Instituto Tecnológico y de Estudios Superiores de Monterrey

Departamento de Computación

Herramientas computacionales: el arte de la analítica

Grupo 570

Actividad 4: Patrones con K-means

Joan Daniel Guerrero García A01378052

Fecha de entrega

13 de enero del 2022

Tabla de contenido

1. Datos generales:

El conjunto de datos seleccionado trata sobre una colección de información acerca de canciones de Spotify, cuantificando diversos aspectos de cada una como su ritmo, volumen, energía, etc. De modo que se cree un modelo predictivo a partir de este. Este conjunto se obtuvo de la página Kaggle (Vergnou, 2021), donde se almacenan y publican diversos archivos para análisis de datos de varios temas.

2. Descripción de datos:

Como se puede ver en el archivo *Estadísticas.py*, se registraron 195 registros dentro de 14 variables distintas, cada una teniendo un rango que varía de 0 a 1, las cuales se describen a continuación:

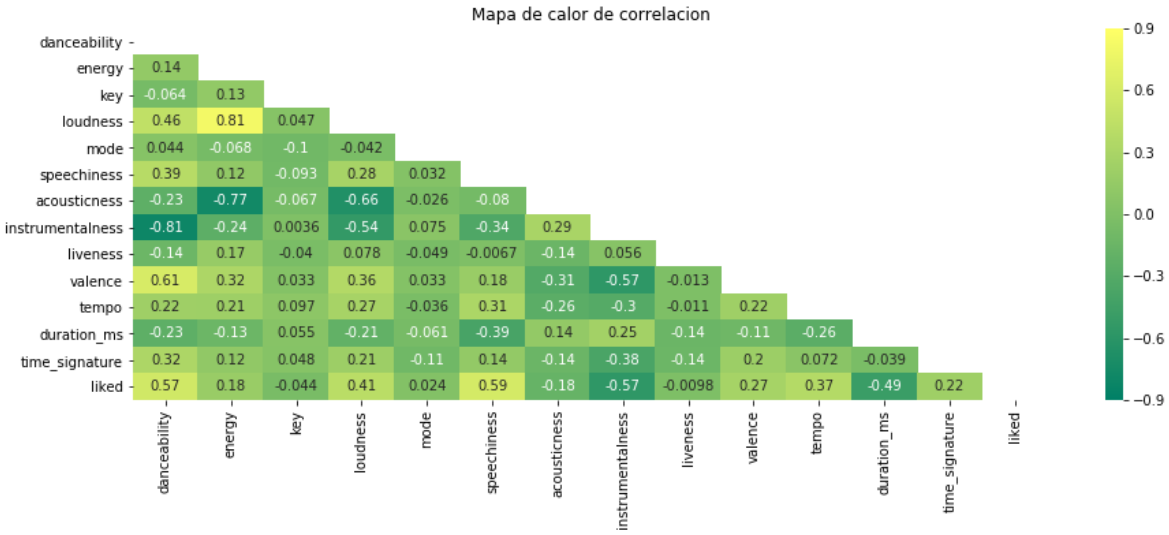
Tabla 1. Variables y sus tipos de datos

Nombre	***Tipo de dato***
Danceability	Float
Energy	Float
Key	Int
Loudness	Float
Mode	Int
Speechiness	Float
Acousticness	Float
Instrumentalness	Float
Liveness	Float
Valence	Float
Tempo	Float
Duration_ms	Int
Time_signature	Int
Liked	Int

3. Análisis comparativo

Dentro de esta actividad se realizó un reporte comparativo entre dos variables de la misma colección de información acerca de canciones de Spotify (Vergnou, 2021). Para realizar las gráficas de mapa de calor, histograma y cajas y bigotes, se seleccionaron durante el análisis un par de variables, distintas en varias veces para obtener un análisis comparativo apropiado. Al final, se seleccionaron las variables *Energy* y *Acousticness*, ya que son de las variables más diferentes entre sí, esto quiere sugerir que, mientras una canción sea más acústica, resulta menos enérgica, pero para ello habría que revisar los resultados obtenidos por las gráficas:

Imagen 1. Gráfica de mapa de calor entre las correlaciones de variables



Observando la gráfica de mapa de calor, los valores más destacables son la diferencia entre Danceability e Instrumentalness, Energy y Acousticness; los cuales son los menores valores. Mientras que valores como Energy y Loudness son los valores más altos. Estas tres comparativas fueron las más apropiadas para este análisis, sin embargo, como el más alto y bajo son diferencias muy altas, es por ello por lo que se escogió analizar Energy y Acousticness. Ya en su reporte comparativo, se puede ver que hay una notoria diferencia entre ellos, incluso desde sus valores de media como se muestra en el diagrama de cajas y bigotes, por lo que la tendencia a separarse una de otra puede llegar a ser evidente. Cabe mencionar, que dado que se ha hecho un preprocesamiento adecuado de los datos, el conjunto de datos originales no parece presentar outliers entre estas dos variables, ya que ambos mantienen un rango igual entre 0 y 1.

Imagen 2. Histograma entre Energy y Acousticness

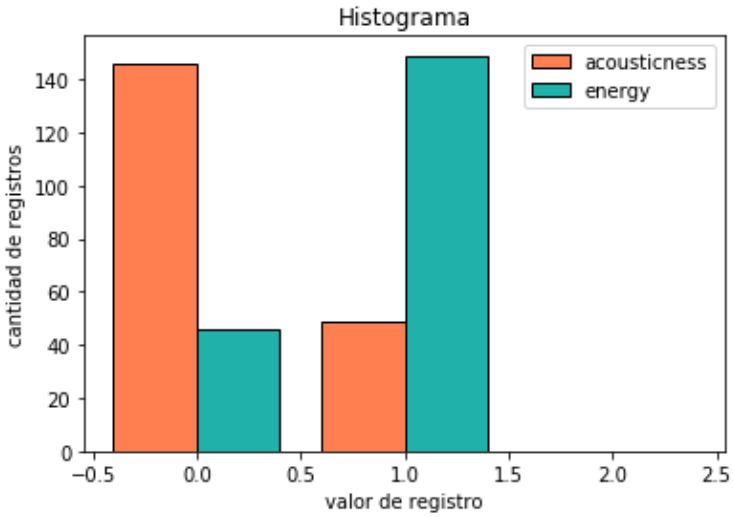
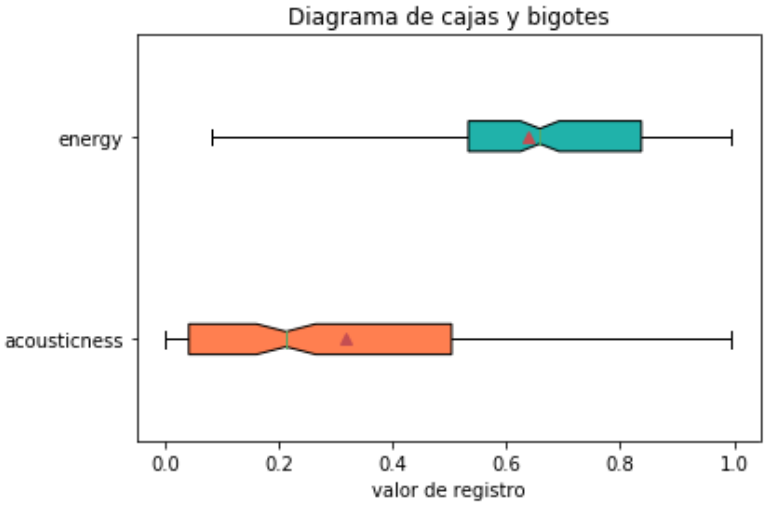


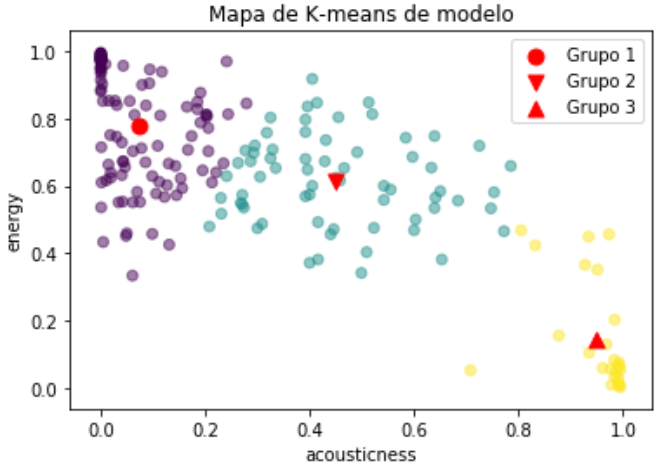
Imagen 3. Diagrama de cajas y bigotes entre Energy y Acousticness



4. Análisis de patrones

Sabiendo que existe una relación entre estas dos variables, se puede realizar un análisis más profundo sobre el comportamiento que tienen entre sí. Para ello, se utilizó en esta sección el algoritmo de K medias, el cual se encarga de agrupar los datos para ambas variables de manera que se pueda particionar en diferentes secciones y encontrar un punto central en cada sección. Dada la imagen 4, se puede ver una correlación inversa, en la que, donde una de las dos variables aumenta, la otra disminuye, y viceversa. Esto se puede ver por el comportamiento entre los tres centroides de los grupos, que conforman una línea con pendiente negativa; esto quiere decir, que las variables de energía y acústica son inversamente proporcionales, en donde, mientras una canción sea más enérgica, es menos probable que se trate de una canción con instrumentos acústicos, y que mientras más enfocada este la canción en ser de tipo acústica, menos energía puede que llegue a tener.

Imagen 4. Mapa de K-means



Tomando en cuenta que los datos son inversamente proporcionales, se realizó el análisis de K medias utilizando un valor de $k = 3$ tomando en cuenta que el significado en la relación entre ambas variables se puede resumir en, aquellas canciones que son muy enérgicas y poco acústicas (siendo estas por ejemplo canciones de metálica, electrónicas, etc.), acústicas y con poca energía, y aquellos que tienen una proporción equilibrada entre ambas. Como se puede ver en la imagen 4, la mayoría de las canciones de este conjunto se ubican en el primer grupo, el cual tiene sentido, ya que en la descripción del archivo csv en Kaggle (Vergnou, 2021), se menciona que las canciones se tratan de mayormente canciones de rap en Frances, rap americano, rock y música electrónica, esto se puede ver dentro de la relación de los centroides, ya que si bien todos los puntos centroides tienen una distancia considerable entre ellos, los puntos del grupo 1 y 2 están más cercanos entre sí, debido a la gran cantidad de canciones dentro de estas categorías.

Ahora, este análisis puede llevarse a cabo de igual forma con varios valores de *k*, sin embargo, la representación de estos datos se volvería más ambiguo o específico para el propósito de la investigación que se está realizando en el comportamiento entre estas dos variables. Por ejemplo, si se redujera la cantidad de *k* en 2 solo se sabría la diferencia entre los extremos de si son energéticos o acústicos, ignorando aquellos que tienen ambas propiedades. Y por el otro lado, si se aumentara la cantidad de *k*, se aumentaría la cantidad de categorías en aquellos que quedan dentro del grupo de en medio, sin embargo, ya no es necesario detallar estas separaciones si tienen en general un mismo comportamiento, al menos, para el alcance de este reporte.

Bibliografía

Vergnou, B. (2021) *Spotify Recommendation*. Kaggle https://www.kaggle.com/bricevergnou/spotify-recommendation