Tecnológico de Monterrey

Reto: Música y Ciencia de datos

Adamaris Leticia de Dios Ramos[1], Santiago Mora Cruz[1], Lilian Alessandra Rangel Barajas[1] and Gabriel Reynoso Escamilla[1]

[1] *Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias*

Noviembre de 2023

Abstract— Con las plataformas de streaming la manera en la que se escucha música ha cambiado radicalmente, la popularidad de las canciones ya no depende de los sellos discográficos y el radio, en este trabajo se busca utilizar métodos estadísticos para analizar algunos factores que influyen en la popularidad de la música.

Keywords—Música, Popularidad, Estadística, Análisis

I. Introducción

Desde finales de la década de 1990 el panorama de la industria musical comenzó a cambiar, con la popularización del internet se inició un intercambio de música entre usuarios dando lugar a piratería, los sellos discográficos descontentos con la situación buscan soluciones y en un primer momento surgen plataformas como iTunes que permitían a los usuarios comprar y almacenar su colección musical de manera digital, sin embargo un gran porcentaje de compradores no estaban dispuestos a adquirir su música digitalmente, con esa necesidad se lanzaron plataformas de streaming que permitían el acceso a toda la música que quisieran solo con una suscripción [1].

O'Dair y Fry [2] sostienen que con en esta era digital, se transforma la manera en la que se consume la música, se facilita el acceso para descubrir géneros nuevos, menor control de las disqueras y mayor facilidad para la producción amateur, pero también cargan el peso al algoritmo de estas compañías de streaming sobre la popularidad que tendrá una canción o artista.

El propósito de este trabajo es conocer como algunos factores como la energía, el tempo o la "bailabilidad" de una canción afectan su popularidad. Esto se logrará con el análisis de una base de datos musicales usando las técnincas de inferencia estadística aprendidas durante el curso, con el fin de ilustrar datos significativos a través de gráficas efectivas y modernas.

II. Metodología

Hay datos faltantes en nuestra base de datos. Para rellenar los espacios faltantes necesitamos calcular la media o la moda según lo necesario. La media representa el punto de equilibrio de la distribución y está influida por los valores extremos. Proporciona una medida de la tendencia general o valor medio de los datos. La moda nos sirve calcularla cuando tenemos muchos valores repetidos.

Queremos identificar las variables numéricas y las variables categóricas. La variable class es una variable categórica, esta identifica el género de música y a cada género le asignamos un número del 0 al 10 para así poder identificarlos.

Al igual que Class, Key es categórica y esta son las notas musicales, esta es reemplazada con las notas y a cada una le asignamos un número del 1 al 12.

Representaciones gráficas

Los histogramas nos sirven para representar las frecuencias de una variable cuantitativa continua, ayudan a ver el centro, la extensión y la forma de un conjunto de datos. En este caso se aplica para la duración de las canciones.

El diagrama de violin proporciona una vista completa de la distribución de los datos, también muestra si se encuentran valores atípicos. Este es una forma de visualizar la distribución de un conjunto de datos, mostrando no sólo la media y la varianza, sino también todo el rango de los datos.

El radar chart también conocido como gráfico polar nos muestra los datos en dos dimensiones representados en un sistema de coordenadas polar. Este tipo de gráfico es útil para comparar múltiples variables en función de varias categorías y para visualizar patrones en los datos.

El gráfico de pastel es una representación gráfica de datos que utiliza un círculo dividido en sectores para ilustrar proporciones numéricas. Cada sector representa una categoría y su tamaño es proporcional a la cantidad o porcentaje que esa categoría representa en relación con el total.

La matriz de covarianza es una medida que describe cómo dos variables aleatorias cambian juntas. En estadísticas y álgebra lineal, esta matriz se utiliza para cuantificar la relación lineal entre diferentes variables en un conjunto de datos.

Los hex bins, o bins hexagonales, son una forma de visu-

alización de datos utilizada en la visualización de datos bidimensionales, especialmente cuando hay una gran cantidad de puntos de datos y se desea obtener una representación gráfica de la densidad de esos puntos en el espacio.

Transformación de Box-Cox

La transformación de Box-Cox es una transformación potencial que corrige la asimetría de una variable, funciona para corregir sesgos en la distribución de errores, para corregir varianzas desiguales y principalmente para corregir la no linealidad en la relación.

Prueba de Grubbs

La prueba de Grubbs se usa para encontrar un solo valor atípico en un conjunto de datos normalmente distribuido. La prueba encuentra si un valor mínimo o un valor máximo es un valor atípico.

Distribuciones

Existen diferentes tipos de distribuciones, como lo son la normal, de bernoulli, geometrica. La distribución normal tiene forma de campana y está completamente determinada por su media y su desviación estándar, la mayoría de los datos se concentran dentro de la media. La distribución geométrica modela el número de ensayos necesarios para obtener el primer éxito, se utiliza en situaciones donde estás interesado en el número de intentos hasta que ocurra un evento específico. La distribución de Bernoulli es una distribución de probabilidad discreta que modela un experimento aleatorio con dos resultados posibles: éxito o fracaso.

Se tiene un conjunto de datos observados y un modelo estadístico con parámetros desconocidos. La función de verosimilitud describe la probabilidad de observar los datos dados los parámetros del modelo. Se aplica para estimar parámetros en modelos de regresión, distribuciones de probabilidad, y en general, en cualquier situación donde se quieran determinar los valores más probables de los parámetros de un modelo a partir de datos observados.

Prueba de Kolmogorov-Smirnov (KS)

La prueba de Kolmogorov-Smirnov (KS) es una prueba no paramétrica utilizada para determinar si una muestra proviene de una distribución específica. El objetivo es evaluar si hay evidencia estadística para rechazar la hipótesis nula de que los datos provienen de la distribución teórica.

Método percentil de bootstrap

Ordena las estadísticas calculadas para formar una distribución empírica de la estadística de interés. Luego, selecciona los percentiles de esta distribución para construir un intervalo de confianza. El intervalo de confianza más comúnmente utilizado es el intervalo de percentil, que está formado por los percentiles 2.5 y 97.5

Método de bootstrap para CIs bias-corrected and accelerated

La idea detrás del método BCa es corregir el sesgo (bias) y ajustar la aceleración (acceleration) del intervalo de confianza bootstrap. El sesgo se refiere a la diferencia entre el valor esperado de la estadística bootstrap y el valor verdadero de la estadística poblacional. La aceleración se refiere a la asimetría en la distribución de la estadística bootstrap.

Prueba de Shapiro-Wilk

La prueba de Shapiro-Wilk es una prueba estadística utilizada para evaluar si una muestra de datos sigue una distribución normal. El estadístico de la prueba de Shapiro-Wilk se calcula utilizando las desviaciones de los valores observados con respecto a la línea de regresión de los cuantiles esperados bajo la hipótesis nula. Si los datos son aproximadamente normales, el estadístico de prueba será cercano a 1, y si los datos se desvían de la normalidad, el estadístico será menor que 1.

ANOVA

La ANOVA compara las varianzas entre los grupos con la varianza dentro de los grupos, y genera una estadística F y un valor p asociado. Si el valor p es menor que un nivel de significancia previamente establecido (como 0.05), se puede concluir que hay diferencias significativas entre al menos dos de los grupos.

III. Desarrollo

a. Exploración del problema

Se importa la base de datos *"music.csv"* (ver Anexo I: Tablas) como *dataframe* en python.

Después de revisar si hay datos faltantes, encontramos a las variables *key, Popularity* e *instrumentalness*. Calculamos la moda para la variable *key* y la media para *Popularity* e *instrumentalness*

El factor *Class* presentaba variables categóricas, para su analisis fueron codificadas de la siguiente manera

Género	Código
Acoustic/Folk	0
Alternative	1
Blues	2
Bollywood	3
Country	4
Hip-hop	5
Indie	6
Instrumental	7
Metal	8
Pop	9
Rock	10

Y para ilustrar su frecuencia se realizo un gráfico de pastel.

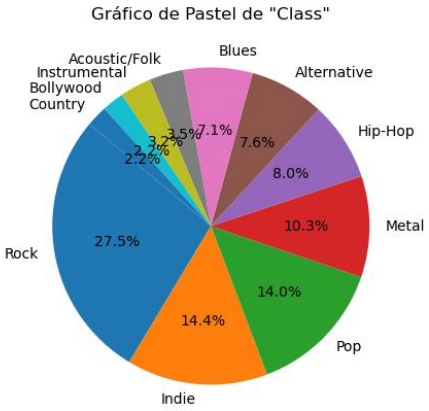

Fig. 1: Gráfica de fracuencia de géneros

Esta gráfica ilustra la proporción de canciones de cada género, los que aparecen con más frecuencia son Rock, Indie

y Pop y los que aparencen con menor frecuencia son Country, Bollywood e Instrumental.

En el factor Key se presenta una situación similar a Class por lo que también fue necesaria su codificación con las claves correspondientes.

Nota	Código
A	1
A♯	2
B	3
C	4
C♯	5
D	6
D♯	7
E	8
F	9
F♯	10
G	11
G♯	12

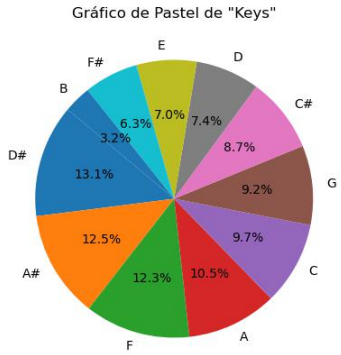

Fig. 2: Gráfica de fracuencia de claves

Este gráfico ilustra la proporción de canciones en cada clave, con mayor frecuencia se presenta D♯, A♯ y F y con menor B, F♯ y E

Tambien se realizó un histograma para explorar la duración de las canciones, en un primer lugar se obtuvo el siguiente.

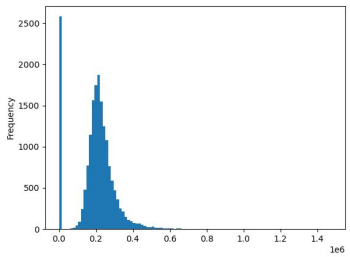

Fig. 3: Histograma de duración de canciones

Se puede observar que la duración de las canciones es incorrecta, no concuerda porque hay algunas que están en minutos y otras en milisegundos, es necesario corregir todos los datos de la columna que estén en minutos a milisegundos, para así tener todos los datos en una misma unidad de medida. Para diferenciar entre cuáles están en minutos y milisegundos, sólo transformaremos aquellas cuya duración sea menor que 100.

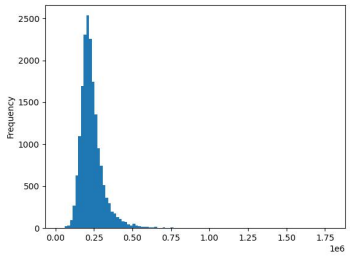

Fig. 4: Nuevo histograma de duración de canciones

Por último se quería conocer la frecuencia de las colaboraciones, ncontramos que el 4.9% son colaboraciones y el 95.1% no lo son.

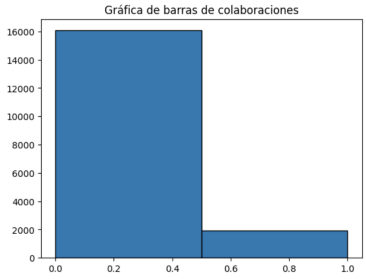

Fig. 5: Gráfico de barras: Frecuancia de las colaboraciones

b. Exploración de los datos

1. Correlación de datos y visualización gráfica

Queremos visualizar como se relaciona la popularidad con algunos factores, utilizaremos la información de la matriz de correlación, y hacer gráficas que nos permitan observar su relación con los cuatro factores con los que tiene mayor correlación

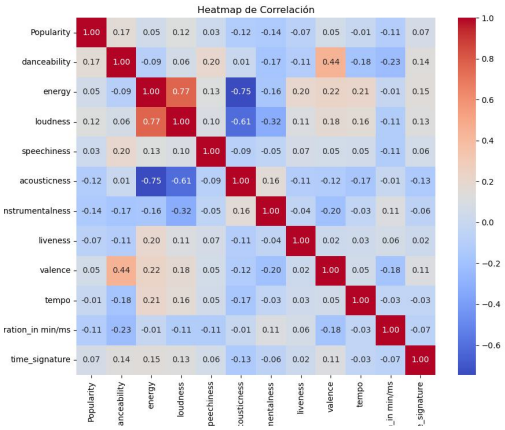

Fig. 6: Correlación de las variables numéricas

Popularidad Está mayormente relacionada con danceability [Fig. 7], energy, loudness y time signature (ver Anexo II Fig. 1) por la cantidad de mediciones por analizar utilizamos *hexagonal binning* para su representación gráfica .

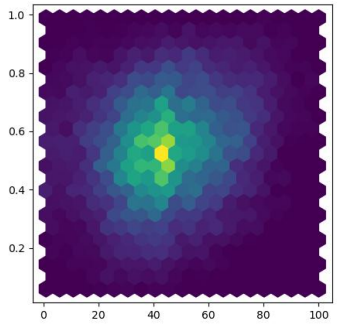

Fig. 7: *Hex-bins* de popularidad

Género Se decidió ilustrar la relacion entre el género y los siguientes factores: danceability [Fig. 8], energy, acousticness y tempo (ver Anexo II Fig. 2), para la visualización se usaron gráficos de violin.

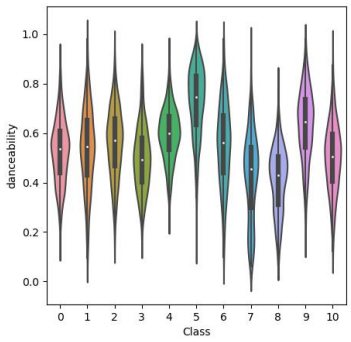

Fig. 8: Gráficos de violín de género

2. Sesgo de los datos y transformaciones de Box-Cox

Primero, exploramos las distribuciones de las variables numéricas de la base de datos para ver cuales tienen exceso de sesgo y cuales tienen valores atípicos (ver Anexo II Fig. 3). Se puede observar, que los datos correspondientes a las variables Energy [Fig. 9], Instrumentalness [Fig. 10], Liveness [Fig. 11] y Duration [Fig. 12] tienen una distribución sesgada, por lo cual les aplicaremos la transformación de Box-Cox: Recordemos que para "duration" se modificaron los datos que estaban en minutos a milisegundos.

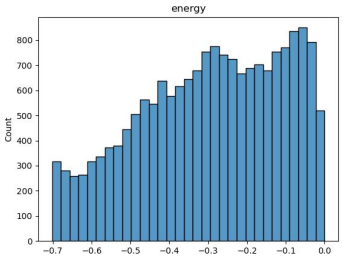

Fig. 9: Nuevo histograma de energy

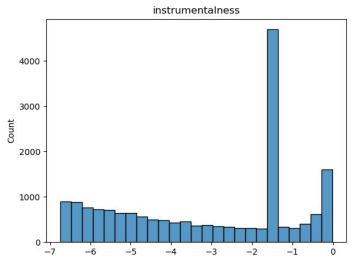

Fig. 10: Nuevo histograma de instrumentalness

3. Canciones más populares

Las diez canciones más poluares son:

1. MONTERO (Call Me By Your Name) de Lil Nas X
2. Beggin' de Maneskin
3. good 4 you de Olivia Rodrigo
4. Kiss Me More (feat. SZA) de Doja Cat
5. Bad Habits de Ed Sheeran

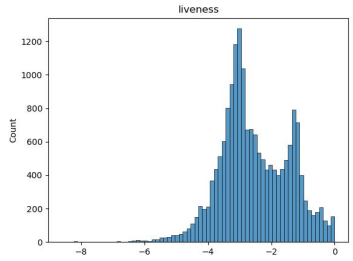

Fig. 11: Nuevo histograma de liveness

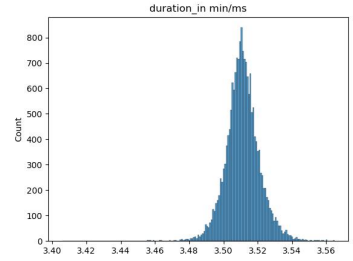

Fig. 12: Nuevo histograma de duración

6. Astronaut In The Ocean de Masked Wolf
7. STAY (with Justin Bieber) de The Kid LAROI
8. RAPSTAR de Polo G
9. Butter de BTS
10. Todo De Ti de Rauw Alejandro

Para ilustrar como las variables influyen en las canciones anteriores hicimo *Radar Charts*, ver (Anexo II Fig. 4), a continuación el *Radar chart* de MONTERO (Call me by your name)

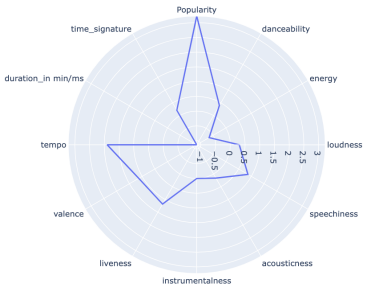

Fig. 13: Radar Chart canción más popular

c. Análisis Estadístico

1. Distribución de las variables

Determina qué distribución de probabilidad sería adecuada para modelar la popularidad, el modo, la valence, el tempo y la duración de las canciones. Visualizando el histograma de cada una de las variables (ver Anexo II Fig. 3), realizamos un KS test con la distribución que más se le parezca, y realizamos un KS test con parámetros estimados, y analizamos el p-valor para ver la validez de nuestra predicción.

Para **popularidad** probamos una distribución normal

- $H_0 : X \sim N(\mu, \sigma^2)$
- $H_a : XnsimN(\mu, \sigma^2)$

El resultado de la prueba fue la siguiente:

- Estadístico de prueba = 0.3333
- P-valor = 0.7777

Poedmos concluir que la popularidad sigue una distribución normal

Para **valence** probamos una distribución normal

- $H_0 : X \sim N(\mu, \sigma^2)$
- $H_a : X \nsim N(\mu, \sigma^2)$

El resultado de la prueba fue la siguiente:

- Estadístico de prueba = 0.3333
- P-valor = 0.7777

Poedmos concluir que valence sigue una distribución normal

Para **tempo** probamos una distribución normal

- $H_0 : X \sim N(\mu, \sigma^2)$
- $H_a : X \nsim N(\mu, \sigma^2)$

El resultado de la prueba fue la siguiente:

- Estadístico de prueba = 0.5842
- P-valor = 0.1675

Poedmos concluir que tempo sigue una distribución normal

Para **duración** probamos una distribución beta

- $H_0 : X \sim \text{Beta}(\alpha, \beta)$
- $H_a : X \nsim \text{Beta}(\alpha, \beta)$

El resultado de la prueba fue la siguiente:

- Estadístico de prueba = 0.3541
- P-valor = 0.7216

Poedmos concluir que la duración sigue una distribución beta.

2. Método de máxima verosimilitud para la estimación de los parámetros de las distribuciones.

Para la **popularidad** los parámetros estimados para la distribución normal son:

- Media estimada: 44.5121
- Desviación estandar estimada: 17.2179

Para la distribución beta de la **duración**, se estimaron:

- Alpha: 8.3807
- Beta: 6641533454996.996
- Loc: 22283.2001
- Scale: 1.6947×10^{17}

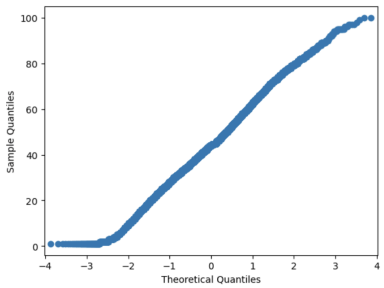

Fig. 14: QQ plot de popularidad

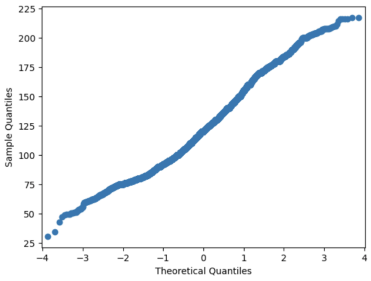

Fig. 15: QQ plot de duración

Para la distribución normal de **valence** se estimaron:

- Media estimada: 0.4862
- Desviación estandar estimada: 0.2402

Por último para la distribución normal de **tempo** se estimaron:

- Media estimada: 122.6233
- Desviación estandar estimada: 29.5707

3. Prueba de Kolmogorov-Smirnov

Para saber su nuestros modelos son útiles realizaremos pruebas de Kolmogorov-Smirnov y QQ plots, como estimamos los parámetros usando nuestra muestra, no podemos utilizar directamente la prueba de Kolmogorov-Smirnov. Tenemos que simular la distribución del estadístico de prueba.

Popularidad [Fig. 14]

Estadistico de prueba: 0.9919819534909304 P-value: 0.0 Se rechaza la hipótesis nula. Los datos no siguen la distribución normal

Duración [Fig. 15]

Estadistico de prueba: 1.0 P-value: 0.0 Se rechaza la hipótesis nula. Los datos no siguen la distribución normal

Valence [Fig. 16]

Estadistico de prueba: 0.5120657712541103 P-value: 0.0 Se rechaza la hipótesis nula. Los datos no siguen la distribución normal

Tempo [Fig. 17]

Estadistico de prueba: 1.0 P-value: 0.0 Se rechaza la hipótesis nula. Los datos no siguen la distribución normal

Podemos observar en las pruebas de Kolmogorov-Smirnov y en los QQ plots generados que, a pesar de que en los histogramas parece, y en los p-values obtenidos de las pruebas de KS nos muestran que no hay argumentos para rechazar la hipótesis nula (que las variables están distribuidas de cierta forma), los QQ-plots nos muestran que las distribuciones que siguen las variables en cuestión no se comportan en su totalidad de manera normal, hipergeométrica o beta, respectivamente. Habrá que hacer más pruebas con estadísticos de prueba para ver si las variables realmente se comportan conforme a las distibuciones en cuestión.

En base a estos resultados decidimos hacer una nueva prueba para conocer a que distribución se aproximaban más, esta fue usando `distfit()` en python, obtuvimos los siguientes resultados:

Propuesta de Distribución	
Variable	Distribución
Popularity	Lognorm
Valence	Beta
Tempo	Beta
Duration in min/ms	Lognorm

Con estos nuevos modelos calculamos algunas probabilidades:

- Probabilidad que la popularida de las canciones sea mayor o igual 70: 0.9306
- Probabilidad de que valence sea mayor o igual que 0.65: 0.7523
- Probabilidad de que el tempo sea menor o igual a 100: 0.2221
- Probabilidad de que la duración de las canciones dure menor o igual de 2 min y medio aproximadamente: 0.0783

4. Intervalo de confianza con Bootstrap

Intervalos de confianza para popularidad:

- Mediana: (44.0, 44.51212431693989)

Intervalos de confianza para valence:

- Mediana: (0.474, 0.488)
- Desviación estándar: (0.23865548, 0.24171956)
- Media: (0.48329683, 0.48919652)

Intervalos de confianza para tempo:

- Mediana: (119.997, 121.01951249999999)
- Desviación estándar: (29.34881435, 29.79738266)
- Media: (122.26326074, 122.99166855)

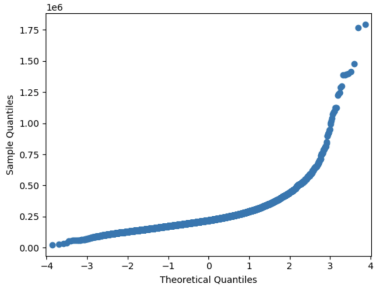

Fig. 16: QQ plot de valence

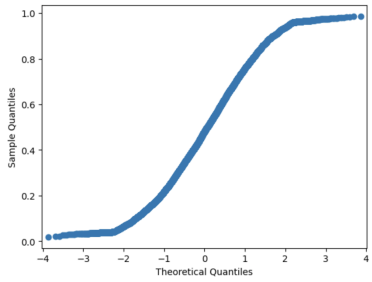

Fig. 17: QQ plot de tempo

Intervalos de confianza para duración:

- Mediana: (218996.0, 221199.5025)
- Desviación estándar: (82533.66474928, 89370.15310393)
- Media: (234784.0776677, 236895.22717343)

5. Prueba de Saphiro-Wilks

La prueba de Saphiro-Wilks es útil para probar si un componente sigue una distribución normal. Su hipotesis nula es una distribución normal, mientras que la alternativa es una distribución no normal, de manera más formal:

$$H_0 \sim N(\mu, \sigma^2)$$

$$H_a \nsim N(\mu, \sigma^2)$$

Se utilizó para probar la normalidad de la popularidad.

- Estadístico de prueba: 0.9966
- P-valor: 8.5423×10^{-20}

Con una significancia de $\alpha = 0.05$, se rechaza la hipótesis nula, no hay evidencia que la popularidad siga una distribución normal.

También hicimos las pruebas al separar la popularidad en canciones en colaboración y en solitario.

Canciones con colaboración

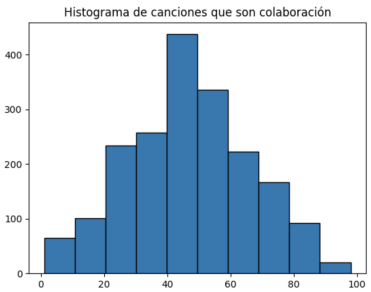

Fig. 18: Histograma: Popularidad de canciones en colaboración

- Estadístico de prueba: 0.9953
- P-valor: 8.5423×10^{-6}

Con una significancia de $\alpha = 0.05$, se rechaza la hipótesis nula, no hay evidencia que la popularidad de las canciones en colaboración siga una distribución normal.

Canciones sin colaboración

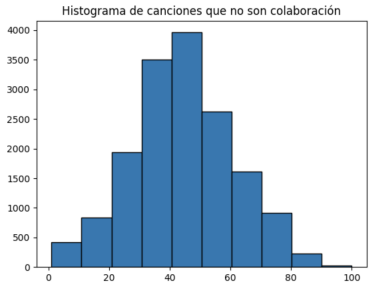

Fig. 19: Histograma: Popularidad de canciones sin colaboración

- Estadístico de prueba: 0.9966
- P-valor: 1.2764×10^{-18}

Con una significancia de $\alpha = 0.05$, se rechaza la hipótesis nula, no hay evidencia que la popularidad de las canciones en colaboración siga una distribución normal.

Queremos conocer si la colaboración afecta la popularidad. Proponemos que la media de las canciones en colaboración es mayor que la de las canciones sin colaboración, para esto realizaremos una prueba de diferencia de medias *t-student*:

- $H_0 : \mu_1 - \mu_2 = 0$
- $H_a : \mu_1 - \mu_a > 0$

Obtenemos los siguientes resultados:

- Estadístico de prueba: 6.6592
- P-valor: 1.4164×10^{-11}

El P-valor es menor que la significancia $\alpha = 0.05$ por lo que se rechaza la hipótesis nula en favor de la alternativa, la media de las canciones en colaboración es mayor que la de las canciones sin colaboración.

6. ANOVA para popularidad y género

El análisis de la varianza (ANOVA) es un método estadístico útil para el estudio de las medias de tres o más conjuntos. Se utilizó este método para encontrar diferencias significativas entre la media de la popularidad de cada género. Las hipótesis:

$$H_O : \mu_0 = \mu_2 = ... = \mu_{10}$$

$$H_a : \text{No todas las medias son iguales}$$

- Estadístico F: 200.0691
- P-valor: 0.0

Como el P-valor es menor que la significancia $\alpha = 0.05$, se rechaza la hipótesis nula, hay diferencia entre las medias.

IV. Conclusión

El propósito del trabajo era analizar cómo diversos factores afectan la popularidad de las canciones, utilizando técnicas estadísticas y visualizaciones modernas.A pesar de algunos desafíos en la modelización de las distribuciones de datos, se exploran y estiman parámetros para comprender mejor la relación entre la música y los datos recopilados.

REFERENCES

[1] J. Zimmer, Scott, "Rise of music streaming." *Salem Press Encyclopedia*, 2021. [Online]. Available: http://0-search.ebscohost.com.biblioteca-ils.tec.mx/login.aspx%3fdirect%3dtrue%26db%3ders%26AN%3d129814641%26lang%3des%26site%3deds-live%26scope%3dsite

[2] M. O'Dair and A. Fry, "Beyond the black box in music streaming: the impact of recommendation systems upon artists." *Popular Communication*, vol. 18, no. 1, pp. 65 – 77, 2020. [Online]. Available: http://0-search.ebscohost.com.biblioteca-ils.tec.mx/login.aspx%3fdirect%3dtrue%26db%3dasx%26AN%3d141719281%26lang%3des%26site%3deds-live%26scope%3dsite

1 Anexo

1.1 Anexo I: Tablas

Arist Name	Track Name	Popularity	...
Bruno Mars	That's What I Like (feat. Gucci Mane)	60.0	...
Boston	Hitch a Ride	54.0	...
The Raincoats	No Side to Fall In	35.0	...
Deno	Lingo (feat. J.I and Chunkz)	66.0	...
⋮	⋮	⋮	⋮

Cálculo y análisis de estadísticas básicas

	Media	Desviación estandar	Varianza	Rango
Popularity	44.51	17.22	296.47	99.0
Danceability	0.54	0.17	0.03	0.9294
Energy	0.66	0.24	0.06	0.9999
Loudness	-7.91	4.05	16.4	41.3070
Mode	0.64	0.48	0.23	1
Speechiness	0.08	0.08	0.01	0.9325
Acousticness	0.25	0.31	0.1	0.996
Instrumentalness	0.18	0.26	0.07	0.9960
Liveness	0.2	0.16	0.03	0.9881
Valence	0.49	0.24	0.06	0.9677
Tempo	122.62	29.57	874.48	186.8590
Duration ms	235823.28	85675.03	7340210798.47	1769840
Time signature	3.92	0.36	0.13	4

Cuartiles

	0.25	0.5	0.75
Popularity	33.0	44.0	56.0
Danceability	0.43	0.55	0.66
Energy	0.51	0.7	0.86
Loudness	-9.54	-7.02	-5.19
Mode	0.0	1.0	10
Speechiness	0.03	0.05	0.08
Acousticness	0.0	0.08	0.43
Instrumentalness	0.0	0.08	0.43
Liveness	0.1	0.13	0.26
Valence	0.3	0.48	0.67
Tempo	99.62	120.07	141.97
Duration ms	187649.75	220000.0	263082.25
Time signature	4.0	4.0	4.0

Coeficientes de asimetría y kurtosis

	Coef. asimetría	Kurtosis
Popularity	0.08	-0.15
Danceability	.0,08	-0.28
Energy	-0.66	-0.32
Loudness	-1.76	5.04
Mode	-0.57	-1.68
Speechiness	3.09	12.67
Acousticness	1.11	-0.18
Instrumentalness	1.76	1.92
Liveness	2.18	5.63
Valence	0.09	-0.92
Tempo	0.38	-0.45
Duration ms	4.03	39.9
Time signature	-4.18	27.89

1.2 Anexo II: Gráficas

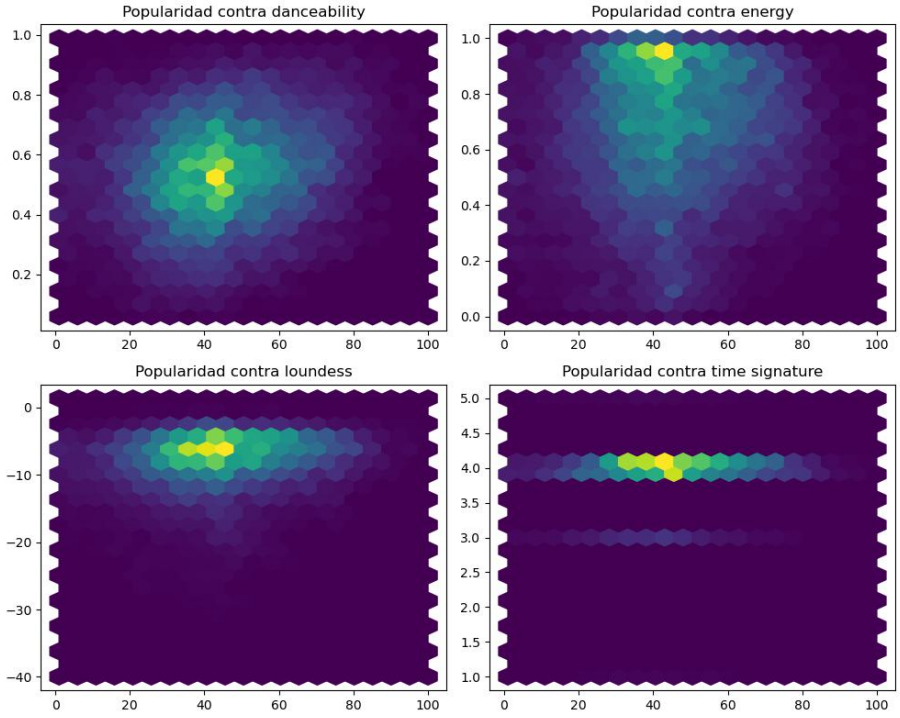

Figure 1: Gráficos *hex-bin* de popularidad

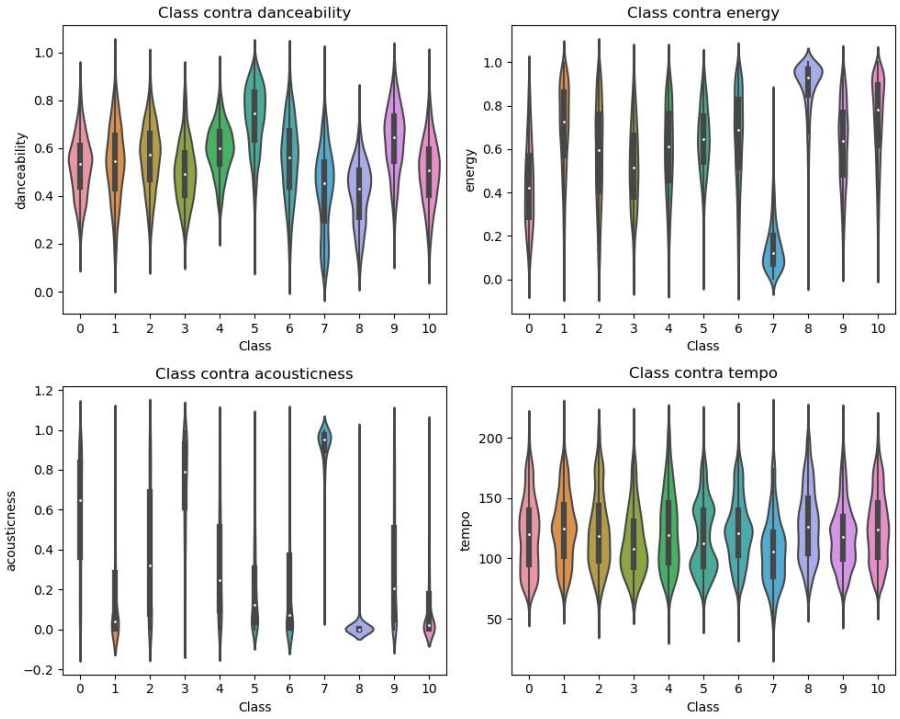

Figure 2: Gráficos de violín de género

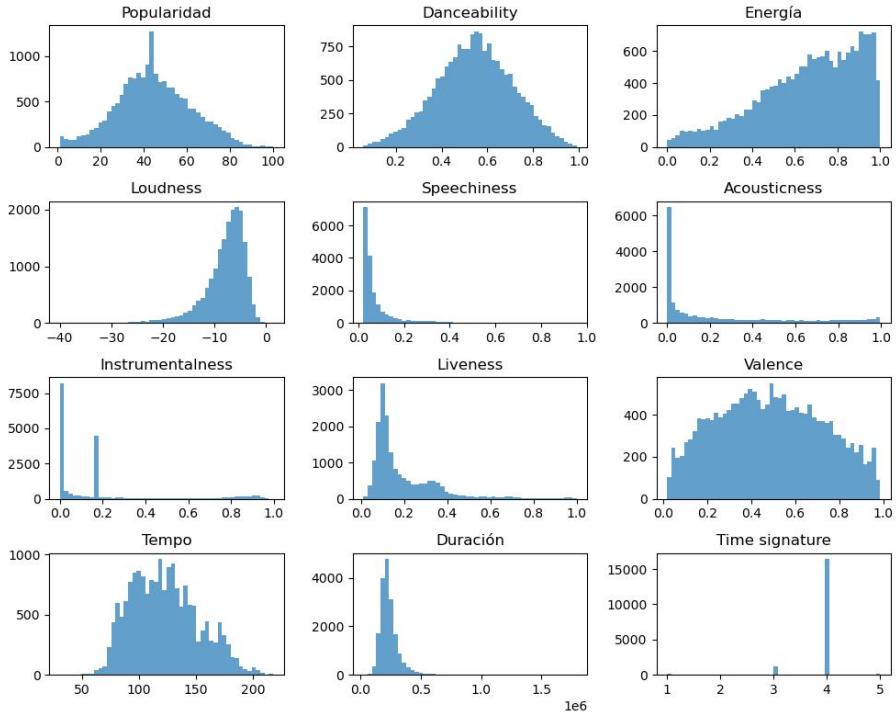

Figure 3: Histogrmas de las variables

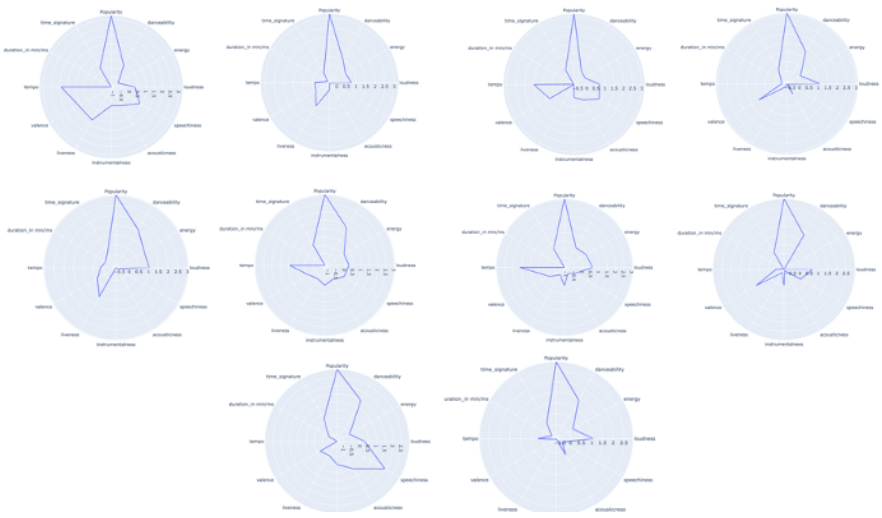

Figure 4: Radar charts de las canciones más populares