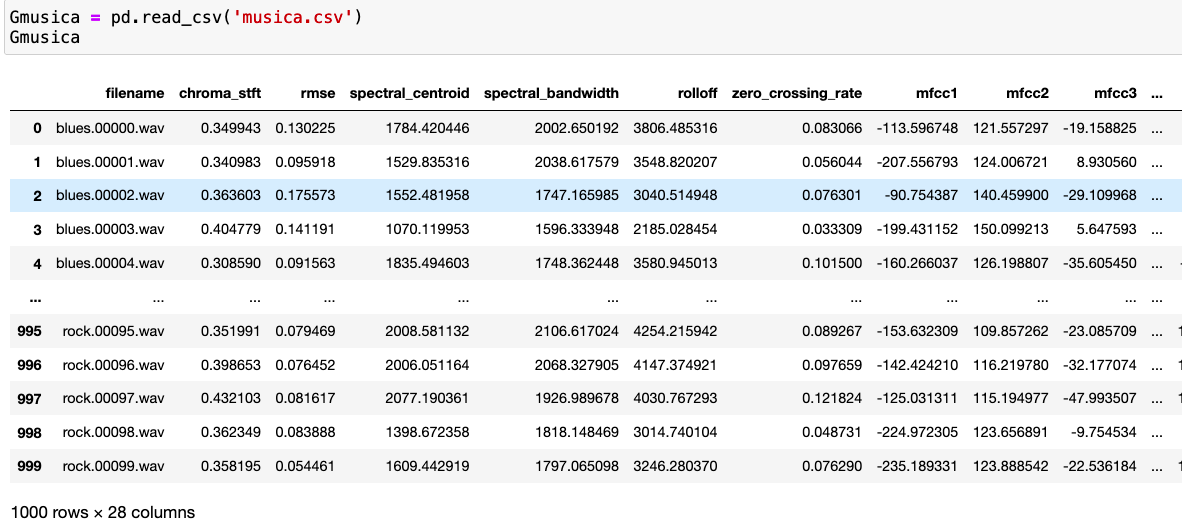
A partir de un conjunto de datos de características físicas de las ondas que tienen diferentes canciones de diferentes géneros, obtener grupos de canciones con características similares que determinen el género, a través de clustering jerárquico y particional.

**Desarrollo**

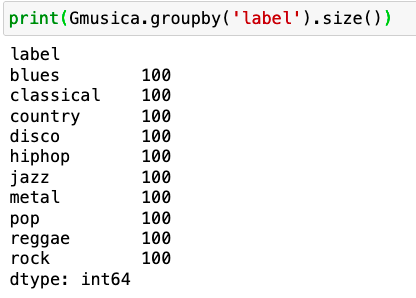
Se importaron las bibliotecas necesarias para el desarrollo de la práctica.



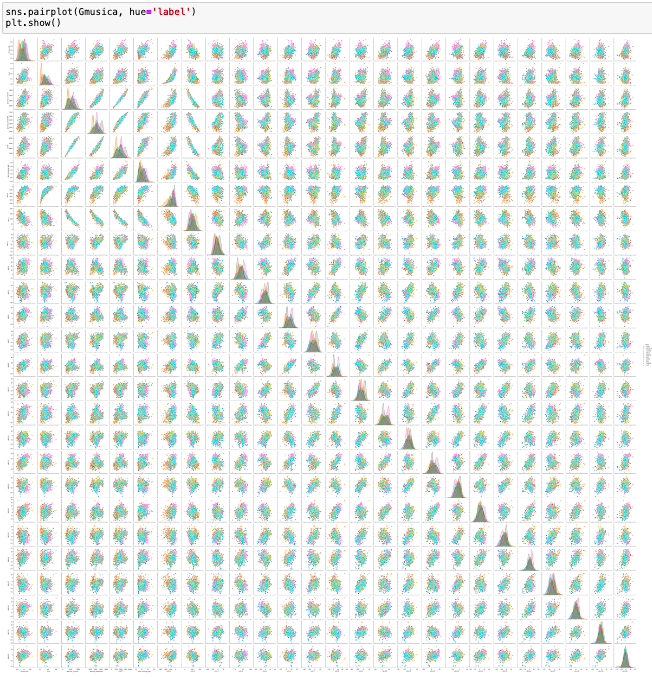
Se leen los datos a utilizar. Se tienen mil datos y 28 columnas de variables.



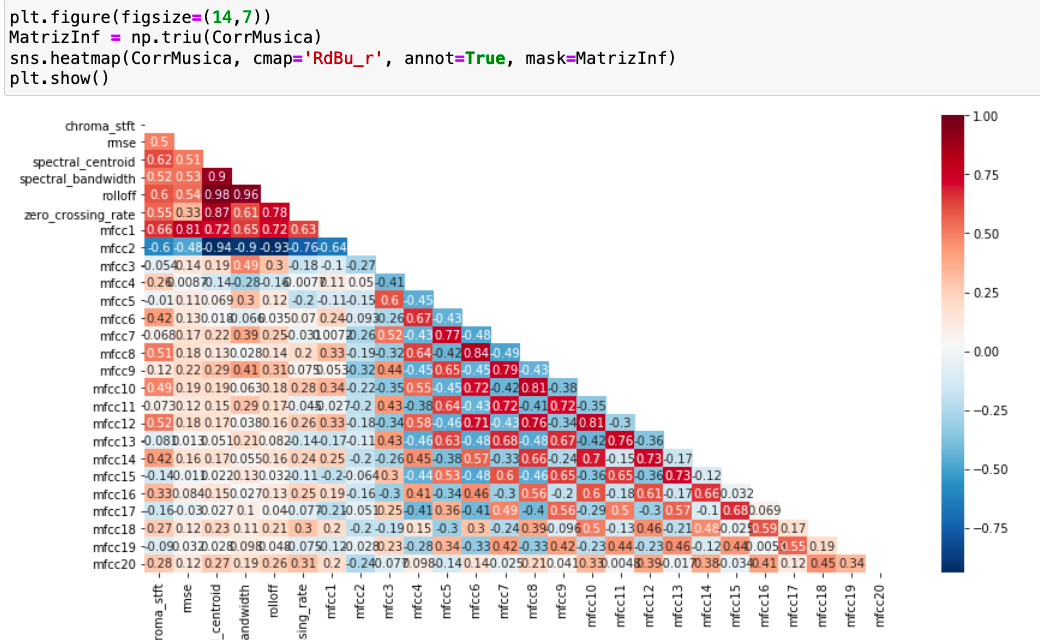
Se agrupan a los datos por género para obtener cuántas canciones de cada género se tienen en el conjunto de datos. Se tienen 100 canciones de 10 géneros distintos. Además, se verificó si existían datos nulos para poder eliminar dichas filas, pero no existe ningún dato nulo.



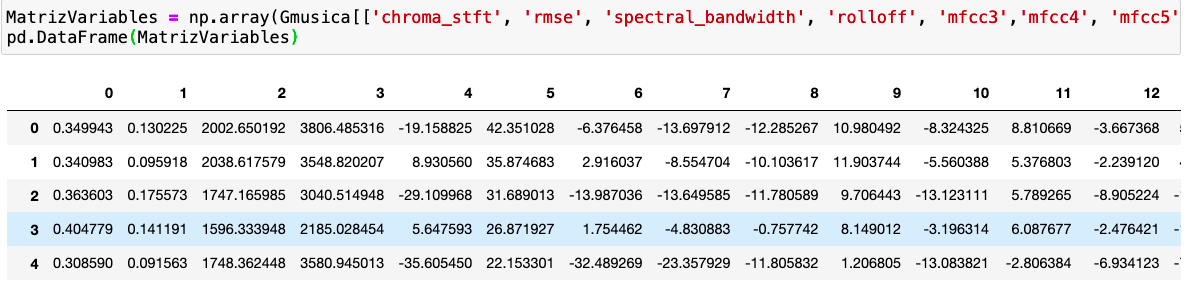
Se generó una evaluación visual de la correlación entre las diferentes variables de los datos pero al ser muchas variables esta evaluación visual no fue tan útil. Se utilizará el valor de las correlaciones para generar una reducción de dimensionalidad.



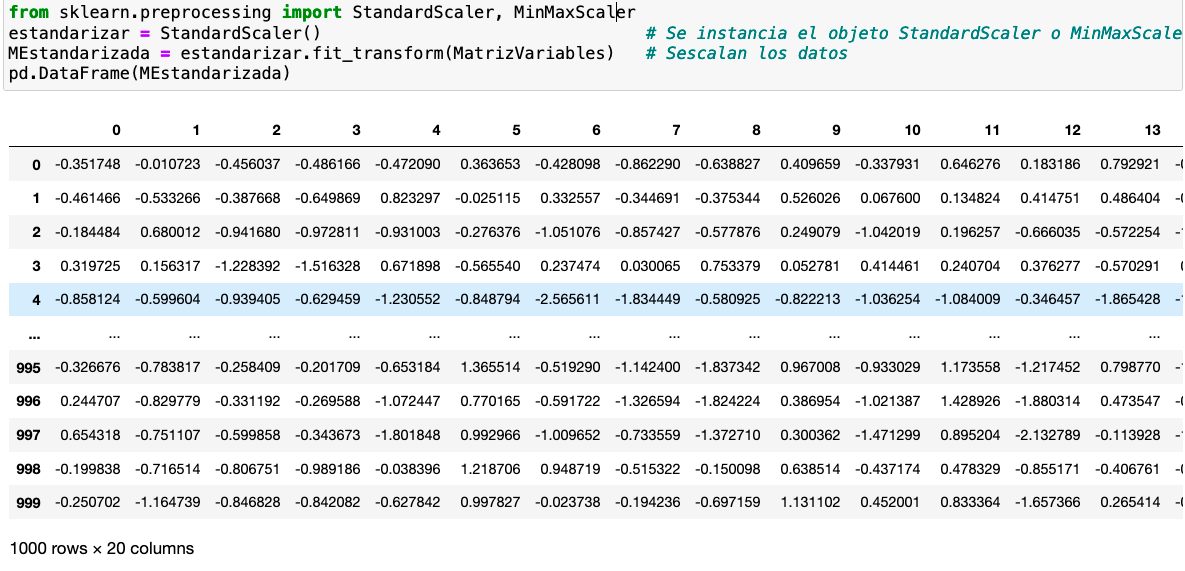
Se generó la matriz de correlaciones de los datos utilizando el método de Pearson y a partir de ella se generó un mapa de calor donde se puede observar qué relaciones tienen alta correlación entre ellas. Con este pasó se generó la selección de variables eliminando aquellas que tienen alta correlación entre ellas.



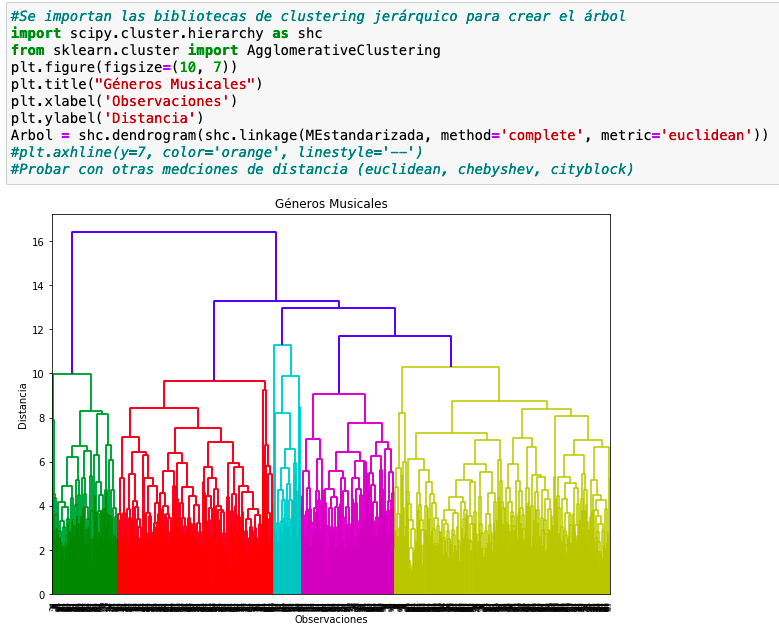
Las variables seleccionadas fueron: chroma\_stft, rmse, spectral\_bandwidth, rolloff, mfcc3 al mfcc5, mfcc7 y mfcc9 al mfcc20. El DataFrame resultante es el siguiente:



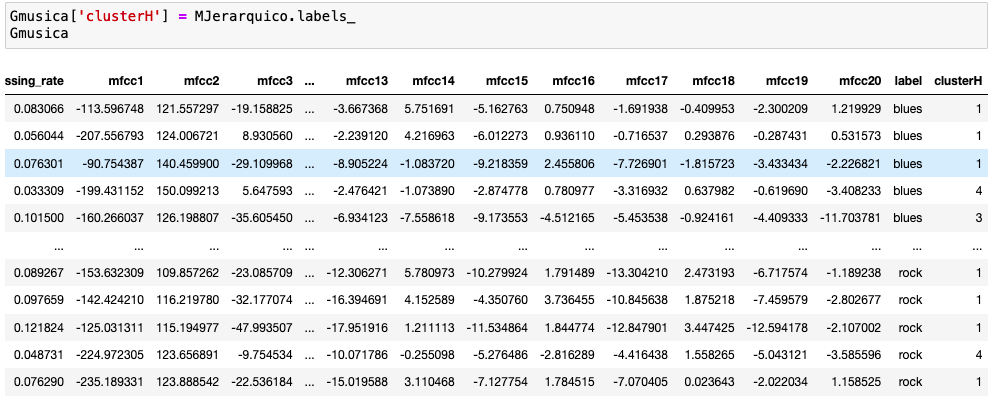
Se estandarizaron los datos para que ninguna tuviera mayor valor que otros. Esta matriz estandarizada será utilizada para los dos métodos de clustering. La matriz de variables estandarizadas es la siguiente.



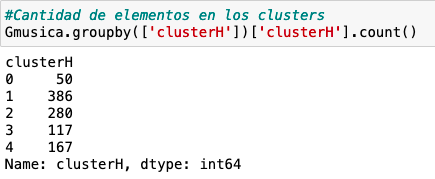
Se creó el clustering jerárquico para saber el número de agrupaciones que utilizará este algoritmo.



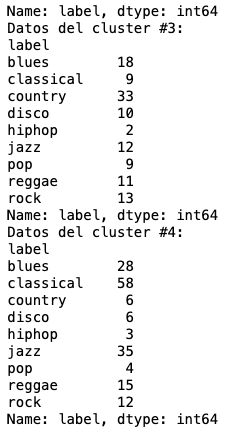
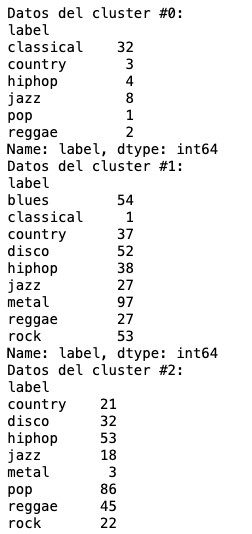
A partir de este modelo se obtuvo que se tenían 5 grupos y se crearon las etiquetas para estos cinco grupos. Dichas etiquetas se añadieron al conjunto de datos originales con el nombre de la columna de clusterH. El conjunto de datos resultante es el siguiente.



Se contó la cantidad de elementos que hay en cada uno de los clústeres para el tamaño de cada clúster.



Para la interpretación de cada uno de los datos se contó de qué genero realmente pertenece para saber qué género sí se está agrupando bien y cuáles no.



Para el clúster 0 se tiene un grupo de canciones donde la gran mayoría son de clásica. Algunas canciones de jazz se metieron a esta agrupación, pero en general es una de las mejores agrupaciones, aunque tiene pocos datos. Debe de ser algún sub-género de clásica que sea muy distinguible.

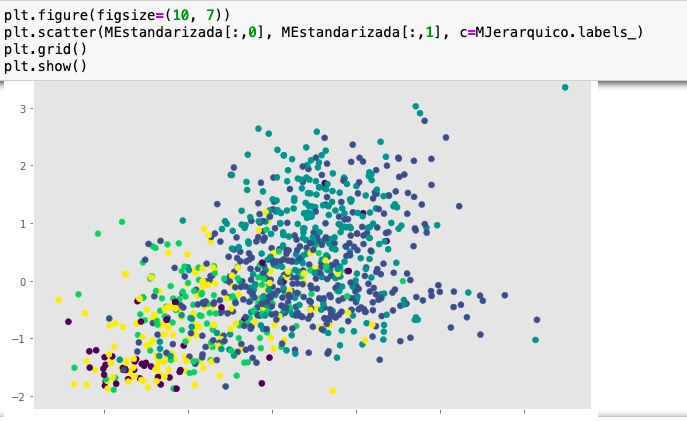
Para el clúster 1 se tiene un grupo de canciones donde se agrupó a la gran mayoría de las canciones. Esta agrupación tiene la peculiaridad de que contiene casi todas las canciones de metal. Esto puede ser un indicador de que esta agrupación capta canciones con mucho ruido y mucho movimiento. Prueba de esto es que también incluye la mitad del disco, la mitad del rock y la mitad del blues.

Para el clúster 2 se tiene un grupo de canciones donde la mayoría son de pop. Este grupo también cuenta con muchas canciones, pero es de los que mejores clasificó los géneros. Cuenta con casi todas las canciones de pop, la mitad de hip-hop y la mitad de reggae.

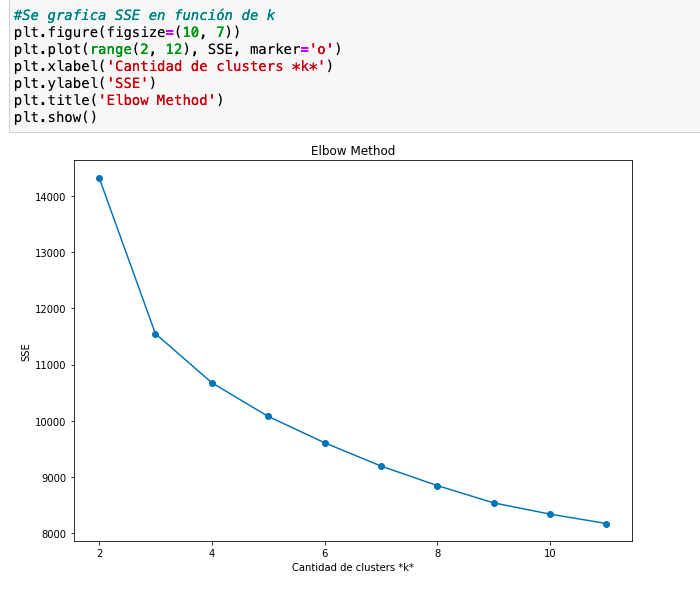
Para el clúster 3 se tiene un grupo de canciones donde la música country. En general, el country fue un género que se encontró en varias agrupaciones y esto se puede deber a que es muy parecido a otros géneros y es muy variado en su sonido.

Para el clúster 4 se tiene un grupo de canciones donde la mayoría son de clásica y de jazz. Este grupo es muy similar al primero y esto podría ser porque representa otro subgénero de la música clásica. El jazz es un género que va ligado con la música clásica por lo que tiene sentido que se encuentre en dos agrupaciones junto a ella.

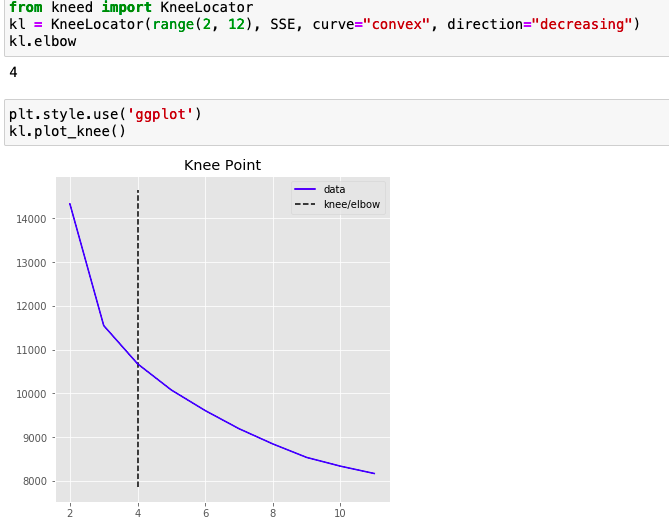
Se generó una gráfica donde se etiquetan los datos con colores. De esta manera no se ve ninguna relación entre las agrupaciones porque el modelo toma en cuenta más dimensiones.



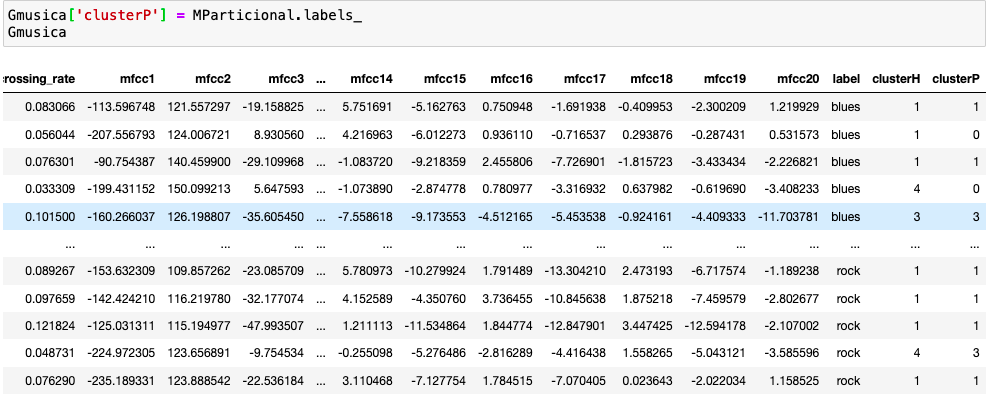
Para el clustering particional se importan las bibliotecas necesarias para generar el modelo utilizando sklearn. Posteriormente, se calcula el SSE para cada uno de los posibles valores de *k* para graficarlo y generar el método del codo. La gráfica generada es la siguiente.



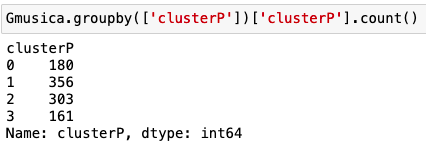
Utilizando la biblioteca como en prácticas anteriores se encontró el punto de inflexión para determinar el número de clústeres. Para esta variación de la agrupación se tienen 4 clústeres.



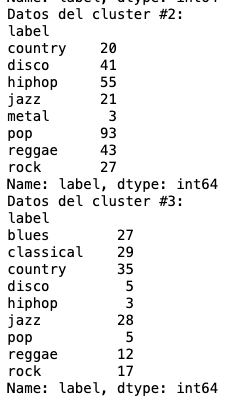
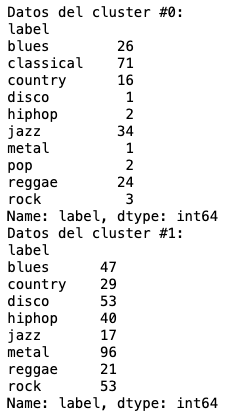
Se crean las etiquetas del modelo y se les adjuntan a los datos originales. La tabla resultante es la siguiente.



Se cuentan las canciones que pertenecen a cada cluster.



Para el clustering particional también se genera una interpretación a partir de el número de canciones que realmente pertenecen a cierto género. Los resultados para cada agrupación son los siguientes.



Para el clúster 0 se tiene un grupo de canciones donde se agrupan a casi todas las canciones de música clásica. Esto es una mejora respecto al modelo anterior donde se tenían dos agrupaciones para la música clásica. En esta agrupación también se encuentra el jazz y otro género que no debería estar relacionado que es el reggae.

Para el clúster 1 se tiene un grupo de canciones que se parece mucho al segundo clúster del modelo anterior. Es la agrupación con más datos y contiene a casi todas las canciones de metal del conjunto de datos. Además, contiene a otras canciones de otros géneros que también llegan a ser ruidosos y enérgicos como el rock, el blues, el hip-hop y el disco.

Para el clúster 2 se tiene un grupo de canciones donde el pop es el género dominante ya que casi todas las canciones del género se encuentran en esta agrupación. Es el segundo clúster con más datos y también contiene canciones de hip-hop, disco y reggae.

Para el clúster 3 se tiene un grupo de canciones donde se tiene la menor cantidad de canciones y no se tiene un género que tengan muchas más canciones que otro. Los géneros que se encuentran en esta agrupación son blues, clásica, country y jazz.

**Conclusiones**

Con estos dos modelos se generaron agrupaciones que permitieron clasificar a diversas canciones por sus características físicas. A partir de dichas clasificaciones se pudo observar cómo existe cierta relación para algunos géneros como el pop, el metal y la música clásica, pero para los demás no existía una agrupación o similitud muy grande. Los resultados fueron buenos y se generaron buenos modelos con mucho sentido. Si se hubieran generado 10 clústeres o muchos más se tendrían modelos que caen en el sobre ajuste, se busca un modelo general que tenga sentido en lo que haga. Aunque ambos modelos tuvieron resultados similares se tuvo uno con menor cantidad de agrupaciones. Gracias a esta práctica se reforzaron ambos algoritmos, aquello que se necesita para llevarlos a cabo y se pudo utilizar un algoritmo de aprendizaje automático para generar resultados a partir de datos propuestos por el alumno.