Análisis de tendencias y popularidad de los podcasts

Diana Cordero

Noviembre 2024

1. Introducción

En la era digital, los pódcasts se han consolidado como una de las plataformas más dinámicas para la difusión de contenido, atrayendo audiencias globales con intereses diversos.

Según Román y Solano (2012, citado en Figueroa 2019), un pódcast se define como ün archivo o una serie de archivos de audio o video digital, previamente grabados, que pueden ser distribuidos a través de internet y descargados automáticamente en un dispositivo portátil".

La creciente popularidad de esta forma de comunicación plantea preguntas sobre los factores que influyen en su éxito y las estrategias necesarias para maximizar su alcance.

Este artículo aborda dos áreas clave relacionadas con el análisis de los pódcasts: las tendencias de popularidad y la optimización para el marketing.

En primer lugar, analizamos cómo varían los rankings de los pódcasts y episodios en diferentes regiones y a lo largo del tiempo, identificando patrones que revelen su estabilidad en los primeros puestos o movimientos rápidos que sugieran tendencias emergentes. Este análisis nos permite comprender la dinámica de las preferencias de la audiencia en distintos contextos geográficos y temporales.

En segundo lugar, examinamos el impacto del tipo de contenido y el formato en la popularidad de los pódcasts, comparando las ventajas de alojar los programas internamente frente al uso de plataformas externas. Además, identificamos a los productores o editores más exitosos que dominan los rankings, proporcionando valiosos insights para quienes buscan optimizar sus estrategias de producción y promoción.

A través de este análisis, se busca no solo entender las fuerzas que impulsan el éxito de los pódcasts, sino también ofrecer herramientas prácticas para que creadores y productores capitalicen las oportunidades de este medio en constante evolución.

2. Antecedentes

En la última década, los pódcasts han revolucionado la manera en que las personas consumen contenido digital.

En 2023, Sanz Julián realiza un estudio de caso analizando varios pódcast de entretenimiento populares en España, con la finalidad de encontrar las claves de éxito de los mismos. En este estudio, Sanz destaca la importancia de los siguientes aspectos:

- Identificación con los contenidos. Sanz indica que los oyentes se sienten parte del pódcast al encontrar un contenido que se alinee con sus intereses y valores.
- Estilo conversacional. En los pódcast analizados, los presentadores adoptan un estilo conversacional y cercano, generando una conexión emocional con la audiencia. Este enfoque íntimo logra que los oyentes se sientan identificados y parte activa del público.
- Duración y accesibilidad. En sus hpótesis, Sanz plantea que la audiencia busca contenidos que permitan descargar, compartir y escuchar donde y cuando quieran, pero con una duracipon concreta de pocos minutos. Luego del estudio, concluye que la duración pudiera no ser un factor determinante, ya que la calidad del contenido y capacidad de los presentadores para mantener una dinámica y ritmo permiten a los oyentes desconectar y sumergirse en una experiencia auditiva gratificante.
- Presencia en redes sociales. Durante el análisis, se observó la relevancia de mantener una presencia activa y dinámica, no solo desde la cuenta oficial del programa, sino también desde las cuentas personales de los presentadores.

3. Conjunto de datos

El conjunto de datos utilizado en este artículo está compuesto por información detallada sobre el top 200 de pódcast en Spotify en 22 regiones alrededor del mundo.

Este conjunto de datos se actualiza diariamente. Los datos se obtuvieron a través de la plataforma kaggle.com. A continuación, se presenta el enlace a este registro de datos: Top Spotify Podcast Episodes

El conjunto de datos consta de 149,600 renglones. Las variables contenidas en este conjunto de datos son las siguientes:

- date: Fecha de registro o actualización del ranking del pódcast.
- rank: La posición del pódcast en el ranking general de popularidad.
- region: La región geográfica en la que se registra el pódcast, lo que permite analizar las diferencias en la popularidad por áreas.
- chartRankMove: El cambio en la posición del pódcast o episodio en comparación con el periodo anterior, lo que ayuda a identificar tendencias y movimientos en el ranking.
- episodeUri: Identificador único del episodio, que permite acceder al archivo de audio o video directamente.
- showUri: Identificador único del programa o pódcast, facilitando la identificación de la serie completa.
- episodeName: Nombre del episodio, proporcionando información sobre el contenido específico del pódcast.
- description: Descripción del episodio.
- show.name: Nombre del pódcast o programa completo.
- show.description: Descripción general del pódcast, resumiendo su temática o enfoque.
- show.publisher: El editor o productor del pódcast, indicando quién gestiona la producción del contenido.
- duration.ms: Duración del episodio en milisegundos, ofreciendo información sobre la longitud del contenido.
- explicit: Indicador de si el episodio contiene contenido explícito, lo que puede influir en la audiencia.
- is.externally.hosted: Indica si el pódcast se aloja en una plataforma externa o está autoalojado.
- is.playable: Indica si el episodio es accesible y puede ser reproducido en la plataforma.
- language: El idioma principal del pódcast o episodio, lo que permite segmentar el contenido por grupos lingüísticos.
- languages: Lista de idiomas disponibles para el episodio o pódcast, en caso de que se ofrezca en varios idiomas.
- release.date: Fecha de lanzamiento del episodio, fundamental para el análisis temporal del contenido.
- show.copyrights: Información sobre los derechos de autor del pódcast.
- show.explicit: Indica si el pódcast en general contiene contenido explícito.
- show.href: Enlace a la página web del pódcast o episodio, facilitando el acceso directo.
- show.html.description: Descripción en formato HTML del pódcast o episodio.
- show.is.externally.hosted: Indica si el pódcast está alojado fuera de la plataforma principal.
- show.languages: Lista de los idiomas en los que se ofrece el pódcast.

- show.media.type: Tipo de medio en que se presenta el contenido (audio, video, etc.).
- show.total.episodes: Número total de episodios disponibles en el pódcast.
- show.type: Tipo de pódcast (por ejemplo, educativo, entretenimiento, noticias).
- show.uri: URI único del pódcast completo.

Los datos de este conjunto contienen información de las siguientes 22 regiones:

- Argentina (.ªr")
- Australia (.ªu")
- Austria (.ªt")
- Brazil ("br")
- Canadá (ça")
- Chile (çl")
- Colombia (ço")
- Francia ("fr")
- Alemania ("de")
- India (ïn")
- Indonesia (ïd")
- Irlanda (ïe")
- Italia (ït")
- Japón ("jp")
- México ("mx")
- Nueva Zelanda ("nz")
- Filipinas ("ph")
- Polonia ("pl")
- España (.ᵉˢ")
- Países Bajis ("nl")
- Reino Unido ("gb")
- Estados Unidos (üs")

4. Análisis inicial

A continuación, se muestran algunos datos relevantes del conjunto

Cuadro 1: Duración en minutos

Duración promedio	56.34
Mínima	0.1149
Máxima	705.81

En la figura 1 se observa el histograma de la duración de los pódcast.
En la figura 2, se observa la relación de pódcast de contenido explícito vs no explícito.

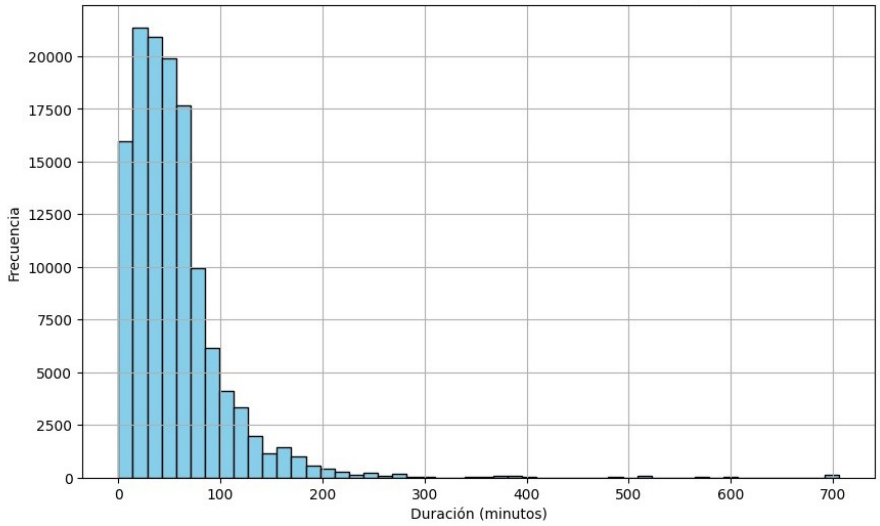


Figura 1: Histograma de duración de los episodios

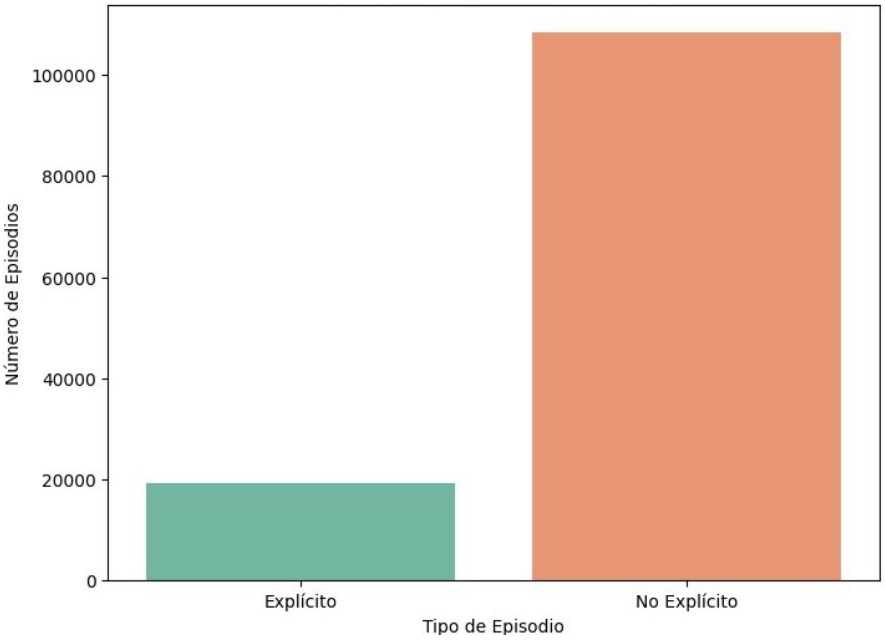


Figura 2: Contenido explícito vs no explícito

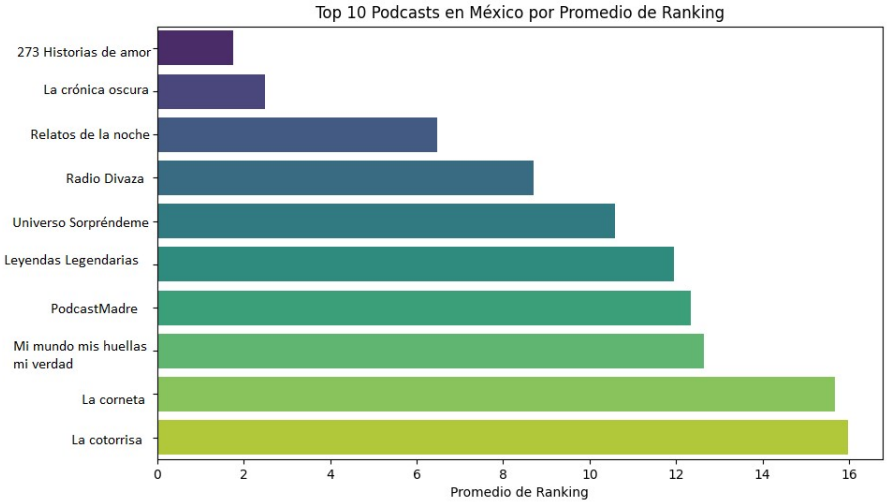
Cuadro 2: Top 10 pódcast a nivel internacional

Nombre	Identificador	Región	Conteo
Mishary Rashid Alafasy	0StCPIsurzmlUZI5eygDaY	id	97
The Joe Rogan Experience	4rOoJ6Egrf8K2IrywzwOMk	us	73
The Joe Rogan Experience	4rOoJ6Egrf8K2IrywzwOMk	ca	72
The Joe Rogan Experience	4rOoJ6Egrf8K2IrywzwOMk	au	71
Relatos de la noche	0SAv0bEYhFndhLODSzPIfL	mx	71
La cruda	2G0HRZba65w6T9NDNScNK2	ar	71
The Joe Rogan Experience	4rOoJ6Egrf8K2IrywzwOMk	nz	71
Café Com Deus Pai — Podcast oficial	5Yl4ao85LeyLc56nvm1E2T	br	69
LEGEND	4xQ0IUSsrfs6pltTnwr3Kk	fr	68
Não Inviabilize	66XCLKbi33MubYTZX2G2jW	br	61

Cuadro 3: Pódcast con mejor ranking promedio mx

Nombre	Ranking promedio
273 Historias de amor	1.75
La crónica oscura	2.5
Relatos de la noche	6.51
Radio divaza	8.71
Universo sorpréndeme	10.6
Podcastmadre	12.33
Mi Mundo Mis Huellas Mi Verdad by Carolina Cruz	12.63
La cotorrisa	15.985
El mundo del sí	16.04
Crímenes	16.75

Figura 3: Top 10 MX por promedio de ranking



Cuadro 4: Pódcast más repetidos en el top 200 mx

Nombre	Veces en top 200
Leyendas legendarias	102
Sonoro	102
Gusgri Podcast	102
La cotorrisa	102
La corneta	102
Hermanos de leche	102
Relatos de la noche	102
Panda Show - Picante	102
Las damitas histeria	102
Paranormal	101

Cuadro 5: Pódcast con más días en el top 10 mx

Nombre	Días en top 10
Relatos de la noche	71
La cotorrisa	38
La corneta	35
Microdosis de amor propio	24
Leyendas legendarias	19
Hermanos de leche	18
Relatos de la noche (Bret Michaels)	10
Nude Project Podcast	9
El mundo del sí	7
El buen pedo	7

En la tabla 2 se muestran los shows que aparecen con mayor frecuencia en el top 10 a nivel internacional.

En la tabla 3 se muestra el top 10 de pódcast por promedio de ranking a nivel nacional.

La figura 3 muestra el top 10 Pódcast en México por promedio de Ranking.

En la figura 3 se observa que el pódcast con mejor promedio de ranking es "273 Historias de amor", sin embargo también se analizaron los pódcast que aparecen más veces en el top 200 de México, que se muestran en la tabla 4.

Finalmente, se filtraron los podcast que aparecen más veces en el top 10, mostrados en la tabla 5.

se analizó el movimiento del ranking para identificar si había tendencias emergentes o pódcast que se viralizaran, pero no se arrojó ningún resultado.

5. Metodología

5.1. Series temporales

Una serie temporal es un conjunto de observaciones que se obtiene midiendo una variable única de manera regular a lo largo de un período de tiempo. Según realicemos la medida de la variable considerada podemos distinguir distintos tipos de series temporales:

- Discretas o Continuas, en base al intervalo de tiempo considerado para su medición.
- Flujo o Stock. Se utilizan mayormente en Economía; se dice que una serie de datos es de tipo flujo si está referida a un período determinado de tiempo. Por su parte, se dice que una serie de datos es de tipo stock si está referida a una fecha determinada.
- De acuerdo con la unidad de medida, podemos encontrar series temporales en euros o en diversas magnitudes físicas.
- Con base en la periodicidad de los datos, se distinguen series temporales de datos diarios, semanales, mensuales, etc.

Una de las razones más importantes para realizar el análisis de las series temporales es intentar predecir los valores futuros de la serie. Un modelo de la serie que explique los valores pasados también puede predecir si aumentarán o disminuirán los próximos valores y en qué medida lo harán.

5.2. Bosques aleatorios

El algoritmo Random Forest (Breiman, 2001) es una técnica de aprendizaje supervisado que construye múltiples árboles de decisión a partir de un conjunto de datos de entrenamiento. Los resultados de estos árboles se combinan para generar un modelo único y robusto, que supera en precisión a los resultados de cada árbol individual.

La construcción de los árboles sigue un proceso en dos etapas:

- Selección aleatoria de predictores: Se generan numerosos árboles de decisión utilizando un subconjunto aleatorio de variables m (predictores) en cada división, donde m¡M y M representa el total de predictores disponibles.
- Bootstrap para muestreo: Cada árbol se construye a partir de un conjunto aleatorio de observaciones seleccionadas mediante la técnica de bootstrap. Este método permite que una misma observación pueda aparecer en múltiples muestras. Las observaciones no seleccionadas en un árbol específico, conocidas como out-of-bag (OOB), se emplean para validar el modelo.

Finalmente, las predicciones de todos los árboles se combinan para producir una única salida final, denominada ensamblado. La combinación de los resultados depende del tipo de problema:

- Regresión: Se calcula el promedio de las predicciones numéricas.
- Clasificación: Se aplica un sistema de votación mayoritaria entre los árboles.

5.3. Gradient Boosting

Gradient Boosting es una técnica de aprendizaje automático diseñada para abordar problemas tanto de regresión como de clasificación. Su enfoque se basa en mejorar iterativamente modelos simples, generalmente árboles de decisión, combinándolos de manera secuencial para construir un modelo final más robusto y preciso.

La esencia de este método radica en que cada modelo nuevo se entrena para corregir los errores cometidos por los modelos previos, optimizando así el rendimiento global del conjunto.

5.4. Árboles de decisión

Un árbol de decisión es un modelo predictivo que organiza los predictores y agrupa las observaciones con valores similares en relación con la variable dependiente o de respuesta. Este modelo pertenece al aprendizaje automático supervisado, ya que requiere una variable dependiente en el conjunto de datos de entrenamiento para que el modelo pueda aprender.

El árbol se construye mediante una serie de decisiones representadas como preguntas, cada una de las cuales genera dos posibles respuestas: sí o no.

- Si la variable dependiente es numérica, el modelo se clasifica como un árbol de regresión.
- Si la variable es categórica, se trata de un árbol de clasificación.

5.5. Regresión

El modelo de regresión lineal (Legendre, Gauss, Galton y Pearson) considera que, dado un conjunto de observaciones $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$, la media μ de la variable respuesta y se relaciona de forma lineal con la o las variables regresoras $x_1, \ldots, x_p$ de acuerdo con la ecuación:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

El resultado de esta ecuación se conoce como la línea de regresión poblacional, y describe la relación entre los predictores y la media de la variable respuesta.

Otra definición frecuentemente encontrada en libros de estadística es:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

En este caso, se hace referencia al valor de y para una observación i concreta. Dado que el valor de una observación puntual no coincide exactamente con el promedio, se incluye un término de error ϵ.

En ambos casos, los elementos del modelo tienen las siguientes interpretaciones:

- β_0: Es la ordenada en el origen, correspondiente al valor promedio de la variable respuesta y cuando todos los predictores son cero.
- β_j: Representa el efecto promedio sobre la variable respuesta de un incremento en una unidad de la variable predictora x_j, manteniendo constantes el resto de predictores. Se les conoce como coeficientes parciales de regresión.
- ϵ: Es el residuo o error, la diferencia entre el valor observado y el estimado por el modelo. Captura el efecto de todas las variables que influyen en y pero no están incluidas como predictores.

En la mayoría de los casos, los valores poblacionales β_0 y β_j son desconocidos, por lo que a partir de una muestra se obtienen sus estimaciones $\hat{\beta}_0$ y $\hat{\beta}_j$. Ajustar el modelo consiste en estimar, a partir de los datos disponibles, los coeficientes de regresión que maximizan la verosimilitud (*likelihood*), es decir, los que generan el modelo más probable para los datos observados.

Los coeficientes de regresión determinan la influencia de cada predictor en el modelo. Sin embargo, su magnitud depende de las unidades en que se midan los predictores, por lo que no está directamente relacionada con la importancia de cada predictor. Para interpretar correctamente la influencia relativa de los predictores, es necesario estandarizarlos antes de ajustar el modelo, asegurando que un coeficiente más cercano a cero indique menor influencia en la variable respuesta.

5.5.1. Regularización

Las estrategias de regularización incorporan penalizaciones en el ajuste por mínimos cuadrados ordinarios (OLS) con el objetivo de evitar *overfitting*, reducir varianza, atenuar el efecto de la correlación entre predictores y minimizar la influencia en el modelo de los predictores menos relevantes. Generalmente, al aplicar regularización se obtienen modelos con mayor poder predictivo (*generalización*).

Dado que estos métodos de regularización actúan sobre la magnitud de los coeficientes del modelo, todos los predictores deben estar en la misma escala. Por esta razón, es necesario estandarizar o normalizar las variables antes de entrenar el modelo.

Ridge

La regularización Ridge penaliza la suma de los cuadrados de los coeficientes:

$$||\beta||_2^2 = \sum_{j=1}^{p} \beta_j^2$$

A esta penalización se le conoce como *L2* y tiene el efecto de reducir proporcionalmente el valor de todos los coeficientes del modelo, sin que lleguen a ser exactamente cero. El grado de penalización está controlado por el hiperparámetro λ. Cuando $\lambda = 0$, la penalización es nula y el resultado es equivalente al de un modelo lineal por mínimos cuadrados ordinarios (OLS). A medida que λ aumenta, mayor es la penalización y menor es el valor de los coeficientes de los predictores.

La función objetivo de Ridge es:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{suma de residuos cuadrados} + \lambda \sum_{j=1}^{p} \beta_j^2$$

Ventajas: La principal ventaja de Ridge frente al ajuste por OLS es la reducción de varianza. En situaciones donde la relación entre la variable respuesta y los predictores es aproximadamente lineal, las estimaciones por mínimos cuadrados tienen bajo sesgo (*bias*) pero alta varianza. Este problema se agrava cuando el número de predictores es cercano al número de observaciones o si $p > n$, en cuyo caso OLS no es aplicable. Ridge permite reducir la varianza sin aumentar significativamente el sesgo, logrando un menor error total.

Desventajas: El modelo final incluye todos los predictores, ya que los coeficientes tienden a cero, pero nunca son exactamente cero (excepto si $\lambda = \infty$). Aunque los predictores menos relevantes tienen menor influencia en el modelo, siguen apareciendo, lo que dificulta su interpretación.

Lasso

La regularización Lasso penaliza la suma de los valores absolutos de los coeficientes:

$$||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$$

A esta penalización se le conoce como *L1*, y tiene el efecto de forzar a que algunos coeficientes de los predictores sean exactamente cero, lo que implica que los predictores correspondientes quedan excluidos del modelo. El grado de penalización está controlado por el hiperparámetro λ. Cuando $\lambda = 0$, el resultado es equivalente al de un modelo OLS. A medida que λ aumenta, mayor es la penalización y más predictores quedan excluidos.

La función objetivo de Lasso es:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{suma de residuos cuadrados} + \lambda \sum_{j=1}^{p} |\beta_j|$$

Comparación entre Ridge y Lasso

La diferencia clave entre Ridge y Lasso radica en que Lasso puede llevar coeficientes a exactamente cero, realizando una selección de predictores, mientras que Ridge no excluye ninguno. Esto hace que Lasso sea preferible en escenarios donde no todos los predictores son importantes y se desea que los menos relevantes sean eliminados.

Sin embargo, en presencia de predictores altamente correlacionados, Ridge distribuye la influencia entre ellos de manera proporcional, mientras que Lasso tiende a seleccionar uno, asignándole todo el peso y excluyendo al resto. Esto hace que las soluciones de Lasso sean más inestables en escenarios con predictores correlacionados.

Frecuentemente, la relación entre la variable de respuesta (y) y la variable explicativa (x) no es estrictamente lineal. Sin embargo, siguiendo el principio de parsimonia, se prefiere ajustar inicialmente un modelo lineal simple, ya que es el modelo más simple y menos parametrizado. Solo si un modelo más complejo (como un modelo no lineal) demuestra ser significativamente superior al modelo nulo, se justifica su adopción.

Una forma sencilla y eficaz de evaluar la desviación del supuesto de linealidad es a través de la regresión polinomial, que extiende el modelo lineal incorporando potencias de la variable explicativa. El modelo polinomial tiene la forma general:

$$y = a + bx + cx^2 + dx^3 + \ldots$$

Aquí, $x^2, x^3, \ldots$ son términos adicionales que capturan diferentes tipos de curvatura en la relación entre y y x.

En la regresión polinomial:

- Se utiliza una única variable explicativa continua (x). - Se incorporan potencias superiores de x ($x^2, x^3, \ldots$) al modelo, además del término lineal (x). - Esto permite describir diversos patrones de curvatura en la relación entre y y x.

6. Resultados y conclusiones

Se realizó inicialmente el análisis de series temporales para visualizar la Evolución de los Rankings Promedio de los 10 Podcasts Más Frecuentes en México, mostrado en la figura 4.

Posteriormente, se hizo la matriz de correlación entre la variable de interés rank y las variables predictoras durationms, explicit6, showmediatype, is_externally_hosted, is_playable, y showtotal_episodes. Dicha matriz se muestra en la figura 5.

De esta matriz se obtienen las siguientes conclusiones:

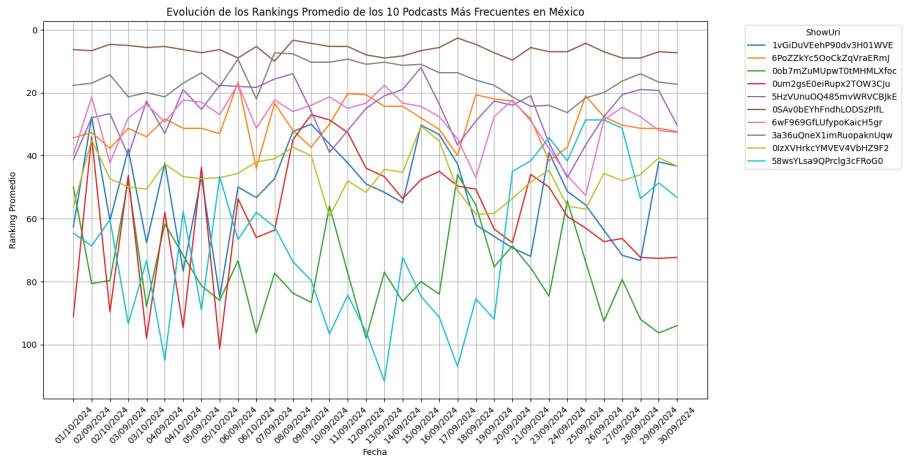


Figura 4: Evolución de los Rankings Promedio de los 10 Podcasts Más Frecuentes en México

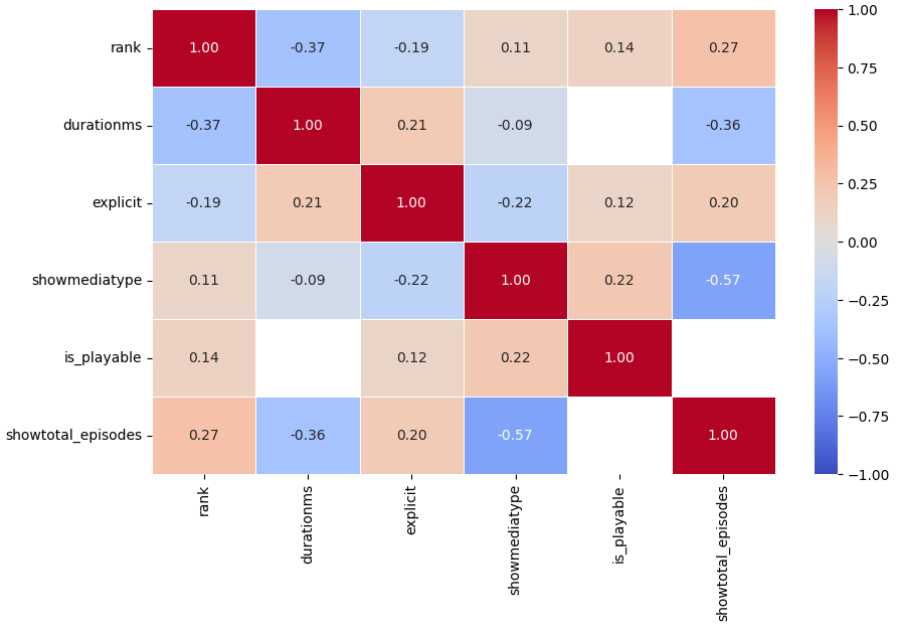


Figura 5: Matriz de Correlación de los 10 Podcasts más Repetidos MX

Cuadro 6: Modelos de regresión

Modelo	RMSE	R2
Regresión lineal	40.87	0.2266
Regresión lineal de Lasso	40.883	0.2266
Regresión Ridge	40.88	0.2266
Regresión polinómica	40.52	0.233
Árboles de decisión	37.33	0.34
Bosques aleatorios	38.16	0.32
Gradient boosting	36.76	0.36
XGBoost	32.86	0.37

- rank tiene una correlación negativa moderada con durationms de -0.37. Esto sugiere que a medida que el rank aumenta, la duración del podcast tiende a ser más corta, aunque la relación no es muy fuerte.
- explicit y durationms tienen una correlación positiva moderada de 0.21, lo que indica que los podcasts explícitos tienden a tener una duración ligeramente mayor.
- showmediatype tiene correlaciones bajas con las demás variables, lo que sugiere que el tipo de medio del show (audio vs. mixto) no está fuertemente relacionado con el rank, la duración, o el hecho de que sea explícito.
- is_playable tiene una pequeña correlación positiva con rank (0.14) y explicit (0.12), lo que sugiere que los podcasts que son accesibles y se pueden reproducir en la plataforma podrían estar ligeramente relacionados con un ranking más alto y con el hecho de ser explícitos.
- showtotal_episodes tiene una correlación positiva con rank (0.27), lo que sugiere que los podcasts con más episodios tienden a estar mejor posicionados. Tambien tene una correlación negativa moderada con showmediatype (-0.57), lo que indica que los podcasts con más episodios tienden a ser de tipo .ªudio.en lugar de "mixto".

Se realizaron pruebas usando distintos modelos de regresión, así como los métodos de árboles de decisión, bosques aleatoios, gradient boosting y XGBoost.

En la tabla 6 se muestran los resultados:

Se observa un mejor rendimiento con el modelo de XGBoost, sin embargo, un RMSE de 32.86 implica que las predicciones del modelo están desviándose, en promedio, unas 33 posiciones del valor real de rank, lo que resulta bastante significativo para los fines de este estudio.

Es importante continuar trabajando con un ajust de hiperparámetros, validación cruzada, o bien, máquinas de soporte vectorial.

7. Bibliografía

- Vinuesa, P. (s.f.). *Tema 9: regresión lineal simple y polinomial: teoría y práctica*. Recuperado de `https://www.ccg.unam.mx/~vinuesa/R4biosciences/docs/Tema9_regresion_presentacionR.html#/`
- *Vista de Big data in radio broadcasting companies: applications and evolution.* (s.f.). Recuperado de `https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/86918/63240`
- Dután, W. O. (2024, 21 de enero). *Periodismo radial y las nuevas tendencias comunicativas en el cantón La Libertad. Caso: panorama informativo de radio Amor 89.3FM*. Recuperado de `https://repositorio.upse.edu.ec/handle/46000/10701`
- Figueroa Portilla, C. S. (2019). *El Pódcast: un medio y una forma de comunicación. Acta Herediana*, 62(2), 129–133. `https://doi.org/10.20453/ah.v62i2.3615`