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 numérico que varía de 0 a 1, las cuales se describen a continuación:

Tabla 1. Variables y sus tipos de datos

Nombre	***Tipo de dato***	***Descripción***
Danceability	Float	Describe que tan apropiado es bailar a una canción dependiendo de diversos factores como el tempo, ritmo, regularidad de la canción, etc.
Energy	Float	La energía representa una medida para la intensidad y actividad en general de la canción, caracterizándose por ser canciones rápidas, con sonidos fuertes y ruidoso.
Key	Int	El *key* o tono tiene que ver con la escala en la que se está tocando la canción de acuerdo con sus notas y composición musical.
Loudness	Float	Esta medida cuantifica que tanto ruido se genera en la canción a partir de la medición de los decibeles (dB) en toda la pista.
Mode	Int	Representa la modalidad de la canción (mayor o menor)
Speechiness	Float	Detecta la presencia de palabras escuchadas en la canción, mientras más palabras, el valor se acerca a 1.0.
Acousticness	Float	Mide que tan confiable es decir que la canción se trata de una canción acústica, 1.0 representa que con mucha seguridad se trata de una canción principalmente con instrumentos acústicos.
Instrumentalness	Float	Predice si la canción contiene no vocales sin contar sonidos producidos por voz que no sean palabras.
Liveness	Float	Detecta que tan probable es que la canción haya sido una grabación de un evento con audiencia en vivo.

Valence	Float	Describe una medida de que sentimiento releva esta canción, sea 1.0 valores como una canción alegre, eufórica o emocionante; mientras que 0.0 sean canciones tristes, con enojo o depresivas.
Tempo	Float	Mide el tempo total de la canción en bits por minuto (BPM), es decir, la rapidez de la duración del bit.
Duration_ms	Int	La duración de la canción en milisegundos.
Time_signature	Int	Mide aproximadamente cuantos bits hay en la canción, en términos generales, el ritmo de la canción.
Liked	Int	Un valor booleano entre 0 o 1 si la canción le gusta al usuario o no.

3. Análisis de datos

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 ello, se seleccionaron variables que puedan llegar a ser significativas para un análisis de correlación entre ellas. Al final, se seleccionaron las variables *Energy* y *Acousticness,* ya que sería interesante realizar una comparativa entre estas dos características en las que, normalmente, se caracterizan por ser contrapuestos en la música, teniendo en cuenta canciones enérgicas como rap, rock, metal, etc. en comparación con canciones acústicas. Para ello, primero valdría la pena revisar los datos que se encuentran en estas dos variables.

Tabla 2. Datos estadísticos de variables

Nombre	***Valor máximo***	***Valor mínimo***	***Media***	***Mediana***	***Desviación estándar***
Energy	0.996	0.0024	0.638	0.659	0.260
Acousticness	0.995	3.05e-06	0.319	0.213	0.321

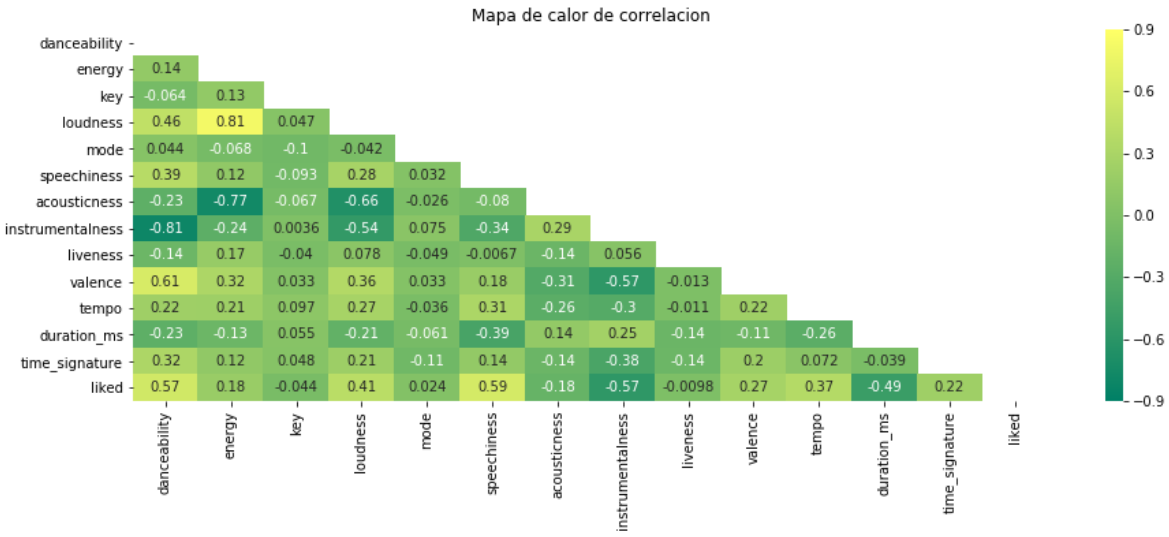
Como se puede ver en la tabla, para empezar, se puede notar que todos los valores obtenidos parecen estar dentro de los rangos esperados, ya que la mayoría de los datos de la tabla contienen valores float que van desde 0 a 1. Adicionalmente, tomando en cuenta la naturaleza de ambas variables, se puede sugerir que las canciones presentes en el conjunto de datos son mayormente enérgicas debido a la media de 0.638, en comparación de una

media de 0.319 de Acousticness, es decir, que en promedio el 30% de las canciones son acústicas, mientras que el resto son enérgicas prácticamente. También es importante constatar que la desviación estándar de ambas variables es relativamente pequeña, con un valor de 0.2 y 0.3, se puede decir que mayormente todos los valores se encuentran muy juntos uno de otro, y que el comportamiento de la mayoría de las canciones sigue estos comportamientos.

4. Análisis comparativo

Para realizar el análisis de la relación que existe entre estas dos variables, fue necesario realizar gráficas de histograma y cajas y bigotes, de modo que se pueda explorar más a fondo la correlación entre ellas, y ver si se puede explicar la correlación negativa que apareció en el mapa de calor al inicio del análisis con las siguientes 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 un posible planteamiento de análisis, sin embargo, para efectos del propósito de la investigación planteada anteriormente, se escogió analizar Energy y Acousticness, con una diferencia bastante

pronunciada entre sus correlaciones de -0.77, haciendo notar que una variable negativa indica que estas dos variables son inversamente opuestas. 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 (Imagen 3), 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

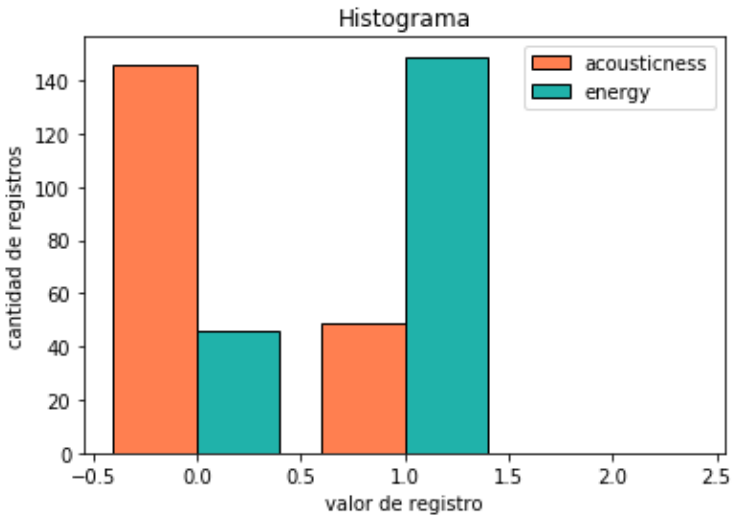
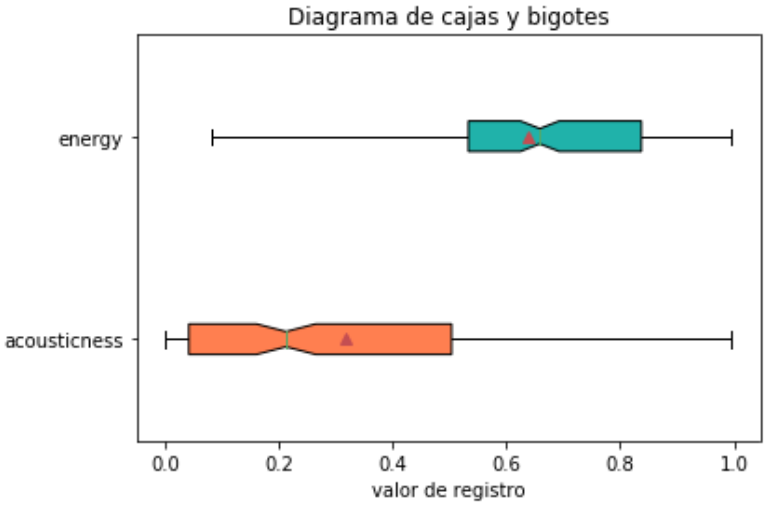


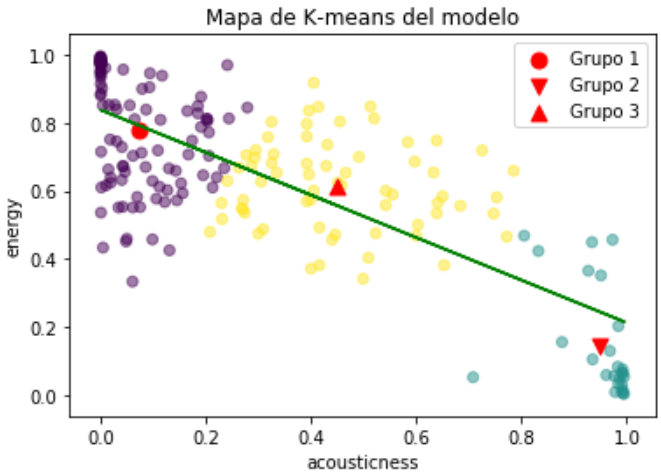
Imagen 3. Diagrama de cajas y bigotes entre Energy y Acousticness



5. 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, gracias a un cálculo de regresión lineal de los datos (código proporcionado por GeeksforGeeks, 2021), 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 con una pendiente de -0.626; 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



$$Intersección\ del\ eje\ y\text{: } 0.839$$

$$Pendiente\text{: } -0.626$$

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

GeeksforGeeks (2021) *Linear Regression (Python Implementation)*. https://www.geeksforgeeks.org/linear-regression-python-implementation/

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