



Actividad: Obtención de estadísticas descriptivas

Alumno: Rafael Hinojosa López

Matrícula: A01705777

TC1002S.570 Herramientas computacionales: el arte de la analítica

Profesor: Gilberto Huesca Juárez

13 de enero del 2021

Campus Nacional

Spotify Recommendation

1. Conjunto de datos

El conjunto de datos utilizado contiene 195 registros de canciones de Spotify donde se evalúan diversos factores de cada canción. El objetivo del conjunto es generar un modelo que prediga si una canción le gustará (o no) a un usuario con base en los datos ya mencionados.

El set de datos que se utilizó en esta actividad fue obtenido de la siguiente liga: <https://www.kaggle.com/bricevergnou/spotify-recommendation?select=data.csv> y pertenecen al usuario de Kaggle Brice Vergnou.

2. Datos y variables

Los 14 campos o variables que se toman en cuenta para la evaluación son las siguientes:

- **acousticness**: medida entre 0.0 y 1.0 que indica si la canción es acústica. 1 representa gran fiabilidad que lo es.
- **danceability**: medida entre 0.0 y 1.0 que indica qué tan “bailable” es una canción. Está calculada a partir de factores como el tiempo, ritmo, estabilidad, regularidad y fuerza del *beat*. 1 representa gran fiabilidad que lo es.
- **duration_ms**: cuánto dura la canción en milisegundos.
- **energy**: medida entre 0.0 y 1.0 que indica qué tan intensa y activa es una canción. Está calculada a partir de la intensidad percibida, timbre, entropía general, tasa de inicio y rango dinámico.
- **instrumentalness**: medida entre 0.0 y 1.0 que indica qué tanto de la canción está representada por instrumentos y sin voz. 1 representa gran fiabilidad que lo es.
- **key**: medida entre 0 y 11 que indica la tonalidad en la que la canción está. 0 representa C (do) y 11, B (si).
- **liveness**: es la probabilidad entre 0 y 1 que haya un público presente en la canción o que haya sido grabada en vivo.
- **loudness**: medida en decibeles (dB) que indica el volumen promedio de la canción, entre -60 y 0 dB.
- **mode**: indica si la tonalidad (*key*) de la canción es mayor o menor. 1 indica que es mayor y 0 que es menor.
- **speechiness**: medida entre 0.0 y 1.0 que indica la presencia de palabras en la canción. Valores de 0 a 0.33 indican un nivel bajo de voz y alto de música,

de 0.33 a 0.66 indica un balance entre voz y música y de 0.66 a 1 indica un nivel alto de voz y bajo de música.

- tempo: indica cuántos golpes o *beats* hay por minuto (bpm) en la canción.
- time_signature: indica cuántos golpes o *beats* hay en cada compás y cuál es la figura musical que representa un golpe.
- valence: medida entre 0.0 y 1.0 que indica qué tan alegre es una canción. 1 representa felicidad, alegría o euforia y 0 representa tristeza, depresión o enojo.
- liked: variable a predecir que es 1 si la canción le gustó al usuario y 0 si no.

Estos datos fueron extraídos a través de la [API de Spotify](#).

El tipo de dato de cada variable se encuentra en la siguiente imagen:

```
Tipos de datos:
danceability      float64
energy            float64
key              int64
loudness          float64
mode             int64
speechiness       float64
acousticness      float64
instrumentalness  float64
liveness          float64
valence           float64
tempo            float64
duration_ms       int64
time_signature    int64
liked            int64
dtype: object
```

Fig 1. Tipos de datos de las variables

El resumen de los datos es el siguiente:

- 14 variables
- 195 registros (canciones)

3. Variables *key* y *tempo*

Escogí las variables *key* y *tempo* por su papel fundamental en la música.

Key

La variable *key* representa tonalidades musicales. La siguiente tabla de conversión ayudará a traducir los valores enteros a letras que representan las mismas.

Valor de la variable <i>key</i>	Valor musical
0	C
1	C#/Db
2	D
3	D#/Eb
4	E
5	F
6	F#/Gb
7	G
8	G#/Ab
9	A
10	A#/Bb
11	B

Tabla 1. Valor musical de los valores de *key*

En la Fig. 2 se puede observar los resultados de la variable *key*, en donde se muestra que la mínima tonalidad es la 0 (C) y la máxima la 11 (B). Junto a este dato se encuentran las tonalidades únicas (*Unique Keys*) que muestran que hay canciones registradas en cada una de las tonalidades posibles. Cabe mencionar que la mínimo valor posible para una *key* es 0, y el máximo, 11.

```
---- Keys ----  
  
Unique Keys:  
[ 0  1  2  3  4  5  6  7  8  9 10 11]  
  
MinKey: 0  
MaxKey: 11  
  
Mean Key: 5.4974358974358974  
Median Key: 6.0  
Std Deviation of Keys: 3.4152090983280465
```

Fig 2. Resultados de la variable *keys*

Observamos que la media de la tonalidad de las canciones escuchadas es 5.49, lo que traducido a tonalidades musicales, es la de F (fa). A pesar de esto, estas tonalidades no se encuentran todas en el mismo modo (mayor o menor), de acuerdo a la variable *mode*. Este hecho puede afectar el análisis que le hagamos a los datos por la diferencia en ambos modos.

Para analizarlos de manera más equitativa lo mejor sería separar las canciones en 2 conjuntos: uno de tonos mayores y otros de tonos menores, o en su defecto, “convertir” los tonos menores cuya escala sea la natural (no se especifica en los datos) a su relativo mayor, mismo que es 3 semitonos (unidades de la variable *key*) mayor a ellos. Ejemplo: el relativo mayor de C menor (*key* = 0) es Eb mayor (*key* = 3).

De igual manera, la mediana es 6.0, que representa la tonalidad de F#/Gb, misma que es la tonalidad que se encuentra en la mitad de las *keys* ordenadas, junto con la de F.

Finalmente, observamos que la desviación estándar es de 3.14, la cual indica que en promedio, la distancia entre alguna tonalidad y la tonalidad media es 3.14.

Tempo

Por otra parte, la variable tempo representa cuántos golpes por minuto hay en la canción. En la Fig. 3 se puede observar los resultados de la variable *tempo*, en donde se muestra que el mínimo presente es 60.171 bpm y el máximo es 180.036 bpm. Podemos observar que la diferencia es muy amplia y que la canción con mayor tempo escuchada es 3 veces más rápida que la más lenta.

```
Unique Tempos:
[ 60.171 60.631 65.023 67.446 68.232 69.363 70.702 71.286 71.428
 71.462 73.679 74.974 75.006 75.026 75.296 75.445 76.503 76.506
 77.507 79.792 79.993 81.548 82.028 82.795 84.907 84.991 86.179
 89.86 90.056 90.664 91.187 92.468 92.628 93.771 94.008 94.032
 94.443 94.992 95.968 96.969 97.346 97.51 97.989 99.046 99.338
 99.97 99.974 100.007 100.047 100.437 101.017 101.052 101.226 102.035
102.757 103.025 103.037 103.048 103.604 103.965 104.964 105.513 106.023
106.275 106.684 107.877 108.004 108.017 108.628 108.674 108.698 108.966
109.041 109.394 110.071 110.547 110.842 110.882 112.019 112.126 112.834
114.223 114.969 115.077 115.908 117.006 119.215 119.303 119.825 119.963
120.026 120.527 120.914 121.063 121.112 122.437 123.041 124.896 124.906
125.059 125.941 126.009 126.025 126.063 127.667 127.693 127.966 128.056
128.553 128.646 129.433 129.466 130.027 131.001 131.716 132.012 132.614
132.979 134.839 134.985 134.992 135.974 136.035 136.059 136.707 136.871
137.043 137.681 138.013 138.027 138.984 139.943 139.959 139.961 139.98
139.981 140.002 140.006 140.025 140.041 140.046 140.061 140.951 142.012
142.03 142.891 142.948 142.95 143.971 143.975 144.481 144.489 145.009
145.121 146.049 146.079 146.154 147.039 147.942 148.084 148.168 149.974
149.978 150.035 150.04 150.067 150.991 151.124 151.329 151.974 152.018
152.085 153.99 154.062 154.986 155.03 155.117 155.999 156.033 156.99
157.738 157.995 159.021 164.032 168.849 169.985 170.054 170.103 172.059
172.068 172.435 176.616 179.661 180.036]

MinTempo: 60.171
MaxTempo: 180.036

Mean Tempo: 121.08617435897436
Median Tempo: 124.896
Std Deviation Tempo: 28.08482882875693
```

Fig 3. Resultados de la variable tempo

Observamos que la media del tempo de una canción es 121.086 y que la mediana es 124.896. Por otro lado, la desviación estándar es 28.08, representando que la variabilidad del tiempo en las canciones es medianamente alto y la variedad de tempos se ve reflejada en la Fig 3.

Podríamos estudiar y analizar cuáles ritmos de música cuentan con un rango similar al obtenido en los datos, basados en muchas canciones, para predecir cuáles fueron los géneros musicales más escuchados por el artista.

Finalmente, se puede concluir que los datos obtenidos muestran gran variedad y variabilidad tanto en las tonalidades y tempos de las canciones o registros. No considero que los valores de las tonalidades y tempos obtenidos puedan definir un

gusto específico por música con cierta tonalidad o tiempo, ya que los datos muestran una importante variabilidad. Así mismo, sería interesante realizar una comparación para observar la relación que hay entre la tonalidad (*key*) y el *tempo*, aunque a nivel musical no estén correlacionadas.

Esto se debe a que la tonalidad a los sonidos que se perciben y cómo se relacionan entre ellos. Por otra parte, el *tempo* es la velocidad con la que debe ejecutarse la pieza musical.

4. KMeans y predicciones

El algoritmo de *k-means* o k-medias se utiliza para clasificar datos en grupos a través de centros aleatoriamente generados y posteriormente relocalizados en la gráfica para estar lo más cerca posible del grupo de puntos que representará.

Continuando con las variables de *key* y *tempo*, las graficamos en una gráfica de puntos y utilizamos el algoritmo de k-medias para clasificar las canciones en grupos y posteriormente predecir en cuál grupo podría estar una canción con cierto *key* y *tempo*.

$k = 3$

Para la primera evaluación, utilizamos la variable k con un valor de 3; es decir, dividiremos el conjunto de datos en 3 subconjuntos.

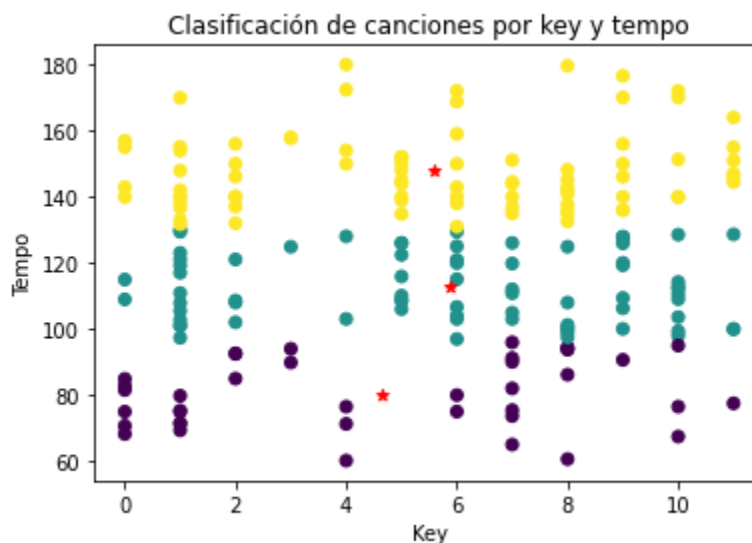


Fig 4. Gráfico de dispersión con $k = 3$

Previamente se comentó que la variable *key* y la variable *tempo* no están correlacionadas musicalmente. En esta gráfica podemos observar la relación que tienen ambas.

Los tres grupos formados están definidos por colores y se observa cómo los grupos cambian según el *tempo*, pero no según la *key*, lo que quiere decir que una canción con cierta *key* o tonalidad puede ser ejecutada en cualquier *tempo* y una canción en cierto *tempo* se puede ejecutar en cualquier *key*; es decir, es una relación N a N entre *tempo* y *key*.

Este hecho apoya a la hipótesis previamente planteada donde se predice que no existe una correlación entre el tiempo y la tonalidad. Esto es porque el cambiar una *key* no significa que se cambiará el *tempo* y viceversa.

$k = 4+$

Pero, ¿qué sucede si el valor de k cambia?

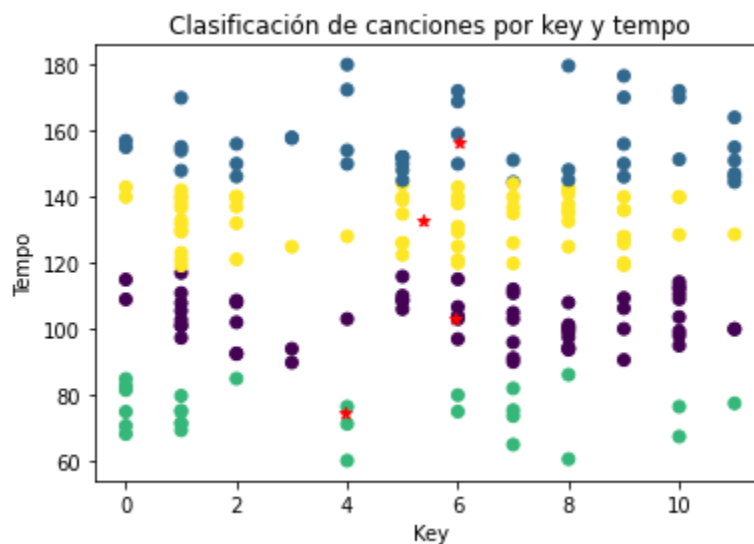


Fig 5. Gráfico de dispersión con $k = 4$

Se logra observar que ahora hay 4 grupos de canciones, pero la correlación entre *tempo* y *key* sigue sin existir.

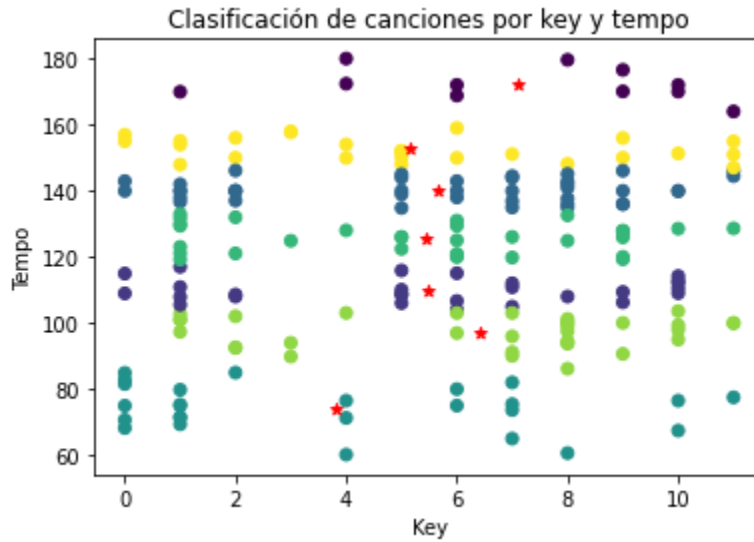


Fig 6. Gráfica de dispersión con $k = 7$

En la Fig 6. Observamos que ocurre lo mismo con $k = 7$, por la misma falta de correlación entre variables explicada previamente. En síntesis, no importa el valor de *key* que utilicemos, el *tempo* podría ser cualquiera.

Centros

Los centros están representados por las estrellas rojas graficadas en las figuras 4, 5 y 6. Estos centros en conjunto están en posiciones que les permiten estar a la distancia mínima promedio de cada uno de sus puntos.

Podemos observar que el centro del grupo con *tempo* más bajo es el más inclinado a la izquierda. Esto se debe a que en el rango de *tempo* en el que está hay más canciones con tonalidades entre 0 y 4 que aquellas con tonalidades entre 5 y 11.

De igual manera, el último centro, correspondiente al *tempo* más elevado se encuentra a la derecha del centro, pues las canciones escuchadas en ese tiempo tiene tonalidades altas.

En contraste, los demás centros se encuentran muy cercanos a la tonalidad 5, que representa a E mayor, y precisamente la media del conjunto de datos es 5.49 (Fig. 2). Esto puede sugerir que si tuviéramos otros datos con una distribución diferente de tonalidades y de tiempos, la hipótesis de que no existe una correlación entre el *tempo* y la *key* podría ser no válida para ese data set; sin embargo, al nosotros contar con canciones con variedad de *keys* y *tempo* podemos observar de manera certera que en este data set no hay una correlación entre ellas.

Cabe mencionar que los grupos de canciones tienen un rango de *tempo* similar entre ellos y que los centros se encuentran cercanos entre sí. Aún así, si tuviéramos muchos *outliers*, todo dependería de dónde están estos para colocar a los centros y definirles grupos. En el análisis de cajas y bigotes, los centros estarían alejados de la mayoría de los valores gracias al *outlier* que afecta a la distancia media que hay entre el centro y los puntos. Si los *outliers* estuvieran agrupados en un mismo sector, entonces sería probable que se definiera un centro para ellos y que los demás centros se mantuvieran con sus grupos.