Sábado, 03 de junio de 2023

Proyecto Integrador Ciencia de Datos

Universidad Panamericana ECID

Número 10

Fabian Adolfo Cervantes Romero

1. Definición del Problema

1.1Contexto

En la década de los 90's la industria de la música vivió uno de sus mejores momentos con enormes ganancias debido a su modelo de negocios basado en la venta de material en formato físico (casetes, discos, vinilos). Las ventas ya superaban los 3,000 millones de copias vendidas a finales de los 90.

Con la invención del formato mp3, vino un cambio revolucionario para esta industria, ya que nacieron los servicios de archivos compartidos, en su momento considerados ilegales, que no proveían ningún tipo de ganancia a los propietarios de los derechos. Toda esta transformación obligo a las grandes disqueras a redefinirse frente a el mundo globalizado y tecnológico, lo que dio como resultado la distribución del material de manera digital y lo que hoy conocemos como servicios de streaming.

Actualmente existen distribuidores de música online que pueden enlazar a los artistas con todos los consumidores a escala global y estar al mismo nivel de accesibilidad de los artistas más cotizados de la música popular.

De este modo cualquier artista puede llevar su proyecto a gran parte del mundo y que su material pueda ser disfrutado por más consumidores.

Hoy en día existen varios servicios de análisis que te permiten conocer tus métricas de alcance y comportamiento de su producto en el mercado. Algunas como: Youtube Sudio, Awario, Chartmetric, etc.

Fig.2. Chartmetric y Awario, herramientas para análisis musical.

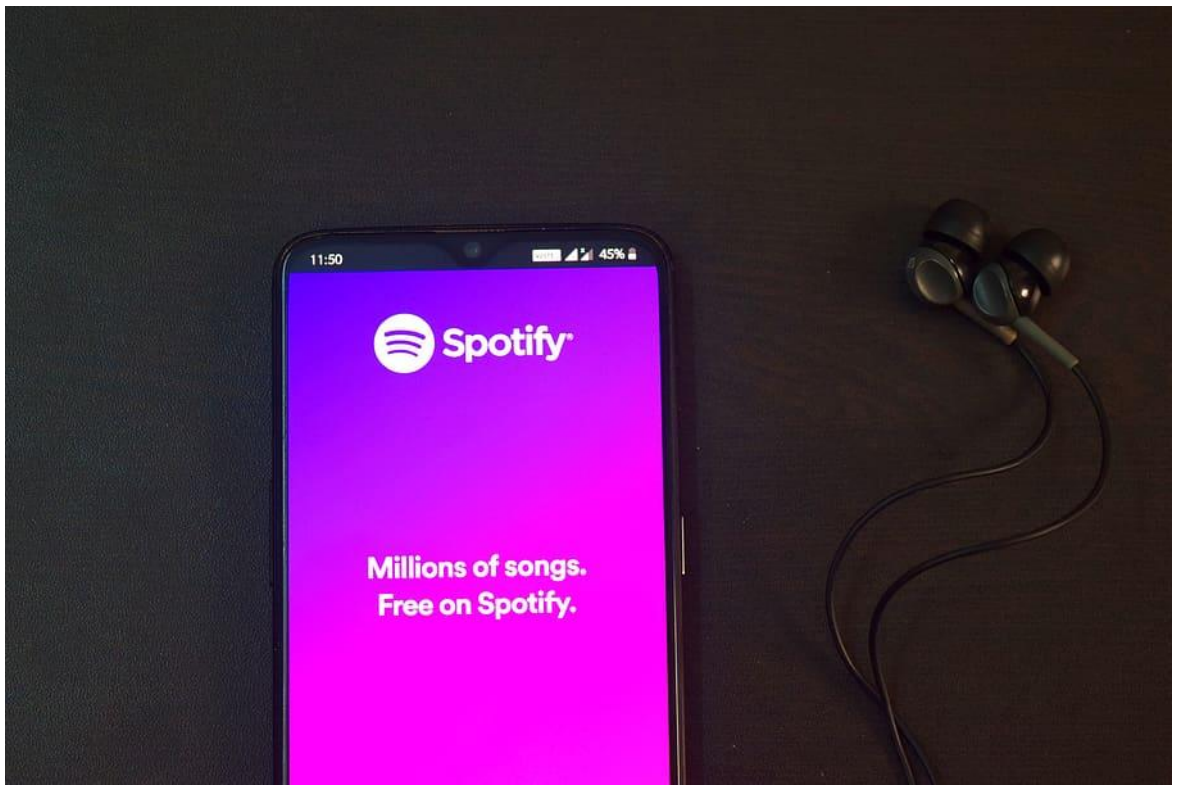

Fig. 1. Spotify. El servicio de donde se obtuvieron los datos para esta investigación.

Fabian Adolfo

1.2 Declaración del problema

El éxito en el mundo del streaming

Así como la tecnología ha abierto nuevas puertas para que los músicos puedan compartir su arte con el mundo, esto ha generado más competitividad entre los que se unen a este mercado buscando destacar y subir hasta la cima.

Esto ha llevado a los artistas o disqueras en la actualidad a buscar ser más eficaz, no solo en talento, composición o escritura de las obras, sino de igual manera en su impacto, la producción, mercadeo y estrategia de ventas. La industria de la música ha dejado a un lado la venta en formato físico por lo que hoy en día la ganancia se centra en el espectáculo, hacer tu proyecto más accesible, más rápido y ligero de consumir para poder avanzar a la par de la sociedad, sus gustos y necesidades.

Para las compañías dedicadas a la creación, distribución o publicación de material musical entender a los consumidores es crucial para lograr posicionarse dentro de los más escuchados y tener mayor volumen de consumo.

Si pudiéramos conocer cuáles son los factores que comparten las pistas y los artistas que más figuran en las listas de éxitos, tal vez, podríamos determinar las variables que arraigan más para que una pieza se vuelva reconocida y goce de gran número de reproducciones, lo que se traduce en ganancias, ya sea por las mismas reproducciones o generando revuelo en la audiencia, lo que te abrirá nuevas puertas en el mundo del espectáculo.

Poder pronosticar el reconocimiento, éxito o, en otras palabras, hacer una estimación de las ganancias que puede traer tu composición para ti o tu sello discográfico es de gran ayuda por lo que podrían analizar mejor su estrategia de venta y promoción de los proyectos musicales.

Con una ruta de publicidad bien planificada se pueden maximizar el impacto y reducir la perdida monetaria en gastos innecesarios a la hora de firmar o llevar la carrera a un artista emergente.

Este trabajo de investigación se realizó con el fin de poner a prueba los temas y los diferentes métodos vistos en la especialidad en ciencia de datos, mediante la realización de este aprendimos nuevas cosas para aplicar a nuestros conjuntos de datos, comportamientos, análisis, metodologías y algoritmos que existen para obtener un mejor desempeño en nuestros modelos de aprendizaje máquina.

El tema de esta investigación basado en la música y su consumo se derivan de mi interés en conocer más sobre hábitos de consumo, comportamiento del mercado y el gusto que tengo esta disciplina del arte. La oportunidad que se nos presentó de combinar cosas que nos apasionan y cosas nuevas para nosotros como lo fue la ciencia de datos resulto muy motivadora. Con esta investigación se intenta ayudar a los artistas, sellos discográficos o promotores, ya sean nuevos o consolidados, que tengan la intención de crecer su mercado y sus oportunidades.

La idea de este proyecto surge de poder otorgar una herramienta para que su debido uso pueda ofrecer mejor entendimiento del negocio y como tu como artista puedes ser más consistente en las escuchas de tu material. La relevancia que puede alcanzar esta investigación es de gran importancia para reducir el nivel de perdidas o aumentar las ganancias de las composiciones musicales, de igual manera puede contribuir a la realización de música más y mejor aceptada por la sociedad, contribuyendo con la satisfacción de esta misma e incluso en la economía.

En cuanto a viabilidad de la investigación, contamos con conocimientos iniciales en ciencia de datos, manejo de librerías para manipular datos y contamos con el apoyo de grandes profesores y profesionistas en el área que nos brindaron ayuda para lograr los resultados de nuestros proyectos.

Esta solución esta basada en estudiar las canciones con mayor numero de reproducciones del 2017 al 2021, por lo que estamos analizando lo mejor que hay para poder encontrar patrones y variables que afecten la popularidad y cantidad de escuchas de las obras para así poder darle un mejor manejo y rumbo a nuestras composiciones o carrera. Encontrar tendencias futuras para poder estar a la vanguardia sobre los gustos del consumidor y poder aprovechar el momento. Los artistas que deseen mantener o iniciar una carrera fructífera se pueden ver beneficiados por este trabajo.

Nuestra solución analítica de datos nos proporciona ayuda para poder encontrar nuevas oportunidades de crecimiento, encontrar nuevos mercados, si es que no hemos explorado y como llegar a ellos, y conocer la relevancia de nuestra inversión en marketing. Por esta razón buscamos tener la información correcta a la hora de tomar decisiones, gracias a los datos podemos acercarnos a las verdaderas causas, si es que existen.

La intención de mi solución es que puedas obtener lo mejor de cada oportunidad de llegar a la cima que se presenta conociendo previamente los factores que favorecen el desarrollo y buen crecimiento dentro de las plataformas. Tener una sana estrategia de evolución en tu carrera para que tu y tus fans solo reciban lo mejor. No solo te ayudará a estar listo y preparado para cada área de oportunidad en tu carrera, sino que también te acompaña a la hora de respaldar tus decisiones frente a ejecutivos y/o cuando estes en busca de promotores y agentes. Si estas a cargo de un sello discográfico y te interesa firmar a los nuevos talentos antes que alguien más lo haga, este trabajo puede darte buena idea de donde buscarlo y que puedas apoyar los artistas nuevos en la escena y sus prometedoras actuaciones.

1.3 Justificación de la solución

Decidí trabajar con una regresión lineal múltiple debido a que es un modelo ampliamente utilizado en el sector de pronósticos y análisis de predicción, nos brinda una herramienta de estadística muy bien estructurada que nos permitirá modelar relaciones lineales entre múltiples variables predictoras independientes y solo 1 variable objetivo dependiente.

Si existe una relación lineal entre las variables predictoras, ya sean características de las pistas, artista, genero, promoción, popularidad, etc. Y el número de streams, la regresión lineal ampliamente nos puede ayudar a captar la realidad de estas relaciones y comprender como es que cada variable influye en el número de reproducciones de cada canción.

Dicho modelo nos permite incorporar múltiples variables predictoras, lo que nos resulta de gran ayuda para estudiar este fenómeno, ya que no esperamos que el éxito de una canción depende completamente de una variable, garantizándonos una mejor estimación y relaciones con la variable objetivo.

La regresión lineal múltiple nos brinda coeficientes para cada variable predictora del modelo, este coeficiente es sumamente útil ya que nos da a conocer el impacto que tiene cada una en el número de reproducciones. De la misma manera estos coeficientes te explican la dirección y la fuerza con la que se relacionan las variables.

La regresión lineal cuenta con varias técnicas de evaluación y validación para determinar la precisión del modelo, por lo que buscaremos dividir nuestro conjunto de datos en sets de entrenamiento, prueba y validación, Con estos diferentes grupos de datos identificaremos la eficacia y la capacidad de generalización del modelo.

2. Implementación

2.1 Análisis descriptivo

Nuestro conjunto de datos lo obtuvimos de Kaggle y este contiene registros de el top 200 de canciones más escuchadas por semana mundialmente desde el 2017 hasta el 2021, lo que nos daba un total de 74,661

registros y cada uno conformado de 40 columnas (2,986,440 items en total).

2.1.1 Exploración de los datos

Los datos vienen capturados en 10 tipos diferentes, como smallint, char, varchar, decimal, float, bool, date, etc. Analizando las columnas con las que contamos en el dataset, encontramos que varias de ellas no agregan ni restan importancia a nuestra investigación o modelo, por ejemplo: enlaces de referencias, imágenes o ids de los artistas o de las mismas canciones. Por eso los primeros pasos de la limpieza de datos consistieron en remover esas columnas, de igual manera identificamos que había una columna don el nombre de "Pivot" que básicamente lo que nos indicaba era el artista principal (valor de 0) de una pisa si es que esta era una colaboración (valor de 1 para artistas invitados); en este paso nos quedamos con los registros que eran igual a 0.

Después reordenamos el conjunto de datos en orden descendiente mediante la columna "streams" y procedimos a eliminar valores repetidos en la columna de "track_name", esto porque teníamos varios registros de la misma canción en diferentes momentos, decidí trabajar con la que más reproducciones había logrado. Una vez eliminado los datos repetidos nos hicimos cargo de los valores nulos, que afortunadamente, solo había uno. Después trabajamos con los datos faltantes que solo se presentaban en 2 categorías y por último transformamos nuestras variables booleanas en 0s y 1s, mientras que las categóricas las transformamos a números, esto para que el modelo pueda trabajar con ellas ya que solo entiende números.

	track_popularity	album_type	album_label	album_popularity	artist_followers	artist_popularity
count	2100.000000	2100.000000	2100.000000	2100.000000	2.100000e+03	2100.000000
mean	62.485238	0.763333	262.224762	61.098095	2.112422e+07	80.524286
std	14.830446	0.964194	152.845405	17.193845	2.331422e+07	10.458242
min	1.000000	0.000000	0.000000	1.000000	4.600000e+02	4.000000
25%	57.000000	0.000000	114.000000	52.000000	4.501457e+06	75.000000
50%	64.000000	0.000000	264.000000	64.000000	1.182438e+07	82.000000
75%	72.000000	2.000000	400.000000	73.000000	2.941922e+07	88.000000
max	92.000000	2.000000	517.000000	86.000000	1.040598e+08	100.000000

Fig.3. Descripción de algunas columnas del conjunto

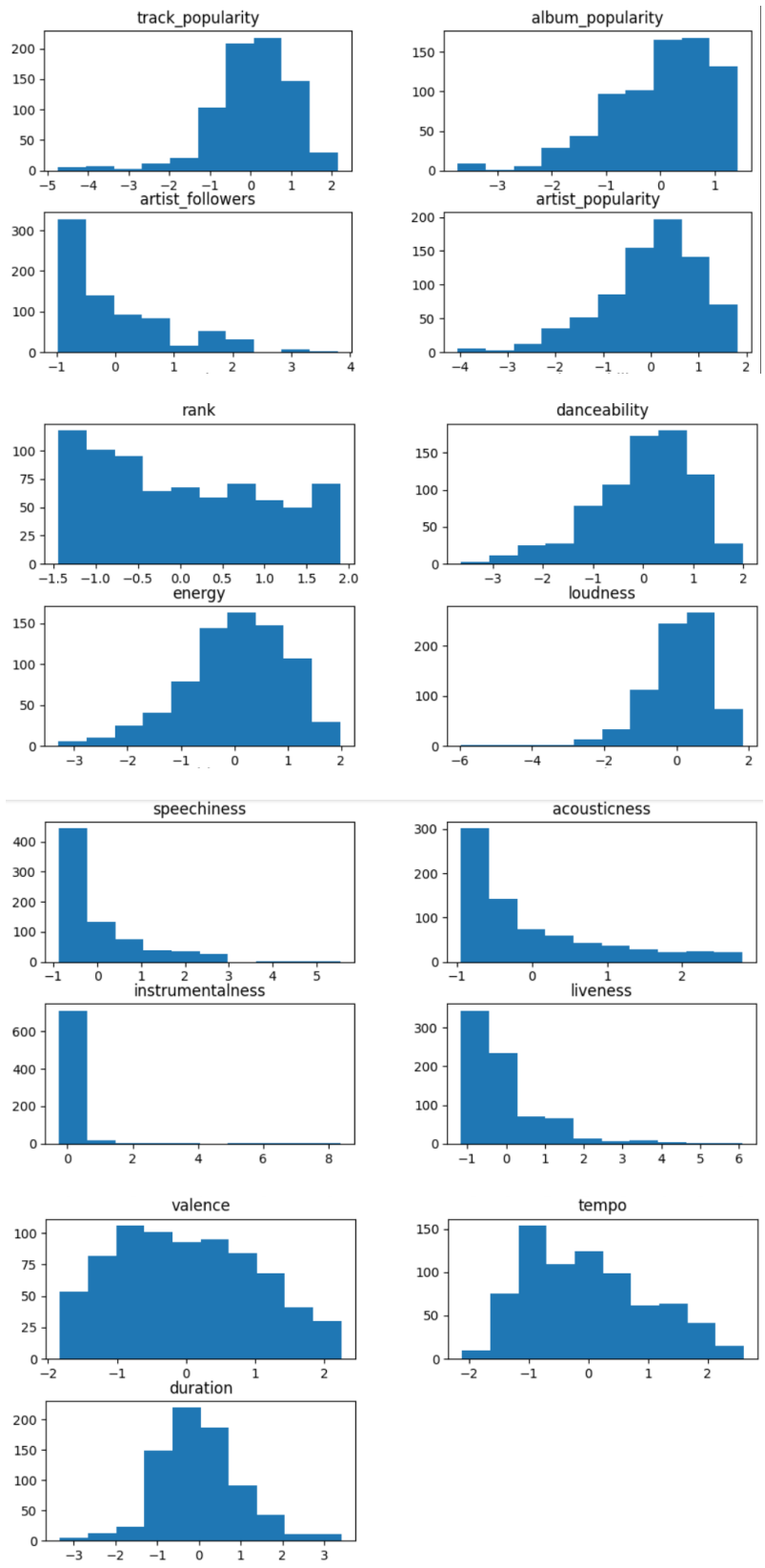

Fig.4. Histogramas de las posibles variables predictoras

2.1.2 Distribución

Fig.8 muestra estadísticos descriptivos de las principales columnas de nuestro dataset, podemos observar que nos quedamos con 2100 registros en total, el promedio de popularidad es de 80 en los artistas que figuran en esa lista, pero hay artistas que alcanzan el 100 y por el contrario tenemos el más bajo que es de 4, este podría ser un caso de alguna canción que se haya hecho muy viral repentinamente y metió a el artista en el top, podría ser el caso.

Los datos no mostraban una distribución normal por lo que aplicamos una normalización para que sea más fácil trabajar con ellos, a continuación, mostraremos los histogramas de las variables.

En la fig.4 se observan las diferentes formas de las distribuciones las variables de nuestro conjunto de datos, podemos ver que la que se ajusta más a la forma de campana es la variable de "duration", que nos indica la duración en milisegundos de las canciones, mientras que las variables que describen de manera más técnica la estructura de las canciones muestran asimetría. En la columna del número de seguidores del artista se puede apreciar que a la lista de canciones más populares también entran artistas con no demasiados seguidores, aumentando la dispersión de los datos. En las columnas de speechiness, acousticness, instrumentalness y liveness vemos que la distribución muestra asimetría, pero muy marcada, lo que me hace pensar que las canciones más escuchadas comparten valores muy similares en esos 4 aspectos. Instrumenalness no me sorprende mucho, ya que entre menor sea el valor,

significa que la canción no es pura instrumental, es decir, a los consumidores les gustan las canciones con vocales, algo que me parece algo de esperar.

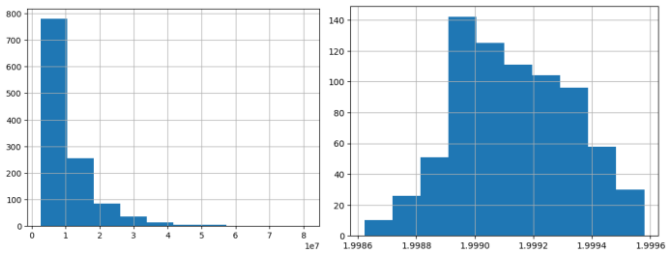

Fig.5. Histogramas de la variable streams(streams)

La Fig.9. muestra los histogramas de la variable streams, nuestra variable objetivo, antes y después de ser transformada. Al principio mostraba asimetría hacía la derecha y dificultaba su manejo a la hora de hacer el modelo, los resultados aumentaron muy considerablemente después de transformarla con el método de BoxCox porque no se comportaba de manera normal.

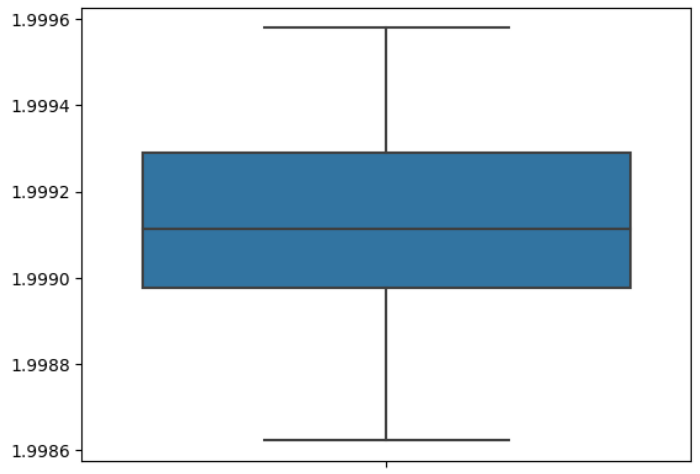

Fig.6. Diagrama de caja y bigote de la variable streams

Como se puede observar en esta gráfica, la variable ya no cuenta con valores atípicos y la media está por el valor de 1.9991.

2.1.3 Correlación

Para la correlación de las variables en nuestro conjunto de datos imprimimos la matriz de correlaciones que nos brinda pandas, con ella pude ver la correlación entre las columnas del conjunto de datos de manera numérica y poder conocer que variables están altamente relacionadas entre ellas y puedan afectar el modelo si es que entran las dos o más al modelo. También con estas graficas de correlación buscamos ver que variables presentan correlación con nuestra variable objetivo y poder darnos una idea de cuales podrían tener mayor significancia a la hora comprender el comportamiento de la variable dependiente.

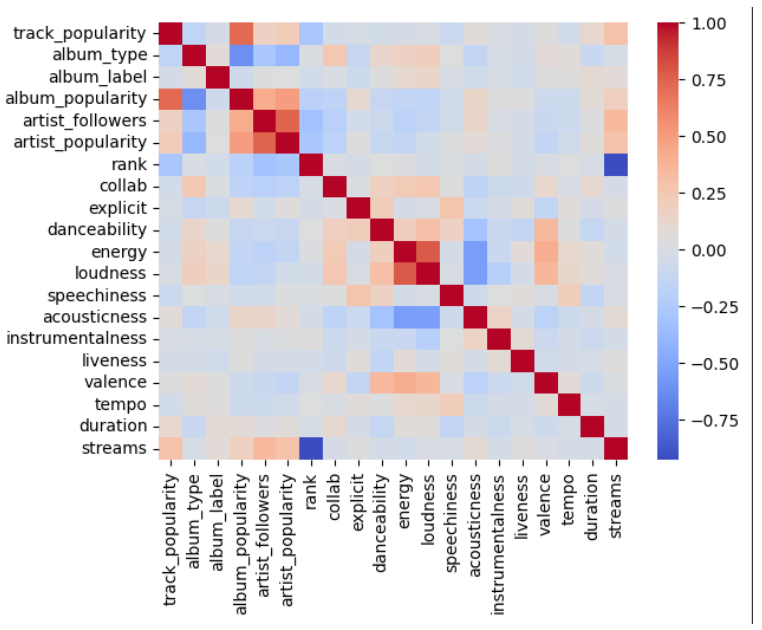

Fig.7. Mapa de calor de las correlaciones dentro de los datos

Fig.7. muestra las correlaciones que hay entre las variables, pero asignándoles un color dependiendo de qué tanto se relaciona con ella positiva o negativamente, con esto quiero decir que la pendiente cuente con estas características. En la imagen podemos apreciar que hay bastante correlación entre las variables que denotan la popularidad, ya sea del artista, del álbum o de la canción. Con esa premisa partiremos a la hora de meter variables predictoras a el modelo, podemos descartar alguna y volver a revisar si aún conserva un valor alto de correlación. Tiene sentido que dichas columnas muestren alta relación entre ellas ya que si el artista tiene muchos seguidores pues a mas personas llega su álbum y de igual manera su popularidad se ve afectada positivamente en la mayoría de los casos. Me llama mucho la alta relación que se muestra entre la popularidad de la canción y la popularidad del álbum, ya que en ocasiones que una canción tenga mucho enganche entre los consumidores no significa que el disco sea también de su agrado, los conocidos “One Hit Wonder” que son éxitos que tiene un grupo o artista, pero sus demás temas no alcanzan los mismos niveles de éxito en el mercado.

Por otro lado, podemos notar que la relación de el rango y los streams es muy fuerte pero negativamente, algo que me hace sentido completamente ya que el rango lo que muestra es la posición de dicha pista en el top 200 de canciones más escuchadas. Entonces lo que se busca es que ese valor vaya disminuyendo, o sea, acercándose a 1, que vendría siendo el top 1, la canción más escuchada del momento y para llegar a esa posición hay que sumar reproducciones.

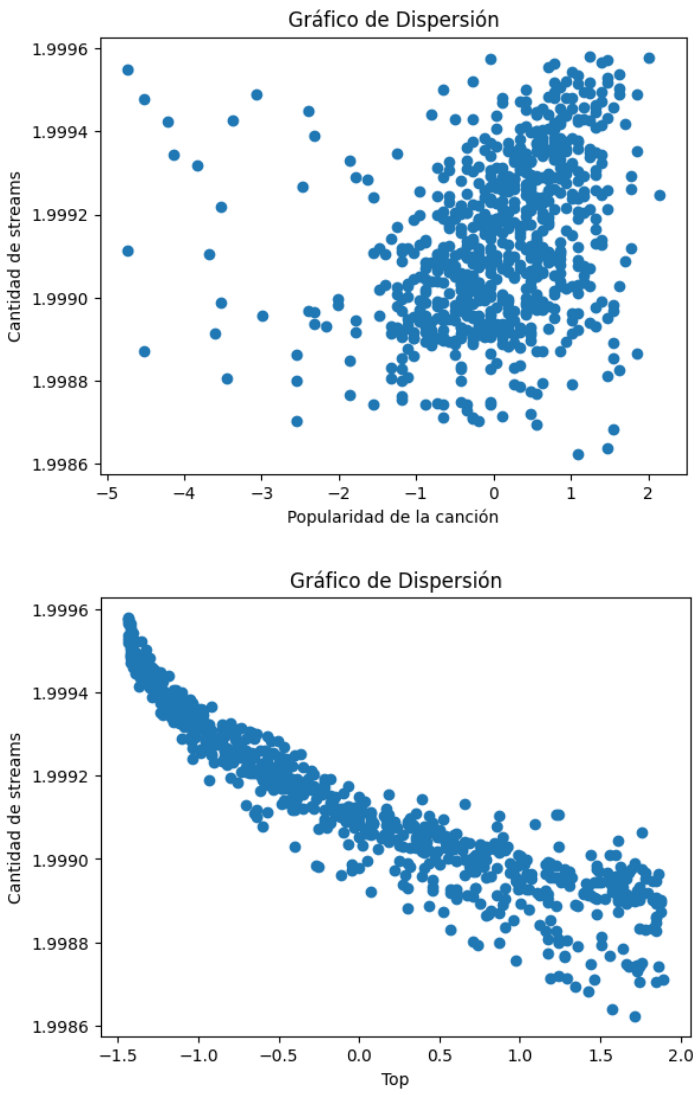

Fig.8. Gráfico de dispersión rank y popularidad de la canción contra la cantidad de streams

En la figura 8 podemos observar los gráficos de dispersión de las 2 variables con correlación más fuerte con en número de reproducciones de las canciones, cada punto representa una observación del conjunto, se puede apreciar más notable la relación lineal entre la posición en la lista y las reproducciones, como si fuera un patrón. Mientras que en la grafica de popularidad de la canción cuesta un poco imaginarla ya que también hay observaciones muy alejadas de nuestra tendencia.

2.2 Explicación de la Metodología

La regresión lineal es una técnica de la estadística que se enfoca en comprender por qué suceden las cosas, explicar comportamientos. Lo que la regresión lineal nos permite es darle forma a un modelo lineal con el que se intenta calcular el valor de una variable, Y, mediante el uso de otras variables que son independientes entre ellas. La variable que se busca pronosticar, recibe el nombre de: objetivo y las variables con las que se intenta llegar a ese valor se les llama: predictores, con la forma X1,X2,X3 por ejemplo.

Haciendo uso de esta técnica podemos no solo pronosticar valores, sino que también nos van a ayudar a notar los factores que tienen peso sobre ella, esto es de mucha ayuda porque podrías encontrar patrones de comportamientos en los fenómenos que te interese estudiar, es como encontrar la formula secreta para llegar a ese algo que estas buscando.

Como toda matemática, los modelos se basan en una ecuación que se ve así:

$$Y_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni}) + e_i$$

Fig.9. Ecuación de los modelos de regresión lineal múltiple

Ahora bien, vamos a hablar sobre los términos y que significan dentro de está ecuación. La variable "Y", como ya sabemos, es el valor que estamos intentando pronosticar correctamente y viene dado por los valores que tomen las demás variables en la formula. El coeficiente B0 es la intercepción, el valor que toma la variable "Y "en el momento que todos los predictores del modelo son 0. Después de ese coeficiente con subíndice 0, los siguientes coeficientes de B incrementan su subíndice y estos nos indican la intensidad y la dirección de la relación entre la variable independiente y la dependiente:

- Cuanto más lejano a 0 es, más fuerte es la relación
- El signo en el coeficiente indica la dirección (si aumenta o disminuye la variable independiente, la variable dependiente sufre un aumento)

El coeficiente "e" es el residuo, la diferencia que existe entre el valor observado y el valor que el modelo estima.

Para poder aplicar esta técnica es necesario cumplir con ciertos requisitos en nuestro fenómeno a estudiar:

1. La variable que deseamos pronosticar, variable dependiente, debe ser numérica u bien ordinal, pero necesita tener varias categorías, de preferencia mayor a 5 categorías.
2. Las variables independientes de igual manera necesitan ser escalares, ordinales o dummy (que se puedan representar con 0 y 1)
3. Las variables independientes que usemos en el modelo no pueden presentar un alto nivel de correlación entre ellas.
4. Las relaciones entre las variables independientes y la variable dependiente deben ser lineales.
5. Todos los residuales de las variables deben seguir la distribución normal y deben mostrar varianzas iguales.

Si bien es importante cumplir con los requisitos de la regresión lineal múltiple, en la realidad se vuelve un poco complicado que los datos satisfagan todas. Analizando los datos puedes ir revisando si tu conjunto de datos cumple con cada uno de los requisitos, pero la 5ta condición es hasta la implementación del modelo que lo sabemos, de igual manera existen diferentes técnicas para tratar por si se incumple con alguna de ellas.

Para analizar una regresión lineal múltiple es de suma importancia comprender estos valores:

i. Significación del F-test: este valor nos va a decir si sí el modelo es estadísticamente significativo y por consecuencia las variables independientes

explican “algo” la variable independiente, ese algo es el valor de R-cuadrada.

ii. R-cuadrada: indica el porcentaje de varianza de la variable dependiente explicado por el conjunto de variables independientes. Entre mayor R-cuadrado más explicativo y mejor será el modelo.

iii. Significancia del t-test: si es menor a 0.05 es que esa variable independiente se relaciona de forma significativa con la variable dependiente, influye sobre ella, es explicativa, ayuda a predecirla.

iv. B-eta: indica intensidad y dirección de la relación entre la variable dependiente y la independiente.

Los dos primeros valores hacen nos son de gran utilidad para estudiar la bondad del modelo, si existe la posibilidad de que el conjunto de variables predictoras se relaciona con el resultado. Los dos siguientes valores nos hablan de la influencia de cada una de las variables independientes. Por ejemplo, si alguna variable te da un valor de t-test mayor a 0.05 y podrías optar por desecharla de tu conjunto de datos y no se vería afectada la bondad del modelo, eso sí, es necesario hacer un análisis antes de desechar una variable. Un ejemplo sería que tengas 2 variables con valor t-test mayor a 0.05, en ese caso puedes desechar la que más correlacionada este con las demás, ya sea analizando la matriz de correlación o con el método del valor Factor de Inflación de Varianza (VIF). El valor VIF se calcula de la siguiente manera:

$$VIF_{\hat{\beta}_j} = \frac{1}{1 - R^2}$$

Fig.10. Formula del factor de inflación de varianza

En este proyecto con las siguientes métricas para dicho valor:

- Si es igual a 1 no existe colinealidad
- Si está entre 1 y 4, existe la posibilidad de estar afectando la regresión

2.2.1 Mínimos Cuadrados Ordinarios (OLS)

En este caso, mis modelos de regresión los hice por medio de el método OLS, esto debido a que tiene el objetivo de minimizar la suma de diferencias cuadradas entre los valores que le para entrenar y los que el modelo predice.

Lo que esta técnica intenta es minimizar el valor del residuo, que se ve algo así:

$$e_i = y_i - \hat{y}_i$$

Fig.11. Cálculo de residuo

Donde Y es el valor observado y Y gorrito es el valor de la predicción, teniendo eso en mente y que el modelo ya está alimentado con los valores de X y los valores de y. Buscamos calcular la recta que mejor se ajuste a esos puntos, ya sabemos que los factores que varían en la ecuación de la recta, Fig.9, son los coeficientes del modelo, las Bs. Con eso en mente necesitamos conocer dichos coeficientes en los que el residuo sea el menor posible, ¿Cómo encontramos eso?, buscando la ecuación que la suma de los residuos nos del valor menor posible.

El problema que podemos llegar a encontrar en esto es que al ser una diferencia aleatoria podemos encontrar valores dentro del rango de los positivos y los negativos lo que podría causar confusiones a la hora de calificar el modelo, por ejemplo, si ves errores cercanos a ceros puedes pensar que el modelo funciona muy bien, pero puede ser el caso contrario. Es por esto por lo que se busca una manera de hacer que los errores con valores menores a cero se eliminen con los valores mayores a cero y elevando al cuadrado justamente es que podemos lograr esto. Entonces nuestro error se convierte en esto:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

Fig.12. Nueva formula para calcular el residuo

En resumen, este método nos permite encontrar la recta de regresión que dé el menor de los valores en la suma de los cuadrados de los residuos.

2.3 Resumen de la Implementación

Una vez que tenemos los datos listos, ya limpios, sin datos nulos, sin valores repetidos, ni valores faltantes, sin valores atípicos y transformados podemos proceder a el modelo.

2.3.1 División del conjunto de datos

Nuestro conjunto de datos comenzó con un total de 74660 registros y después de todo el proceso de limpieza nos quedo de un tamaño de 1178 filas, este es el conjunto que usaremos para separar en el set de prueba y validación.

Dividimos de la siguiente manera los conjuntos:

- Dividimos en 80% y 20% para entrenamiento y prueba
- Dividimos en 80% y 20% de entrenamiento y validación

Al final nuestros conjuntos se veían así:

```
tamaño del conjunto de datos original:  (1178, 19)
tamaño del conjunto de de entrenamiento :  (753, 19)
tamaño del conjunto de test:  (236, 19)
tamaño del conjunto de validación:  (189, 19)
```

Fig.13. Tamaño de los conjuntos de datos utilizados

Al momento de dividirlos utilicé los parámetros de test_size = .02 para que le asignara ese 20% a el set de entrenamiento y validación, random_state para que

pueda dividírmelos de la misma manera cada que ejecutemos la instrucción, y, por último, shuffle = para mezclar los datos antes de la separación de los conjuntos.

Antes de llegar a el modelo final hicimos muchos otros para probar diferentes técnicas y procedimientos para quedarnos con las mejores variables. Vamos a hablar de el modelo que decidí usar al final porque es el que utiliza el mayor número de herramientas y procesos, que a fin de cuentas es la intención de esto, aprender y experimentar con los diferentes modelos y recursos que hemos estado viendo en clases e investigando.

2.3.2 Selección de Variables

Para este modelo final utilicé sequential feature selector (SFS) que nos va a ayudar a determinar qué variables son las predictoras son las que más peso tienen en este modelo. Este proceso a lo que nos ayudó es a no tener tantas variables en el modelo y poder predecir de manera más fácil y eficiente. La dirección en que realizamos esta búsqueda o eliminación de variables fue "Backward", esto quiere decir que el modelo inicia con todas las variables predictoras y va eliminando las que menos significativas son hasta que cumple con un criterio de parada, en mi caso fue encontrar 6 variables predictoras. El SFS fue de bastante ayuda ya que evita el proceso de estar creando modelos e ir sacando variables una por una para ver como va reaccionando el modelo. Otro parámetro que nos permite modificar SFS es el de "Cross-Validation" qué lo que va a hacer es validación cruzada en cada iteración del proceso y en mi caso le dije que hiciera 10 'folds' o particiones para validar el rendimiento del modelo con diferentes subconjuntos de los sets de prueba y validación. La ventaja de utilizar 'cross-validation' en mi modelo es que me ayuda controlar, dentro de lo que se puede, el sesgo al dividir en más conjuntos mis datos.

Al final el procedimiento de sequential feature selector nos regresó las mejores 6 variables predictoras que encontró:

1. album_label
2. artista_followers
3. rank
4. acousticness
5. valence
6. duration

Acto seguido creamos un dataset con esas columnas para crear nuestro modelo.

2.4 Resultados

Una vez corrido y entrenado el modelo estos fueron los resultados que me dieron:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.867
Model:	OLS	Adj. R-squared:	0.866
Method:	Least Squares	F-statistic:	809.7
Date:	Sun, 04 Jun 2023	Prob (F-statistic):	1.14e-322
Time:	03:27:45	Log-Likelihood:	6106.7
No. Observations:	753	AIC:	-1.220e+04
Df Residuals:	746	BIC:	-1.217e+04
Df Model:	6		
Covariance Type:	nonrobust		

Fig.14. Resumen del modelo

En la Fig.14. podemos ver que el modelo nos da una R^2 de 86 que considero es un valor alto, pero también existía la preocupación de que el modelo estuviera sobre entrenado y a la hora de encontrarse con nuevos datos no diera un buen rendimiento, pero eso lo veremos más adelante. Recordemos que la fig.14 muestra el modelo del que partimos para llegar a el mejor, ese no es el final.

Después revisamos que no existiera alta correlación entre mis variables predictoras:

	vif
track_popularity	1.121546
album_label	1.015298
artist_followers	1.143939
rank	1.194548
acousticness	1.024509
duration	1.021890

Fig.15. Factor de Inflación de la Varianza de las variables

Podemos ver que no existe multicolinealidad entre las variables porque mis valores VIF son de 1, nada por lo que preocuparse.

Después trabajamos con los puntos de influencia, que vendrían siendo los valores que más impacto tienen en la estimación de parámetros y pueden afectar el modelo. Para hacernos cargo de estos puntos de apalancamiento utilizamos "La Distancia de Cook's", la cual nos ayuda a visualizar los puntos que tiene un impacto enorme en los resultados del modelo.

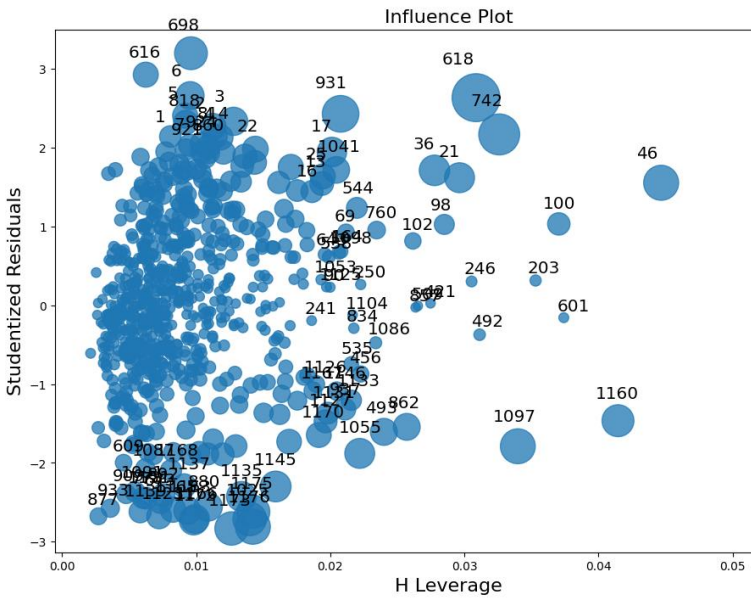

Fig.16. Puntos de influencia: Cook distance

Con esta grafica pude identificar qué puntos tenían mayor impacto, los que su tamaño es mayor, y gracias a que el número que muestra la gráfica es el índice, los pude localizar en el conjunto de datos para analizarlos y ver que tenían en común.

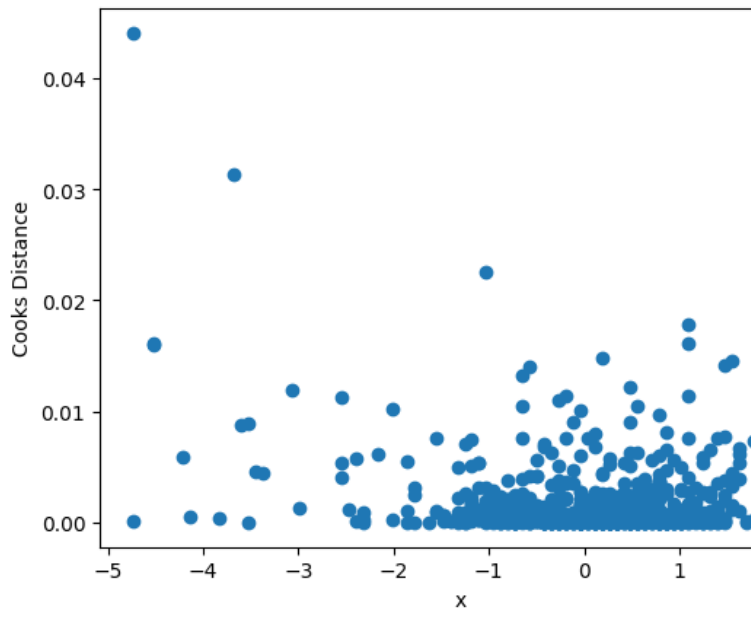

Fig.17. Grafica de Cook's distance

Ya con esta información removimos los puntos con alta influencia y creamos el nuevo conjunto de datos para nuestro modelo final, el que decidimos utilizar.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.878
Model:	OLS	Adj. R-squared:	0.877
Method:	Least Squares	F-statistic:	870.9
Date:	Sun, 04 Jun 2023	Prob (F-statistic):	0.00
Time:	03:27:48	Log-Likelihood:	5983.6
No. Observations:	731	AIC:	-1.195e+04
Df Residuals:	724	BIC:	-1.192e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.9991	2.52e-06	7.94e+05	0.000	1.999	1.999
track_popularity	9.774e-06	2.92e-06	3.343	0.001	4.03e-06	1.55e-05
album_label	6.327e-06	2.5e-06	2.526	0.012	1.41e-06	1.12e-05
artist_followers	1e-05	2.85e-06	3.513	0.000	4.41e-06	1.56e-05
rank	-0.0002	2.81e-06	-62.369	0.000	-0.000	-0.000
acousticness	5.708e-06	2.67e-06	2.138	0.033	4.66e-07	1.09e-05
duration	-5.847e-06	2.59e-06	-2.261	0.024	-1.09e-05	-7.7e-07

Omnibus:	3.513	Durbin-Watson:	1.957
Prob(Omnibus):	0.173	Jarque-Bera (JB):	3.637
Skew:	-0.092	Prob(JB):	0.162
Kurtosis:	3.292	Cond. No.	1.65

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fig.18. Resumen del modelo final

Como podemos ver en la Fig.18. el valor de R^2 de nuestro modelo aumentó a 87

Tiene un F-statistic de 0.00, lo que significa que el modelo es estadísticamente significativo y las variables independientes explican 87% de la varianza de la variable independiente.

Todas las variables que utilizamos en el modelo tienen valor t-test menor que 0.05 por lo que esas variables independientes se relacionan de forma significativa con la variable dependiente. Los valores AIC y BIC son cercanos a cero por lo que son una buena opción para considerar a la hora de escoger modelo. Un valor bajo en estas variables nos da idea de un mejor ajuste del modelo y una mayor capacidad de generalización.

Volvemos a revisar la multicolinealidad de las variables:

	vif
track_popularity	1.167736
album_label	1.018740
artist_followers	1.163112
rank	1.234262
acousticness	1.031662
duration	1.020922

Fig.19. Nuevo Factor de Inflación de la Varianza de las variables

Notamos que los valores son de 1 entonces no hay evidencia de alta correlación entre mis variables predictoras.

2.4.1 Una vez hecho y validado mi modelo vamos a revisar las suposiciones del modelo:

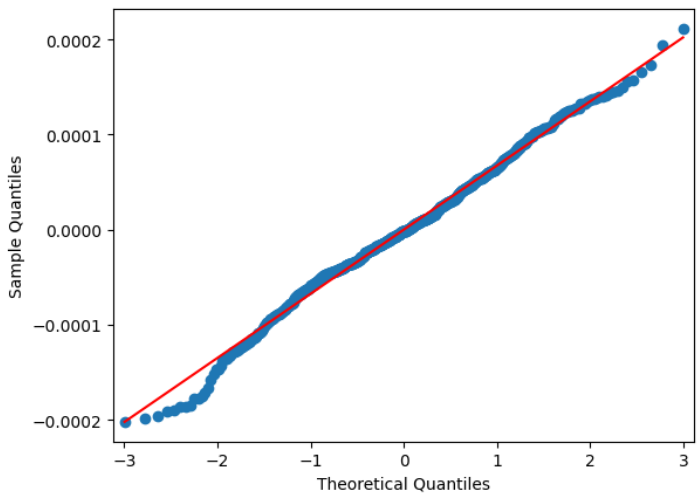

Fig.20. Normalidad de los residuales

Con esta grafica de la Fig.20. podemos ver que los residuales se comportan de manera normal pero lo vamos a corroborar con el test de Shapiro y este nos regresa un valor P = 0.004 por lo que podemos decir que sí son normales.

Ahora vamos a revisar la linealidad del modelo.

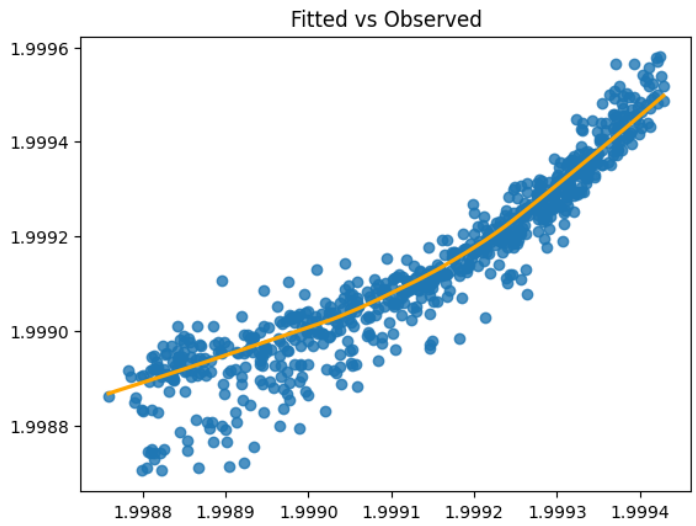

Fig.21. Linealidad del modelo con los datos ajustados

Aquí en la Fig.21 podemos ver que el modelo no se ve perjudicado por valores reales y no se ven afectados los supuestos del modelo.

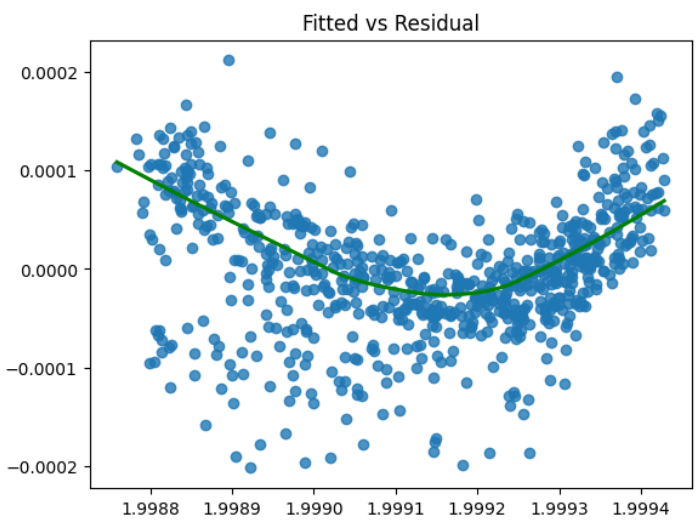

Fig.22. Valores de entrenamiento vs residuales

En esta grafica de la Fig.22. checamos la Homocedasticidad, que las varianzas se distribuyan a lo largo de la gráfica, es importante no poder identificar un patrón, que las varianzas estén distribuidas de manera homogénea. En este caso podemos ver que al inicio de la gráfica si muestran aleatoriedad, pero al final de ella se puede llegar a notar un patrón de agrupamiento, mostrando una tendencia decreciente.

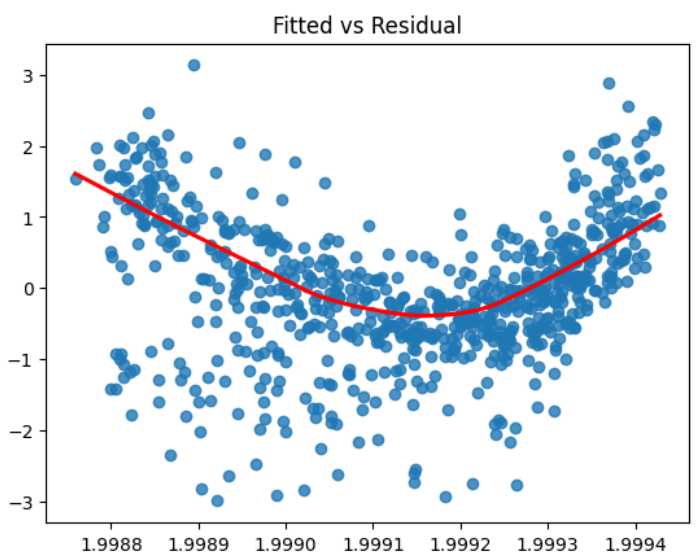

Fig.23. Valores de residuales estandarizados

En la Fig.23. revisamos la homocedasticidad del modelo, pero en este caso con los valores de los residuales estandarizados para tener resultados más certeros.

A continuación, corrimos una prueba de Breush-Pagan para ver si nuestro modelo presenta heterocedasticidad y nos dio un valor P= 5.194235193254562e-05. Lo cual indica que hay evidencia de heterocedasticidad.

Por último, hacemos las predicciones del modelo final para ver su comportamiento con el conjunto de prueba y el de validación. Lo que esperamos ver aquí en un buen rendimiento del modelo frente a datos nuevos.

```
Coeficiente de determinación
R^2 test model con backwise, sin atipícos, ni leverage points:  0.8614727905540197

Mean squared error
MSE test model con backwise, sin atipícos, ni leverage points:  4.5412109057543276e-09
```

Fig.24. Rendimiento del modelo en el conjunto de prueba

El modelo parece ser que puede explicar un 86% de la varianza de la variable dependiente con el conjunto de variables independientes dadas. Creo que es una buena actuación del modelo, pero ahora sigue verlo en acción con el conjunto de validación.

```
Coeficiente de determinación
R^2 validacion model con backwise, sin atipícos, ni leverage points:  0.8710970568159448

Mean squared error
MSE validación model con backwise, sin atipícos, ni leverage points:  5.011216010575327e-09
```

Fig.25. Rendimiento del modelo en el conjunto de validación

La fig.25. nos muestra el rendimiento del modelo frente a los datos de validación, que nunca había visto el modelo, y parece ser que es capaz de explicar el 87% de la varianza de la variable dependiente. Considero que es un buen resultado para un trabajo de este estilo, en el que cada cambio te puede afectar mucho el modelo.

Cuadernillo del Proyecto

3.1 Conclusiones

En la realización de este proyecto me di cuenta de muchas cosas, en la ciencia de datos gran parte del problema y lo que puede llegar a costar más es entender tus datos, limpiarlos e interpretarlos. Realmente correr los modelos y los análisis estadísticos ya están implementados en librerías y todo, lo importante está en saber lo que quieres lograr y la información que te va a ayudar a dar con esa explicación.

Lo que me quedo de este proyecto es el amplio conocimiento que adquirí al estar investigando y realizando prueba y error a la hora de entrenar y crear modelos, fue toda una aventura, para llegar a mi modelo final, antes tuve que crear aproximadamente unos 8 que no funcionaban tan bien, primero era porque no había realizado las transformaciones necesaria, ni había seleccionado bien las variables, pero de todo se aprende y logré llegar a modelos con mejores rendimientos peleándome contra los datos.

Algo que me hubiera gustado tener más claro a la hora de iniciar con el proyecto de ciencia de datos es si es que existe una ruta a la hora de empezar a trabajar en el modelo, me refiero a saber si primero se quitan los valores de 0 y luego se puede estandarizar o si hacer uno primero otro afecta a mi modelo.

Por ultimo agregar que trabajar con es interesante, porque como nos decía un profesor el semestre pasado:" Hay que hacer a los datos hablar" y en este trabajo creo que logré algo similar, los datos nos dijeron un poco que se necesita para lograr éxito en el mundo de la música por streaming. Como pasos a futuro me gustaría encontrar más y mejores datos para seguir reforzando este aprendizaje y poder conocer si hay factores más técnicos musicalmente hablando que hagan que una canción sea bien recibida por el público.