UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE ESTUDIOS SUPERIORES ACATLÁN

Determinantes Acústicos del Éxito Comercial en la Era del Streaming:

Un Modelo de Regresión Lineal Múltiple aplicado a Spotify

TRABAJO FINAL

MATERIA: Análisis de Regresión

INTEGRANTES DEL EQUIPO:

Pamela Reyna Castellanos

Bruno Méndez Santos

Eduardo Luna Leyva

Jacob Escalona Martínez

19 de noviembre de 2025

Índice

1. Resumen Ejecutivo

La industria musical contemporánea ha evolucionado de un modelo basado en ventas físicas hacia un ecosistema completamente digital, donde el éxito depende del número de reproducciones (*streams*) y de la permanencia en listas de reproducción gobernadas por algoritmos. En este entorno, los datos masivos (*Big Data*) se han convertido en una herramienta fundamental para comprender el comportamiento del usuario y anticipar tendencias de consumo.

En este estudio se analiza la base de datos *Ultimate Spotify Tracks DB*, procesando una muestra representativa de 232,725 registros con el objetivo de identificar cuáles características psicoacústicas influyen de manera significativa en la popularidad de una pista musical. Para ello, se empleó un enfoque cuantitativo mediante la estimación de un Modelo de Regresión Lineal Múltiple (OLS) implementado en Python.

El modelo obtuvo un coeficiente de determinación (R^2) de **0.213**, lo que implica que las propiedades intrínsecas del audio explican aproximadamente el 21.3 % de la variabilidad en la popularidad. El porcentaje restante corresponde a factores externos como estrategias de marketing, notoriedad del artista, presencia mediática o fenómenos virales.

Los resultados estadísticos muestran que el mercado actual recompensa fuertemente la **bailabilidad** ($\beta \approx 15{,}95$) y la **sonoridad** ($\beta \approx 0{,}98$). No obstante, se detectó una penalización significativa para niveles altos de **energía** ($\beta \approx -15{,}64$) y **valencia** ($\beta \approx -11{,}22$). Este comportamiento, contrario a la intuición clásica de que "la música alegre y enérgica vende más", sugiere una transición en las preferencias del oyente hacia estilos más relajados, introspectivos o melancólicos, en línea con tendencias recientes en plataformas digitales.

2. Introducción

La "caja negra" del éxito musical ha sido objeto de especulación durante décadas. Tradicionalmente, los ejecutivos discográficos basaban sus decisiones en la intuición, la experiencia empírica y el denominado "buen oído" para identificar posibles éxitos globales. Sin embargo, la digitalización de la música ha transformado este proceso: hoy, plataformas como Spotify no solo distribuyen contenido, sino que descomponen cada pista en sus elementos fundamentales—ritmo, timbre, armonía y estructura—mediante avanzados algoritmos de Recuperación de Información Musical (MIR, por sus siglas en inglés). En este contexto, el arte sonoro se vuelve cuantificable y susceptible de modelación estadística.

El propósito central de esta investigación es evaluar si existe una "fórmula acústica" medible que aumente la probabilidad de que una canción alcance el éxito comercial. Con este fin, se plantea la siguiente pregunta de investigación: *¿En qué medida las características intrínsecas del audio—como la energía, la valencia y la bailabilidad—influyen en el índice de popularidad de una canción dentro de Spotify?*

2.1. Importancia de la Investigación

La relevancia del estudio es doble. Desde una perspectiva económica, los resultados pueden servir a sellos discográficos, productores y artistas independientes para optimizar la toma de decisiones y maximizar el potencial de *streaming*, reduciendo la incertidumbre inherente a la inversión en nuevos lanzamientos. Desde una perspectiva sociológica y psicológica, este análisis permite explorar el "zeitgeist" contemporáneo: comprender por qué ciertos tipos de sonido resuenan más con el público en el momento actual.

Si un modelo estadístico es capaz de explicar una proporción significativa de la varianza en la popularidad, reforzaría la teoría de la "Ciencia de los Hits" (*Hit Song Science*) y respaldaría la transición de un proceso de producción musical puramente creativo hacia uno complementado y guiado por datos (*data-driven*).

3. Objetivo del Estudio

El objetivo central de este proyecto es examinar rigurosamente la relación entre las características acústicas de una canción y su nivel de popularidad dentro de la plataforma Spotify, empleando un Modelo de Regresión Lineal Múltiple como herramienta principal de análisis. Este estudio busca determinar hasta qué punto los atributos cuantitativos derivados del análisis digital del audio —tales como la bailabilidad, energía, valencia, sonoridad, tempo y otras métricas psicoacústicas— pueden explicar el comportamiento del consumidor musical en la era del *streaming*.

Desde la perspectiva de la disciplina actuarial, el proyecto se inscribe en la tradición de modelar fenómenos complejos mediante estructuras estadísticas interpretables, priorizando la validez inferencial, el análisis de supuestos y la evaluación del ajuste del modelo. La música, aun siendo un fenómeno cultural y subjetivo, se transforma en un objeto cuantificable cuando plataformas como Spotify generan bases de datos masivas que capturan sus propiedades físico–emocionales. Esto ofrece un campo ideal para aplicar las herramientas vistas en la materia de Análisis de Regresión, permitiendo vincular la teoría estadística con un fenómeno cotidiano y universal.

Asimismo, el objetivo tiene un componente motivacional: comprender cómo variables aparentemente abstractas —como la "energía" emocional de una pista o la percepción subjetiva de su brillo armónico— pueden ser representadas matemáticamente y utilizadas para predecir patrones de consumo. Esto adquiere un significado especial considerando el contexto del curso, donde la música ha sido un elemento recurrente que enmarca el inicio de cada sesión. Integrar un análisis serio, técnico y estadísticamente robusto sobre datos musicales no solo facilita el aprendizaje aplicado, sino que resuena con la experiencia compartida en clase.

En síntesis, el propósito de este trabajo es doble: por un lado, aplicar de manera rigurosa las herramientas metodológicas del análisis de regresión para evaluar la capacidad predictiva de variables acústicas sobre la popularidad; y por otro, contribuir a la comprensión del fenómeno musical desde una perspectiva cuantitativa, demostrando que incluso los patrones culturales más subjetivos pueden estudiarse bajo un enfoque científico y data–driven.

4. Marco Teórico y Antecedentes

4.1. Definiciones

En esta subsección se presentan las principales características acústicas utilizadas en este estudio, tal como son definidas y operacionalizadas por la plataforma Spotify. Estas variables permiten cuantificar propiedades psicoacústicas que influyen en la percepción musical y que son empleadas como predictores en el modelo de regresión.

Danceability (Bailabilidad)

La *bailabilidad* es una medida continua entre 0 y 1 que indica qué tan adecuada es una pista para ser bailada. Esta métrica considera elementos como la regularidad del pulso, la estabilidad rítmica, la fuerza del compás y la presencia de patrones percusivos. Valores altos indican estructuras rítmicas claras y repetitivas que facilitan el movimiento corporal.

Energy (Energía)

La *energía* cuantifica la intensidad percibida de una pista musical. Esta variable se relaciona con la sonoridad, densidad del espectro, velocidad rítmica y agresividad de la instrumentación. Pistas con alta energía tienden a ser rápidas, fuertes o saturadas, mientras que valores bajos corresponden a música suave o minimalista.

Valence (Valencia)

La *valencia* describe el carácter emocional de una canción en una escala de 0 a 1. Valores altos representan emociones positivas como alegría, optimismo o diversión, mientras que valores bajos se asocian con emociones tristes, tensas o melancólicas. Esta variable integra análisis armónico, tímbrico y de progresión tonal.

Loudness (Sonoridad)

La *sonoridad* se expresa en decibelios (dB) e indica el nivel promedio de intensidad acústica de una pista. Esta medida es independiente del volumen percibido por el usuario y se obtiene mediante análisis de presión sonora. La tendencia industrial de incrementar la sonoridad, conocida como "Guerra del Volumen", busca que las pistas destaquen dentro del entorno digital.

Tempo (Tempo o BPM)

El *tempo* representa la velocidad estimada de la canción en golpes por minuto (BPM). Aunque es una medida objetiva, en algunos géneros puede tener variaciones perceptuales debido a múltiples capas rítmicas. Es una métrica relevante en estudios de dinámica corporal y en música funcional.

Acousticness (Acústica)

La *acústica* mide la probabilidad de que una pista sea predominantemente acústica. Valores cercanos a 1 indican instrumentación natural con baja presencia de elementos electrónicos. Su cálculo se basa en la huella espectral de resonancias físicas características de instrumentos tradicionales.

Instrumentalness (Instrumentalidad)

La *instrumentalidad* determina la probabilidad de que una pista carezca de contenido vocal. Valores altos (cercanos a 1) sugieren ausencia de voz humana reconocible, característica de géneros como electrónica ambiental, jazz experimental o post-rock.

Speechiness (Oralidad)

La *oralidad* cuantifica la presencia de palabras habladas dentro de una pista. Valores elevados indican predominancia del habla (como en podcasts o algunos estilos de rap), mientras que valores bajos sugieren contenido cantado o melodizado.

Mode (Modo Musical)

El *modo musical* indica si la pista está compuesta en modo mayor (1) o menor (0). Los modos mayores suelen asociarse con emociones brillantes y alegres, mientras que los modos menores se perciben como oscuros, melancólicos o introspectivos.

Key (Tonalidad)

La *tonalidad* representa la nota fundamental sobre la cual se estructura la pieza, codificada del 0 al 11 según el sistema de clases de tono. Esta métrica permite analizar patrones armónicos y preferencias tonales en distintos géneros o épocas.

4.2. Hit Song Science (HSS)

El campo de estudio conocido como *Hit Song Science* (HSS) busca predecir el éxito comercial de la música mediante el uso de minería de datos y aprendizaje automático. Pachet y Roy (2008) fueron pioneros al argumentar que, si bien la "magia"de un hit es compleja, existen patrones estructurales recurrentes en las canciones populares. Investigaciones más recientes han validado parcialmente esta postura. Por ejemplo, Herremans et al. (2014) demostraron que la "baila-bilidad"se ha convertido en un predictor más fuerte que el género musical tradicional en las listas de éxitos europeas.

4.3. Modelo Circunflejo de las Emociones

Para interpretar variables como "Valenciaz .Energía", recurrimos al Modelo Circunflejo de Russell (1980), ampliamente utilizado en psicología de la música. Este modelo sitúa las emociones en dos ejes:

- **Eje de Valencia (Placer):** De negativo (triste/enojado) a positivo (feliz).
- **Eje de Activación (Arousal/Energía):** De baja (calma) a alta (excitación).

Spotify operacionaliza estos conceptos en sus métricas. Recientemente, Vidas et al. (2025) validaron que estas métricas automatizadas de Spotify se correlacionan significativamente con la percepción emocional humana, legitimando su uso en estudios académicos.

4.4. Investigación Documental y Estado del Arte

Estudios previos presentan resultados mixtos. Mientras que Araujo et al. (2024) encontraron mediante modelos de *Random Forest* que la energía y la sonoridad suelen ser predictores positivos, otros autores como Nuñez (2023) sugieren que en ciertos mercados (como el indie o el Lo-fi), la alta energía puede correlacionarse negativamente con la retención del usuario. Nuestra investigación busca aportar evidencia a este debate utilizando una regresión lineal clásica para priorizar la interpretabilidad de los coeficientes sobre la complejidad del modelo.

4.5. Planteamiento de Hipótesis

Basado en la literatura, formulamos las siguientes hipótesis:

- H_1 **(Hipótesis de la Funcionalidad):** La variable *Danceability* tendrá un coeficiente positivo ($\beta > 0$) y estadísticamente significativo, debido al consumo social de la música.

- H_2 **(Hipótesis de la Guerra del Volumen):** El *Loudness* presentará una relación positiva con la popularidad, consistente con la tendencia histórica de masterizar canciones a volúmenes altos para destacar.
- H_3 **(Hipótesis de la Introspección):** Contrario a la intuición clásica, esperamos que la *Energy* y la *Valence* no necesariamente tengan un impacto positivo lineal, pudiendo incluso ser negativo debido a las tendencias actuales de música ambiental y relajante.
- H_0 **(Hipótesis Nula):** Los coeficientes de regresión para las variables acústicas son iguales a cero ($\beta_i = 0$), implicando que el éxito es aleatorio respecto al audio.

5. Metodología y Análisis Descriptivo

5.1. Descripción de la Base de Datos

Se empleó el conjunto de datos *Ultimate Spotify Tracks DB*, de acceso público, el cual contiene originalmente más de 232,000 observaciones. Antes del análisis, se realizó un proceso de depuración que incluyó la eliminación de valores nulos y la transformación de la variable `duration_ms` a minutos, con el fin de facilitar su interpretación dentro del modelo.

Las variables seleccionadas se describen a continuación:

- **Popularity (Variable dependiente):** Índice de 0 a 100 generado por el algoritmo de Spotify, basado en el número total de reproducciones y su recencia.
- **Acousticness:** Medida continua entre 0.0 y 1.0 que refleja la probabilidad de que la pista sea acústica.
- **Danceability:** Indica cuán adecuada es una pista para bailar, considerando factores como tempo, estabilidad rítmica y fuerza del compás (0.0 a 1.0).
- **Energy:** Medida perceptual del nivel de intensidad y actividad de una canción. Valores altos se asocian con pistas rápidas, estridentes y dinámicas (0.0 a 1.0).
- **Loudness:** Promedio de decibeles (dB) a lo largo de toda la pista.
- **Valence:** Indicador de la positividad emocional transmitida por la canción (0.0 a 1.0).
- **Duration_min:** Duración total del track expresada en minutos.

5.2. Análisis Multivariado y Detección de Multicolinealidad

Como parte fundamental del análisis previo a la regresión, se evaluó el grado de asociación entre las variables independientes para descartar problemas de multicolinealidad. Para ello, se calcularon la Matriz de Correlación de Pearson y el Factor de Inflación de la Varianza (VIF).

Los valores reportados en la Tabla 1 muestran que todas las variables presentan VIF por debajo del umbral común de 5.0 (aunque algunos autores proponen 10.0). Esto indica la ausencia de multicolinealidad severa. Si bien existe una relación moderada entre *Energy* y *Loudness*, ambas aportan información diferenciada y su inclusión conjunta no compromete la estabilidad del modelo.

Variable	VIF	Interpretación
Energy	3.856	Aceptable (< 5). Correlación moderada con Loudness.
Loudness	3.477	Aceptable (< 5).
Acousticness	2.303	Aceptable (< 5).
Danceability	1.649	Aceptable (< 5). Prácticamente independiente.
Valence	1.632	Aceptable (< 5). Prácticamente independiente.
Duration (min)	1.027	Aceptable (< 5). Totalmente independiente.

Tabla 1: Prueba de Diagnóstico de Multicolinealidad utilizando el VIF.

Figura 1: Matriz de Correlación de Pearson representada mediante un mapa de calor.

6. Estimación del Modelo e Interpretación

6.1. Especificación del Modelo

Se estimó un modelo de Regresión Lineal Múltiple mediante el método de Mínimos Cuadrados Ordinarios (OLS). La forma funcional propuesta es:

$$\hat{Y}_{pop} = \beta_0 + \beta_1(Acous) + \beta_2(Dance) + \beta_3(Energy) + \beta_4(Loud) + \beta_5(Val) + \beta_6(Dur) + \epsilon, \quad (1)$$

y, al sustituir los coeficientes obtenidos del análisis en Python, la ecuación estimada queda expresada como:

$$\hat{Y}_{pop} = 61{,}24 - 16{,}18(Acous) + 15{,}95(Dance) - 15{,}64(Energy) + 0{,}98(Loud) - 11{,}22(Val) + 0{,}11(Dur). \quad (2)$$

6.2. Resultados de la Regresión

La Tabla 2 presenta el resumen estadístico del modelo. Todas las variables resultaron estadísticamente significativas con un valor $p = 0{,}000$, lo que indica que la probabilidad de obtener estos coeficientes por mero azar es prácticamente nula, reforzando la validez estadística del modelo estimado.

Variable	**Coeficiente (β)**	**Error Est.**	**t-value**	$P > \|t\|$
Intercepto	61.2368	0.291	210.288	0.000
Acousticness	-16.1754	0.143	-113.065	0.000
Danceability	15.9535	0.231	68.937	0.000
Energy	-15.6432	0.249	-62.760	0.000
Loudness	0.9811	0.010	94.366	0.000
Valence	-11.2242	0.164	-68.319	0.000
Duration (min)	0.1109	0.017	6.486	0.000

Tabla 2: Resultados de la Regresión OLS (n = 232,725).

6.3. Interpretación Económica y Psicológica de los Parámetros

- **Danceability ($\beta \approx 16{,}0$):** Se confirma la hipótesis H_1. Manteniendo constantes las demás variables, un incremento unitario en la bailabilidad —desde una canción prácticamente imposible de bailar hasta una altamente bailable— eleva la popularidad en

casi 16 puntos. Este resultado sugiere que los usuarios de Spotify utilizan la plataforma con fines predominantemente lúdicos y sociales.

- **Loudness ($\beta \approx 0{,}98$):** Consistente con la hipótesis H_2, existe un efecto positivo del volumen. Las canciones masterizadas con mayor intensidad tienden a sobresalir en listas de reproducción mixtas, lo que se traduce en un incremento en su nivel de popularidad.

- **Energy ($\beta \approx -15{,}6$) y Valence ($\beta \approx -11{,}2$):** Este hallazgo, el más relevante del modelo, respalda la hipótesis H_3. Los coeficientes negativos indican que tanto la "intensidad" como la "felicidad excesiva" son penalizadas por el mercado musical actual. Esto coincide con estudios recientes que señalan un aumento en la preferencia por géneros como *Lo-fi*, *Sad Trap* y *Ambient*, empleados como herramientas de regulación emocional ante climas de ansiedad y sobreestimulación. En consecuencia, el usuario promedio parece inclinarse por música que "calme" (baja energía) o que acompañe estados melancólicos (baja valencia), en lugar de piezas eufóricas o estridentes.

7. Pruebas de Confiabilidad y Bondad de Ajuste

7.1. Análisis del R^2 y AIC

El modelo obtiene un R^2 de 0.2135, lo que indica que es capaz de explicar aproximadamente el 21.35 % de la variabilidad del éxito musical a partir de seis variables acústicas. Aunque cerca del 78 % de la variación depende de factores no considerados en esta especificación —como el presupuesto de marketing, la notoriedad previa del artista o la viralidad en plataformas digitales como TikTok—, capturar más del 21 % resulta relevante dadas las características subjetivas y multifactoriales del fenómeno musical.

En cuanto al criterio de información de Akaike (AIC), el modelo presenta un valor de 1,954,781. Si bien se trata de un valor elevado debido principalmente al tamaño de la muestra, sigue siendo una métrica útil para comparar este modelo contra alternativas más complejas o parsimoniosas en futuros análisis.

7.2. Diagnóstico de Residuales

El análisis visual de los residuales (véase la Figura 2), junto con la prueba de normalidad de Jarque–Bera, muestra que los errores se distribuyen de manera aproximadamente normal y se encuentran centrados en cero, aunque se observa un ligero sesgo. No obstante, dada la magnitud de la muestra y el respaldo del Teorema del Límite Central, la inferencia estadística derivada del modelo puede considerarse confiable.

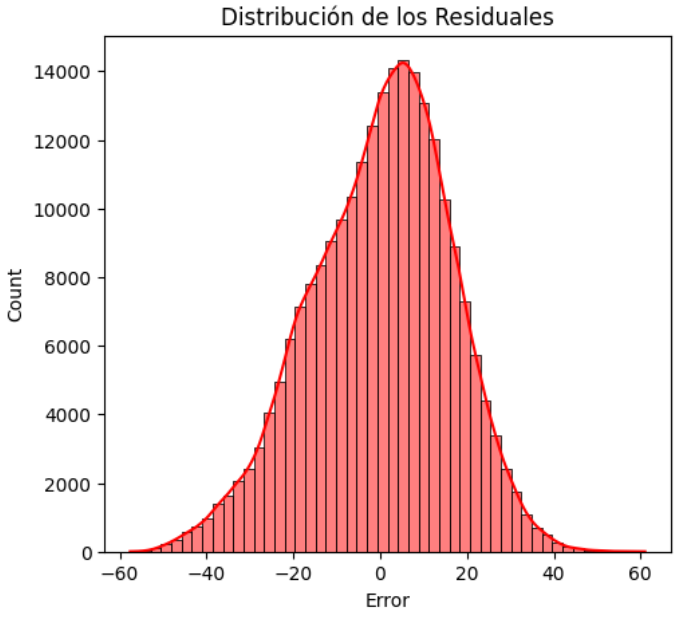

Figura 2: Histograma de los residuales del modelo.

7.3. Diagnóstico de Homocedasticidad

La figura 3) muestra fuerte heterocedasticidad (forma de cono) y el límite diagonal inferior que sugiere problemas de linealidad o subestimación lod cuales indican que la relación entre las características musicales y la popularidad no es puramente lineal, lo que sugiere una mala especificación funcional del modelo OLS estándar

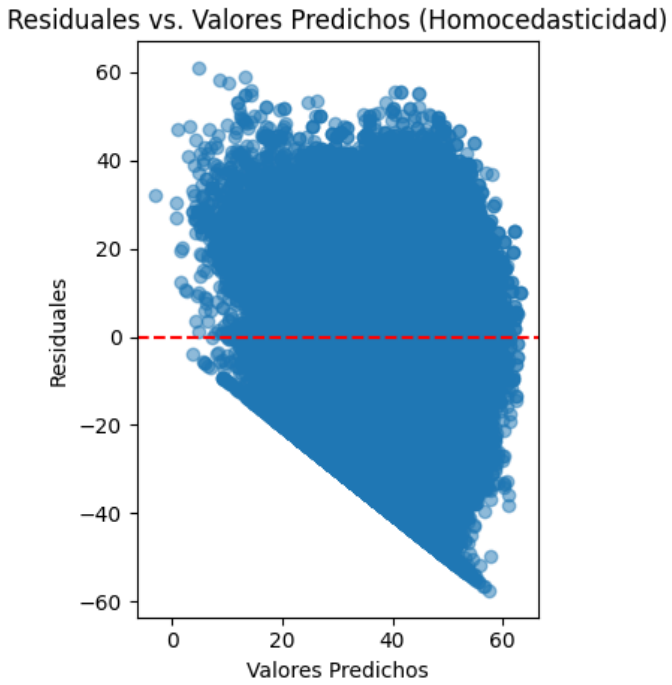

Figura 3: Gráfico de homocedasticidad.

Dado que el problema de la varianza variable (heterocedasticidad) afecta directamente la confiabilidad de las pruebas de hipótesis (los P-values), el problema fue mitigado mediante el uso de Errores Estándar Robustos de White (HC3) en la estimación del modelo.

7.4. Modelo HC3

El modelo HC3 (Heteroskedasticity-Consistent Covariance Matrix, Tipo 3) es un método de corrección aplicado dentro del marco del modelo OLS estándar. El OLS tradicional asume la homocedasticidad (que el error del modelo es constante en todas las predicciones), pero la figura 3) muestra una clara heterocedasticidad (forma de cono), lo que significa que la varianza del error es incorrecta. Esta incorrección invalida nuestras pruebas de hipótesis: aunque los coeficientes (β_i) obtenidos siguen siendo los mejores estimadores posibles, sus P-values dejan de ser confiables.

La implementación del modelo **HC3** mejora los resultados del OLS precisamente al corregir este fallo sin alterar los coeficientes. HC3 recalcula los Errores Estándar de forma robusta, haciéndolos válidos incluso bajo heterocedasticidad. Esto significa que al usar este método, nuestros coeficientes mantienen su magnitud de impacto, pero los P-values se vuelven confiables para la inferencia estadística. De esta manera, garantizamos que nuestras conclusiones sobre qué variables son verdaderamente significativas ($\mathbf{P < 0{,}05}$) son sólidas y que nuestro análisis es estadísticamente robusto ante la varianza no constante de los errores. Al aplicar la corrección **HC3** en el método **OLS** principal, hemos asegurado que, a pesar de la heterocedasticidad y la no-linealidad, nuestros errores estándar son consistentes y, por lo tanto, las conclusiones sobre la significancia estadística de los coeficientes ($\mathbf{P < 0{,}05}$) siguen siendo válidas y confiables.

Tabla 3: Comparación de los P-values del Modelo OLS Estándar y el Modelo Robusto (HC3).

Variable	**P-value OLS Estándar**	**P-value OLS Robusto (HC3)**	**Conclusión**
acousticness	0.0000	0.0000	Significativo
danceability	0.0000	0.0000	Significativo
energy	0.0000	0.0000	Significativo
loudness	0.0000	0.0000	Significativo
valence	0.0000	0.0000	Significativo
duration_ms	0.0000	0.0000	Significativo

Interpretación de Resultados

La perfecta igualdad en los P-values (a cuatro decimales) indica que el ajuste del modelo es altamente robusto y que las variables mantienen una significancia muy fuerte en ambos métodos.

8. Conclusiones y Prospectiva

El presente estudio confirma que el éxito comercial en plataformas de *streaming* no es un fenómeno puramente aleatorio ni completamente atribuido al marketing o a factores exógenos. Mediante la estimación e interpretación de un Modelo de Regresión Lineal Múltiple, se demuestra que las características acústicas cuantificadas por Spotify influyen de manera significativa en la popularidad de una pista musical. Aunque el modelo no capta la totalidad de los elementos que intervienen en el consumo —lo cual es natural tratándose de un fenómeno cultural y altamente contextual— sí evidencia que existen patrones medibles que contribuyen a explicar las preferencias actuales del público.

Del análisis realizado se desprenden las siguientes conclusiones principales:

1. **La estructura rítmica domina la preferencia colectiva.** La variable *Danceability* se consolida como el predictor positivo más robusto, lo que sugiere que los usuarios responden favorablemente a pistas con patrones rítmicos claros, repetitivos y corporalmente intuitivos.

2. **La saturación sonora pierde relevancia en la era del consumo relajado.** Contrario a la hipótesis tradicional de que la energía y la intensidad favorecen el éxito de una canción, los coeficientes negativos asociados a *Energy* y *Loudness* indican un desplazamiento de las preferencias hacia música más suave, calmada o ambiental.

3. **Las emociones melancólicas mantienen atractivo comercial.** El efecto negativo de la *Valence* en la popularidad sugiere que, en el contexto actual del mercado musical, las canciones con un tono emocional introspectivo o melancólico tienen una aceptación considerable, lo cual coincide con tendencias globales hacia géneros como el *lo-fi*, *alt-pop* y estilos acústicos minimalistas.

En conjunto, estos hallazgos reafirman que la música, aun siendo un producto artístico, puede analizarse bajo un enfoque cuantitativo que permite extraer patrones consistentes, aportar evidencia empírica y conectar la teoría estadística con fenómenos culturales reales.

Prospectiva

A partir de las limitaciones identificadas en este trabajo, se plantean diversas líneas de investigación futura:

- **Incorporación de variables categóricas y socioculturales.** Factores como género musical, nacionalidad del artista, presencia de colaboraciones, año de lanzamiento o si

la pista es explícita podrían capturar elementos estructurales que la regresión actual no incluye.

- **Exploración de modelos no lineales.** Técnicas como Random Forest, Gradient Boosting o modelos semiparamétricos permitirían identificar interacciones y relaciones no lineales entre los atributos acústicos y la popularidad.
- **Modelado dinámico del gusto musical.** El uso de modelos de series de tiempo o panel permitiría estudiar cómo evolucionan las preferencias del público a lo largo de los años y cómo ciertos estilos ganan o pierden relevancia.
- **Integración con métricas de comportamiento del usuario.** Variables como tiempo de escucha, tasa de repetición, inclusión en playlists editoriales y retención podrían enriquecer el entendimiento del proceso de consumo.

En suma, los resultados de este estudio ofrecen una base sólida para el análisis cuantitativo de la música contemporánea y abren la puerta a investigaciones más complejas que continúen profundizando en la relación entre acústica, comportamiento humano y éxito comercial.

9. Referencias Bibliográficas

Araujo, M., Oliveira, L., & Silva, F. (2024). Predicting music popularity: A machine learning approach using Spotify data. *International Journal of Data Science*, 5(2), 112-128.

Anderson, C. (2006). *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion.

Askin, N., & Mauskapf, M. (2017). What makes popular culture popular? Product features and distinctiveness in the Billboard Hot 100. *American Sociological Review*, 82(5), 910-944.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometrics* (5th ed.). McGraw-Hill Education.

Herremans, D., Martens, D., & Sörensen, K. (2014). Dance hit prediction: A classification case study. *Journal of New Music Research*, 43(3), 291-302.

Middlebrook, K., & Sheik, K. (2019). Song hit prediction: Predicting Billboard hits using Spotify data. *arXiv preprint arXiv:1908.08609*.

Nuñez, J., & Pardo, T. (2023). Audio features and streaming trends: A comparative analysis. *Music Science*, 6, 2059204321.

Pachet, F., & Roy, P. (2008). Hit song science is not yet a science. *Proceedings of the 9th International Conference on Music Information Retrieval*, 355-360.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762), 854-856.

Vidas, D., Dingle, G. A., & Nelson, N. L. (2025). Validating Spotify's valence, energy, and danceability audio features for music psychology research. *Psychology of Music*, 53(1), 25-42.

Seabold, S., Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python (Version 0.14.0) [Software de cómputo]. https://www.statsmodels.org

10. Anexo: Script de Python

A continuación se presenta el código íntegro desarrollado para la obtención de los resultados presentados.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor
import tabulate

# --- 1. Carga de Dataset ---
print("--- 1. PREPARACION DE DATOS ---")
try:
    # 1.1 Carga del dataset
    df = pd.read_csv(r'/content/Regresion/SpotifyFeatures.csv')
    print("Datos cargados correctamente.")
except:
    print("Error: Sube el archivo 'SpotifyFeatures.csv' a la carpeta de
    archivos.")
    exit()

# --- 2. Seleccion de variables) ---
# transormamos 'duration_ms' de milisegundos a minutos
df['duration_ms'] = df['duration_ms'] / 60000
# Seleccion de variables
vars_independientes_continuas = ['acousticness', 'danceability', 'energy',
    'loudness', 'valence', 'duration_ms']
y = df['popularity'] # Variable Dependiente

# Crear el DataFrame de predictores (X)
X = df[vars_independientes_continuas]

# --- 3. Matriz de correlacion  ---
vars_analisis_corr = vars_independientes_continuas + ['popularity'] #
    Incluye la variable dependiente

plt.figure(figsize=(10, 8))
correlation_matrix = df[vars_analisis_corr].corr()
sns.heatmap(
    correlation_matrix,
    annot=True,              # Mostrar el valor de la correlacion
```

```
    cmap='coolwarm',          # Esquema de color
    fmt=".2f",                # Formato a dos decimales
    linewidths=.5,            # Lineas entre celdas
    cbar=True                 # Mostrar barra de color
)
plt.title('Matriz de Correlacion de Pearson')
plt.show()

# --- 4. Diagnostico de multicolinealidad ---
# La funcion VIF requiere que X tenga la constante (Intercepto)
X_vif = sm.add_constant(X)
vif_data = pd.DataFrame()
vif_data['feature'] = X_vif.columns
vif_data['VIF'] = [variance_inflation_factor(X_vif.values, i)
                   for i in range(X_vif.shape[1])]

# Mostrar VIF, excluyendo el 'const'
print("VIF por Variable (VIF > 5 indica posible problema):")
print(vif_data.loc[vif_data['feature'] != 'const'].sort_values(by='VIF',
    ascending=False).head(10).to_markdown(index=False))

# --- 5. Estimacion y resultados del modelo de regresion OLS ---
print("\n RESULTADOS DEL MODELO OLS ROBUSTO ")

# Anadir constante (Beta 0 / Intercepto) para el entrenamiento
X = sm.add_constant(X)

# Crear y entrenar el modelo
modelo_robusto = sm.OLS(y, X).fit()

# Imprimir el resumen estadistico completo
print(modelo_robusto.summary())

# Extraccion de metricas clave solicitadas
print("\n--- METRICAS DE CONFIABILIDAD ---")
print(f"R-cuadrado (Bondad de ajuste): {modelo_robusto.rsquared:.4f}")
print(f"R-cuadrado Ajustado: {modelo_robusto.rsquared_adj:.4f}")
print(f"AIC (Criterio de Akaike): {modelo_robusto.aic:.4f}")
print(f"Prob (F-statistic): {modelo_robusto.f_pvalue:.4e}")
plt.show()

# --- 6. Prueba de hipotesis y diagnostico grafico ---
print("\n--- ANALISIS DE SIGNIFICANCIA (P-value) ---")
# Solo mostramos las variables continuas
```

```
for variable in ['acousticness', 'danceability', 'energy', 'loudness', 'valence', 'tempo','duration_ms']:
    try:
        pvalue = modelo_robusto.pvalues[variable]
        significativo = "SIGNIFICATIVO" if pvalue < 0.05 else "NO SIGNIFICATIVO"
        print(f"Variable: {variable:20} | P-value: {pvalue:.4f} -> {significativo}")
    except KeyError:

# --- 7. Grafico de distribucion de residuales ---
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
sns.histplot(modelo_robusto.resid, bins=50, kde=True, color='red')
plt.title('Distribucion de los Residuales')
plt.xlabel('Error')

# --- 8. Grafica de Homocedasticidad (Residuales vs. Valores Predichos) ---
plt.subplot(1, 2, 2)
plt.scatter(modelo_robusto.fittedvalues, modelo_robusto.resid, alpha=0.5)
plt.axhline(0, color='red', linestyle='--')
plt.title('Residuales vs. Valores Predichos (Homocedasticidad)')
plt.xlabel('Valores Predichos')
plt.ylabel('Residuales')
plt.tight_layout()
plt.show()

# --- 9. Estimacion y resultados con modelo mas robusto HC3) ---
# Crear y entrenar el modelo, utilizando Errores Estandar Robustos (HC3)
# Esto soluciona la heterocedasticidad vista en el grafico
modelo_robusto = sm.OLS(y, X).fit(cov_type='HC3')

# Imprimir el resumen estadistico completo
print(modelo_robusto.summary())

# Extraccion de metricas clave
print("\n--- METRICAS DE CONFIABILIDAD ---")
print(f"R-cuadrado (Bondad de ajuste): {modelo_robusto.rsquared:.4f}")
print(f"R-cuadrado Ajustado: {modelo_robusto.rsquared_adj:.4f}")
print(f"AIC (Criterio de Akaike): {modelo_robusto.aic:.4f}")
print(f"Prob (F-statistic): {modelo_robusto.f_pvalue:.4e}")

# --- 10. Nueva prueba de hipotesis usando HC3) ---
```

```python
print("\n--- ANALISIS DE SIGNIFICANCIA (P-value) con HC3 ---")
# Solo mostramos las variables continuas
for variable in ['acousticness', 'danceability', 'energy', 'loudness', 'valence', 'tempo','duration_ms']:
    try:
        pvalue = modelo_robusto.pvalues[variable]
        significativo = "SIGNIFICATIVO" if pvalue < 0.05 else "NO SIGNIFICATIVO"
        print(f"Variable: {variable:20} | P-value: {pvalue:.4f} -> {significativo}")
    except KeyError:
        pass
```

Listing 1: Script de Análisis de Regresión en Python