Universidad Miguel Hernández

FACULTAD DE CIENCIAS SOCIALES Y JURÍDICAS DE ELCHE

GRADO EN ESTADÍSTICA EMPRESARIAL

Desarrollo de un Sistema de Recomendación de Canciones Basado en Análisis de Componentes Principales y Métodos de Clasificación Avanzados

Trabajo Fin de Grado

Autora: Rocío Pérez Jiménez

Tutora: Mª Asunción Martínez Mayoral

Curso 2024/2025

Índice

Resumen ***3***

Agradecimientos ***3***

Palabras clave ***3***

1. Antecedentes ***4***

2. Objetivos ***5***

3. Información disponible ***6***

4. Metodología ***10***

4.1 Análisis exploratorio y preprocesado **10**

4.1.1 Análisis Univariado 11

4.1.2 Análisis Bivariado 12

4.2 Aprendizaje no supervisado para la reducción de la dimensión **12**

4.2.1 Análisis de Componentes Principales (PCA) para Definir una Variable de Éxito 13

4.3 Aprendizaje supervisado para la predicción del éxito **15**

4.3.1 Modelo lineal 15

4.3.2 Árboles de decisión 17

4.3.3 Random Forest 19

4.4 Software **22**

5. Resultados ***24***

5.1 Análisis exploratorio y preprocesado **24**

5.1.1 Análisis Univariado 24

5.1.2 Análisis bivariado 29

5.2 Aprendizaje no supervisado para la reducción de la dimensión **31**

5.3 Aprendizaje supervisado para la predicción del éxito **34**

5.3.1 Modelo lineal 35

5.3.2 Árboles de decisión 38

5.3.3 Random Forest 41

6. Conclusiones ***47***

7. Referencias ***48***

Resumen

En este estudio abordamos el análisis de un banco de datos proveniente de *Spotify* y *YouTube*, con un conjunto de temas y cantantes top10, con los que pretendemos desarrollar un análisis aproximado a lo que podría plantear un sistema de recomendación, para primero reconocer el éxito en un único indicador, luego caracterizarlo en función de la información disponible sobre las canciones, y terminar agrupando canciones por afinidad, para poder hacer recomendaciones a los usuarios en función de sus preferencias al consumir productos en la plataforma. Todo esto lo trabajamos mediante técnicas de aprendizaje automático o *machine learning*.

Agradecimientos

Quisiera expresar mi más sincero agradecimiento a mi tutora, Mª Asunción Martínez, por su dedicación, guía y apoyo constante a lo largo de todo el proceso de realización de este trabajo. Su implicación y sus orientaciones han sido fundamentales para poder llevarlo a buen término.

También quiero dar las gracias a mis amigos, por su compañía, sus palabras de ánimo y por estar siempre dispuestos a ayudarme cuando más lo necesitaba.

Y, por supuesto, a mi familia, por su apoyo incondicional, su paciencia y por creer en mí en cada etapa de este camino académico.

Palabras clave

Indicadores de éxito, análisis de componentes principales, modelos de predicción, árboles de decisión, Random Forest, recomendación basada en contenido.

1. Antecedentes

Las plataformas digitales de música, como *Spotify*, acumulan una vasta cantidad de datos de sus usuarios para así personalizar experiencias, mejorar servicios y optimizar publicidad. *YouTube* y *Spotify* son dos de las **plataformas de streaming** más utilizadas globalmente. Por un lado, *YouTube* tiene más de 2 mil millones de usuarios activos mensuales y es un líder en vídeos, incluyendo música, mientras que *Spotify*, centrado en la música, cuenta con más de 550 millones de usuarios activos mensuales, 220 millones de suscriptores de pago, y un catálogo de más de 100 millones de canciones. Ambas plataformas son clave en la distribución y consumo de música digital.

El **etiquetado de los productos** que se distribuyen en estas plataformas resulta fundamental para estructurar y organizar adecuadamente la información sobre las canciones. Estas etiquetas incluyen el título de la canción, el nombre del artista principal y los colaboradores, así como el nombre del álbum al que pertenece. También se registra el género musical, el año de lanzamiento, los compositores y los productores responsables de la grabación. Otros datos clave son la duración de la canción, el idioma, el número de pista, la versión (original, remix o acústica) y la letra completa, si está disponible. Este conjunto de etiquetas facilita la correcta **identificación y análisis de las canciones**, permitiendo una base sólida sobre la cual desarrollar distintos enfoques analíticos.

Estas plataformas además recopilan **datos demográficos** (edad, género, ubicación, etc) y **preferencias musicales** (géneros y artistas favoritos) para segmentar a los usuarios. También registran patrones de uso, como frecuencia y duración de escucha, interacciones con contenido (reproducción, saltos, descargas, etc) y búsquedas realizadas en la plataforma. Finalmente, analizan el rendimiento de las canciones y artistas para ajustar sus algoritmos y mejorar la experiencia del usuario a través de los sistemas de recomendación personalizados.

Los **sistemas de recomendación personalizados** comenzaron a desarrollarse en los años 90 con proyectos académicos como *Tapestry* (1992) y *GroupLens* (1994), basados en filtrado colaborativo. A partir de los 2000, empresas como *Amazon* y *Netflix* impulsaron su evolución, aplicándolos a gran escala en comercio electrónico y entretenimiento. En 2006, el *Netflix Prize* marcó un hito al promover mejoras en algoritmos predictivos,

sentando las bases para los sistemas actuales utilizados por plataformas como *Spotify* y *YouTube*.

Estos sistemas son herramientas basadas en algoritmos que tienen como objetivo predecir las preferencias o intereses de los usuarios y ofrecerles **sugerencias personalizadas** de productos o contenidos. Surgieron gracias a avances en análisis de datos, inteligencia artificial y computación, y se han consolidado como componentes clave en **plataformas digitales**. Utilizan información tanto de los productos como de los usuarios para generar recomendaciones relevantes.

Con este tipo de sistemas, *YouTube* analiza el historial de visualización y las interacciones del usuario para sugerir vídeos de interés, mientras que *Spotify* emplea datos de escucha y patrones de comportamiento para crear listas de reproducción personalizadas.

Las plataformas de gestión de contenido multimedia utilizan diversas técnicas de ***machine learning*** y **aprendizaje automático** para analizar grandes volúmenes de datos y mejorar la experiencia del usuario a través de recomendaciones adaptadas a sus gustos y hábitos.

A continuación, se detallan los objetivos que guían el desarrollo de este estudio.

2. Objetivos

La base de datos que hemos elegido para trabajar está disponible en *Kaggle*, identificada con el nombre "*Spotify and Youtube*". Contiene información sobre las características de las 10 mejores canciones de varios artistas de *Spotify* y sus vídeos de *YouTube*.

El **objetivo principal** del trabajo es analizar esta base de datos con el fin de conseguir **agrupaciones de canciones afines con las que diseñar la estructura de un sistema de recomendación** basado en contenido, con el que agrupamos contenido afín, y así poder ofrecer sugerencias a los usuarios, similares o afines a los temas/canciones que consumen o les gustan.

Los objetivos específicos que trabajamos son:

- Calcular un único **indicador de éxito** para cada canción a partir de variables disponibles relacionadas con el éxito de los temas y correlacionadas entre sí.
- **Predecir el éxito** en función del resto de variables, tanto numéricas como categóricas, disponibles sobre cada canción, identificando cuáles de las variables disponibles aportan información sobre el éxito.
- Conseguir una **agrupación de las canciones** en perfiles de similaridad de éxito, con las que plantear la base de un sistema de recomendación al usuario, basado en el contenido.

3. Información disponible

Los datos, como hemos comentado anteriormente, han sido extraídos de ***Kaggle***, una plataforma en línea conocida por proporcionar una extensa colección de conjuntos de datos públicos. Específicamente, la información para este proyecto proviene del conjunto de datos *"Spotify and YouTube"*, publicada en *[https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube/data]*, y proporcionados por *Salvatore Rastelli, Marco Guarisco y Marco Sallustio.* Los datos han sido descargados en formato csv, tal y como estaban publicados en *Kaggle*.

Cabe señalar que los datos utilizados fueron recopilados el **2 de julio de 2023** y actualizados por última vez hace un año.

Este conjunto de datos ha sido analizado por los usuarios de *Kaggle* para **desarrollar diferentes proyectos**, utilizando principalmente los lenguajes de programación Python y R. Se han realizado trabajos como análisis exploratorios (Abhishek Pal, 2024), bosques aleatorios de regresión (Random Forest) para predecir valores numéricos (Decherisey, H, 2023), y técnicas de reducción de dimensionalidad, como el Análisis de Componentes Principales (PCA) (F. Zumpano, 2024). Adicionalmente, se han desarrollado proyectos de investigación que contrastan canciones antiguas con las más recientes (Pilgaonkar, 2023).

El conjunto de datos incluye información detallada sobre las 10 mejores canciones en *Spotify* y sus respectivos vídeos en *YouTube* de distintos artistas de todo el mundo, lo que

nos permite realizar un análisis exhaustivo de las características que pueden estar relacionadas con el **éxito** de dichas canciones.

Contamos con una base de datos de 2.718 entradas con 26 variables (14 numéricas y 12 categóricas) para cada una de las canciones recogidas en las plataformas. De algunas de estas variables se ha decidido prescindir en el análisis, dado que no aportan información relevante para los objetivos propuestos en este trabajo. Las variables que hemos utilizado se describen a continuación:

VARIABLES NUMÉRICAS:

- ***Danceability***: es una medida numérica que indica lo bueno que es un tema musical para bailar. Se calcula considerando factores como el tempo, la regularidad del ritmo y su fuerza. Su escala de variación está entre 0 y 1. Una puntuación de 0 indica que la canción es poco bailable, mientras que 1 significa que es muy bailable.

- ***Energy***: Mide la intensidad y actividad percibida en una canción. Las canciones enérgicas son rápidas, ruidosas y tienen un sonido con un valor más enérgico. Su escala de variación está entre 0 y 1. El death metal es un ejemplo de un género con alta energía (valores próximos a 1), mientras que la música clásica suele tener una energía más baja (valores próximos a 0).

- ***Speechiness***: Mide el grado en que una pista de audio está compuesta principalmente por voz humana. Un valor de 1 indica una grabación exclusivamente vocal, mientras que un valor cercano a 0 sugiere una pista predominantemente musical. Los rangos intermedios (0.33-0.66) indican una combinación de voz y música, como en el rap. Los valores inferiores a 0,33 representan música y otras pistas que no son de voz.

- ***Loudness***: la sonoridad global de una pista en decibelios (dB). Los valores oscilan entre -60 y 0 db.

- ***Acousticness***: Mide el grado en que una canción es acústica. Un valor cercano a 1 indica una alta probabilidad de que la pista haya sido grabada principalmente con instrumentos acústicos y con mínima intervención de producción electrónica. La variable toma valores continuos en una escala de 0 a 1.

- ***Instrumentalness***: Mide la probabilidad de que una pista esté compuesta exclusivamente de instrumentos musicales. Un valor cercano a 1 indica una alta probabilidad de que la pista no contenga voces. Su escala de variación está entre 0 y 1.

- ***Liveness***: Mide la probabilidad de que una canción haya sido grabada en vivo frente a una audiencia. Valores cercanos a 1 indican una alta probabilidad de que la grabación haya sido en vivo. Su escala de variación está entre 0 y 1.

- ***Valence***: Mide la positividad emocional percibida en una canción. Los valores cercanos a 1 indican que la pista transmite sensaciones alegres, felices o eufóricas, mientras que los valores próximos a 0 reflejan emociones negativas, como tristeza o enfado. Su escala de variación está entre 0 y 1.

- ***Tempo***: el tempo general estimado de una pista en pulsaciones por minuto (BPM). En terminología musical, el tempo es la velocidad o el ritmo de una pieza determinada y se deriva directamente de la duración promedio del tiempo.

- ***Duration***: la duración de la pista en minutos.

- ***Stream***: número total de veces que la canción ha sido reproducida en *Spotify*.

- ***Views***: cantidad de visualizaciones registradas para la canción en *Spotify*.

- ***Likes***: número de "me gusta" que ha recibido la canción por parte de los usuarios en la plataforma.

- ***Comments***: cantidad de comentarios que los usuarios han dejado sobre la canción en *Spotify*.

VARIABLE CATEGÓRICA:

- ***Key***: Indica la tonalidad principal o nota tónica de una pista musical, codificada numéricamente según la notación estándar de clase de tono. A continuación las 12 clases de tono musicales en la *Tabla 1*:

  Tabla 1. Relación entre las clases de tonos musicales y sus correspondientes notas.

Clase	Tono musical
-1	no se pudo identificar
0	C
1	C♯/D♭
2	D
3	D♯/E♭
4	E
5	F
6	F♯/G♭
7	G
8	G♯/A♭
9	A
10	A♯/B♭
11	B

Tras la definición de los objetivos, el siguiente paso es seleccionar y aplicar las técnicas adecuadas para abordarlos de forma rigurosa.

En el siguiente apartado se describe el enfoque metodológico seguido, incluyendo las fases de preprocesado, análisis exploratorio y construcción de modelos supervisados y no supervisados. Se detallan las herramientas estadísticas utilizadas, así como los criterios para evaluar y comparar el rendimiento de los modelos desarrollados

4. Metodología

El análisis del **éxito musical** se ha abordado mediante distintas técnicas estadísticas y de aprendizaje automático, seleccionadas en función del objetivo de cada etapa del estudio.

Para sintetizar la información disponible sobre el grado de éxito alcanzado para un tema (en función de las reacciones de los usuarios), se aplicó el **Análisis de Componentes Principales (PCA)**, que permite transformar un conjunto de variables correlacionadas en un número reducido de componentes no correlacionadas, conservando la mayor parte de la variabilidad de los datos originales. Esta técnica fue clave para construir un **indicador continuo de éxito musical**.

Posteriormente, con el objetivo de predecir este indicador en función de las caracterizaciones disponibles sobre los temas en la base de datos, se emplearon métodos supervisados como **árboles de decisión** y ***Random Forest***, tanto en su versión de regresión como de clasificación. Estos modelos permiten estimar el nivel de éxito de una canción a partir de sus características musicales, evaluar su precisión mediante métricas específicas y detectar qué variables influyen más en el resultado. Además facilitan como resultado perfiles de agrupación de temas, que se pueden convertir en la base de un sistema de recomendación a usuarios, basado en el contenido, pero también a productores para conseguir temas de éxito.

4.1 Análisis exploratorio y preprocesado

Para entender la estructura, la calidad y la naturaleza de los datos antes de aplicar modelos predictivos, debemos hacer un **análisis exploratorio de los datos**, que permita comprender la estructura de los datos, identificar posibles problemas de calidad y tomar decisiones informadas sobre el preprocesado, y con este último garantizamos la idoneidad de los datos para los objetivos del estudio.

Para ello distinguimos entre **técnicas univariantes**, con las que describir de modo diferenciado cada una de las variables en la base de datos y realizar el procesado necesario, y **técnicas inferenciales**, con las que indagar sobre las posibles relaciones entre ellas, y en especial con las variables de éxito. Las presentamos a continuación.

4.1.1 Análisis Univariado

El análisis exploratorio univariado se conjuga con el preprocesado para revisar la información contenida en las diversas variables y prepararlas para su incorporación efectiva en los modelos inferenciales. De hecho, el análisis exploratorio permite obtener una primera aproximación a la estructura interna de los datos y facilita la toma de decisiones en cuanto al preprocesamiento posterior.

En el procesado de los datos consideramos las siguientes tareas:

- Identificación de valores faltantes: Se detectan y contabilizan los valores faltantes en cada variable.

- Imputación de valores faltantes, a través del método basado en medias para variables numéricas, y en las respuestas más frecuentes para la variable categórica, relativamente robusto cuando el número de valores faltantes no es excesivo en comparación con el total de datos.

- Estandarización de variables numéricas: Se examinan las escalas de variación para identificar en qué variables será necesaria una estandarización de los datos, para evitar un efecto de escala en los modelos de aprendizaje. La estandarización que se utiliza consiste en restar la media y dividir por la desviación típica.

- Viabilidad del uso de variables categóricas en función de que tengan un número de respuestas o niveles razonable para el análisis, o contemplar una posible recodificación.

Para diferenciar el tratamiento en el análisis que vamos a llevar a cabo, se clasificaron las variables del conjunto de datos en función de su rol:

- Variables de respuesta, asimilables como indicadores de éxito: ***Likes, Comments, Views y Stream***. Estas variables cuantifican el nivel de popularidad o rendimiento de cada elemento del conjunto de datos, y con ellas intentamos construir un único indicador de éxito. Están dimensionadas en escalas numéricas distintas.

- Variables predictoras: que describen características musicales y técnicas que podrían influir en el éxito de una canción. Están dimensionadas en diferentes escalas (***Loudness, Tempo y Duration),*** y dimensionadas con valores entre 0 y 1

(***Danceability, Energy, Speechiness, Acousticness, Instrumentalness, Liveness, Valence).*** También tenemos una variable categórica, que es ***Key.***

Para describir las **variables numéricas**, hemos empleado estadísticas numéricas como percentiles, e histogramas y diagramas de cajas para entender mejor su comportamiento y distribución de forma gráfica.

Para representar la información en la variable categórica *Key*, hemos utilizado diagramas de barras con los porcentajes.

4.1.2 Análisis Bivariado

El análisis exploratorio bivariado es una herramienta fundamental en la ciencia de datos que nos permite investigar las **relaciones entre dos o más variables** en un conjunto de datos. Dado que la mayoría de las variables a relacionar son de tipo numérico, utilizamos **correlogramas** y **mapas de calor** para investigar las relaciones (de tipo lineal) entre las variables numéricas disponibles. También se plantean **diagramas de dispersión** para identificar el tipo de relación entre dichas variables.

Además, proporcionan la fundamentación para un análisis de reducción de la dimensión, identificando correlaciones altas entre las variables que están relacionadas directamente con la popularidad o el éxito de las canciones.

4.2 Aprendizaje no supervisado para la reducción de la dimensión

Para resolver el objetivo específico 1 se aborda el análisis de **reducción de la dimensión**, una técnica crucial para simplificar conjuntos de datos complejos. Este enfoque permite identificar y conservar las características más relevantes de los datos, eliminando redundancias y reduciendo el ruido. Al reducir la cantidad de variables, se **facilita el análisis, se mejora la eficiencia de los modelos y se potencia la capacidad de interpretación**, sin sacrificar la calidad de la información esencial para la toma de decisiones.

El análisis de la reducción de la dimensión se trata de una técnica de **aprendizaje no supervisado**, pues se lleva a cabo sin la necesidad de etiquetas o variables de salida. Este tipo de aprendizaje permite explorar la estructura subyacente de los datos de forma

automática, encontrando patrones o representaciones más compactas, sin necesidad de supervisión externa. Entre las técnicas más empleadas en este contexto destaca el **Análisis de Componentes Principales (PCA)**, una herramienta estadística que transforma los datos originales en un nuevo conjunto de variables no correlacionadas, que capturan la mayor cantidad de variabilidad presente en el conjunto de datos.

A continuación se presenta esta técnica con mayor profundidad, explicando su metodología, su proceso de construcción y su utilidad en el análisis de datos.

4.2.1 Análisis de Componentes Principales (PCA) para Definir una Variable de Éxito

El Análisis de Componentes Principales (PCA) es una técnica de aprendizaje no supervisado ampliamente utilizada en el análisis exploratorio de datos. Su objetivo principal es **reducir la dimensionalidad** de un conjunto de variables cuantitativas, **manteniendo la mayor cantidad posible de información**. Esta técnica es especialmente útil cuando se trabaja con muchas variables que pueden estar correlacionadas entre sí, lo cual sugiere la existencia de información redundante. En estos casos, el **PCA** permite transformar el conjunto original en un número reducido de nuevas variables, llamadas **componentes principales**, que son combinaciones lineales de las variables originales y que entre sí no están correlacionadas.

Desde un punto de vista geométrico, el PCA puede interpretarse como una rotación del sistema de coordenadas original hacia una nueva base ortogonal, de forma que los nuevos ejes coincidan con las direcciones de mayor varianza en los datos. Cada componente principal representa una dimensión del nuevo espacio reducido, ordenadas según la cantidad de variabilidad que explican. De este modo, es posible simplificar la estructura del conjunto de datos, facilitando su interpretación, visualización y análisis posterior, sin perder información relevante.

En el contexto de este trabajo, se aplica PCA sobre cuatro variables que están directamente relacionadas con el éxito de una canción: *Streams, Views, Likes y Comments*. Estas variables reflejan diferentes aspectos del rendimiento o impacto de un contenido, y es razonable suponer que **contienen información redundante** o altamente correlacionada entre sí. El propósito del análisis es construir una nueva variable —la primera componente principal— que actúe como un indicador sintético de **"éxito"**,

condensando en una sola dimensión la información más relevante de las cuatro variables originales.

El procedimiento seguido es el siguiente:

Estandarización de las variables:

Antes de aplicar PCA, es necesario que las cuatro variables se encuentren estandarizadas, es decir, con media cero y varianza unitaria, ya que están en distintas escalas y PCA es sensible a las magnitudes de las variables.

Cálculo de la matriz de covarianza (o correlación):

A partir de los datos estandarizados, se construye la matriz de correlación, ya que en este caso se utiliza para evaluar las relaciones lineales entre las variables.

Construcción de las componentes principales:

Se calculan los autovalores y autovectores de la matriz de correlación. Los **autovalores** indican la cantidad de varianza explicada por cada componente principal, y los **autovectores** determinan la combinación lineal de variables originales que define cada componente.

El primer componente principal se obtiene como una **combinación lineal de las variables originales estandarizadas** que captura la mayor parte de la variabilidad conjunta de las cuatro variables. A continuación, se construyen ortogonalmente otras componentes, también como combinaciones lineales de las variables originales, que van acumulando la variabilidad restante.

Interpretación y validación:

Se considerará válida la utilización del primer componente principal como un indicador del éxito siempre que este explique **al menos el 70% de la varianza total** de los datos. En caso de que no cumpla con este criterio, se optará por utilizar los dos primeros componentes principales, de forma que se capte una proporción razonable de la variabilidad original, y por lo tanto no se pierda información sustancial.

El uso de PCA en este caso no solo permite reducir la dimensión del problema, sino también construir una nueva variable latente que resume eficazmente la información

contenida en múltiples indicadores del éxito de un tema. Esta técnica facilita tanto la interpretación como el uso posterior de los datos en modelos más complejos o en análisis descriptivos.

4.3 Aprendizaje supervisado para la predicción del éxito

Una vez disponible un único indicador de éxito, se aplican distintos métodos de aprendizaje supervisado con el objetivo de **predecir el éxito** de una canción a partir de sus características musicales. Se presentan tres enfoques: *un modelo lineal, árboles de decisión* y *el método Random Forest*. Cada uno de ellos permite explorar la **relación entre las variables predictoras y la variable de éxito** desde distintas perspectivas, comparando su rendimiento y capacidad interpretativa.

4.3.1 Modelo lineal

El modelo lineal es una técnica básica y fundamental en el aprendizaje supervisado, que permite establecer un modelo relacional simple y lineal entre una **variable respuesta continua** y un conjunto de variables predictoras. En este caso, el objetivo es modelar el éxito de una canción —representado por la variable continua— a partir de diferentes características musicales, todas ellas numéricas y previamente estandarizadas. Este enfoque se basa en la suposición de una relación lineal entre las variables, y su interpretación y aplicación han sido ampliamente desarrolladas en la literatura estadística (James, Witten, Hastie & Tibshirani, 2021).

Este modelo asume que existe una relación lineal entre la variable dependiente y cada predictor, es decir, que el efecto de cada variable sobre el éxito puede estimarse como una suma ponderada de los predictores. Cada coeficiente estimado en el modelo indica cuánto se espera que cambie el éxito ante una variación unitaria en el predictor correspondiente, manteniendo constantes los demás.

La construcción del modelo comienza ajustando una versión completa (modelo basal) que incluye **todas las variables predictoras**. A continuación, se aplica un proceso de **selección automática,** que permite identificar un subconjunto óptimo de variables basándose en el criterio de información de **Akaike (AIC)**. Este criterio penaliza la complejidad del modelo, buscando un equilibrio entre la calidad del ajuste y el número

de parámetros incluidos. Específicamente, el AIC estima la pérdida de información al utilizar un modelo determinado para representar la realidad, favoreciendo aquellos modelos que explican adecuadamente los datos con la menor cantidad posible de variables (Akaike, 1974).

Para evaluar la calidad del modelo se utilizan indicadores estadísticos como:

- El **coeficiente de determinación** (R^2), que refleja la proporción de la variabilidad total de la variable dependiente que es explicada por el modelo. Su valor oscila entre 0 y 1, donde valores cercanos a 1 indican un mayor poder explicativo. Su versión ajustada (**R^2 ajustado**) corrige este valor teniendo en cuenta el número de predictores incluidos en el modelo, penalizando aquellos que no aportan mejora sustancial.

  La fórmula del R^2 ajustado es:

$$R^2_{ajustado} = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - p - 1}\right)$$

  Donde:

  - R^2: coeficiente de determinación,
  - n: número total de observaciones,
  - p: número de predictores del modelo.

- El **estadístico *F de bondad de ajuste***, que contrasta la significación global del modelo a partir de las sumas de cuadrados de la regresión y las sumas de cuadrados total, corregidas por sus respectivos **grados de libertad;** un pvalor < 0.05 nos indica que se rechaza un modelo nulo a favor del modelo ajustado

- Otra métrica para evaluar la calidad del ajuste del modelo lineal es el **error cuadrático medio de la raíz (RMSE)**, que mide el error promedio entre los valores predichos por el modelo y los valores observados. Su fórmula es:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)}$$

donde y_i representa los valores observados, $\hat{y}_i$ los valores predichos, y n el número total de observaciones.

Cuanto mayor sea el RMSE, mayor será la discrepancia entre las predicciones del modelo y los datos reales, lo que indica un peor ajuste. Por el contrario, un RMSE más bajo refleja un mejor rendimiento predictivo del modelo, indicando que este es capaz de aproximarse con mayor precisión a los valores reales de la variable dependiente.

4.3.2 Árboles de decisión

Los árboles de decisión son una técnica de **aprendizaje supervisado** ampliamente utilizada tanto para problemas de regresión como de clasificación que nos permite representar visualmente el proceso de toma de decisiones y facilita la interpretación de los resultados.

El algoritmo funciona mediante un proceso de **partición recursiva del espacio de datos**, seleccionando en cada paso la variable predictora más relevante y el punto de corte que mejor separa los datos **según la variable objetivo**. El objetivo es maximizar la homogeneidad dentro de cada nodo resultante, creando divisiones que minimicen la impureza (en clasificación) o la varianza (en regresión). A medida que el árbol crece, se forman nodos hijos que contienen subconjuntos cada vez más específicos del conjunto original.

Para evitar que el árbol crezca de forma excesiva y se sobreajuste a los datos de entrenamiento, es necesario establecer ciertos **criterios de parada** u **hiperparámetros** que regulen su complejidad. Entre los más relevantes se encuentran:

- La **profundidad máxima del árbol** (maxdepth), que limita el número de niveles jerárquicos.

- El **número mínimo de observaciones** en un nodo para permitir su división (minsplit).
- El **número mínimo de observaciones** que debe haber en un nodo terminal (minbucket).

Estos parámetros permiten construir modelos más generalizables, reduciendo el riesgo de ajustar ruido en los datos y facilitando interpretaciones más sencillas. Ajustar correctamente estos valores es esencial para alcanzar un **equilibrio entre precisión y simplicidad del modelo.**

Para construir estos modelos, y dado el gran volumen de datos en la base de datos disponible, se divide el conjunto de datos en dos partes: una de entrenamiento, utilizada para ajustar el árbol, y otra de prueba, destinada a evaluar su capacidad predictiva. Esta partición garantiza una valoración objetiva del rendimiento del modelo y permite controlar el sobreajuste. El **sobreajuste** (overfitting) ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento, capturando tanto los patrones reales como el ruido o las fluctuaciones aleatorias, lo que reduce su capacidad de generalización a nuevos datos.

En este trabajo, se han aplicado dos enfoques: el **árbol de regresión** al considerar el indicador de éxito como variable respuesta de tipo numérico, y el **árbol de clasificación**, al categorizar este indicador y generar un indicador politómico con tres niveles de éxito (leve, moderado y alto).

Árbol de regresión

Los árboles de regresión constituyen una técnica de modelado supervisado utilizada para predecir una variable dependiente de tipo continua a partir de un conjunto de variables explicativas.

Cuando se desea construir un árbol de regresión, se debe especificar el argumento *method = "anova"*, lo que indica que el objetivo del modelo es predecir una variable continua. En este caso, el criterio que guía la construcción del árbol es la **minimización de la varianza intra-nodo**, es decir, se seleccionan aquellos predictores y sus divisiones que maximizan la reducción del **error cuadrático medio (MSE)** en los nodos hijos. Cada división busca

generar subconjuntos que sean lo más homogéneos posible con respecto a la variable respuesta.

Una vez generado el árbol, es posible interpretar visualmente las reglas de decisión y los valores de predicción en cada nodo terminal. Para evaluar cuantitativamente el rendimiento del modelo, se emplea nuevamente el **error cuadrático medio de la raíz (RMSE)**, ya descrito en el apartado del modelo lineal, como medida de la precisión de las predicciones obtenidas a partir del árbol.

Árbol de clasificación

En cuanto a los árboles de clasificación, permiten **predecir una variable categórica** a partir de un conjunto de variables predictoras. Su funcionamiento se basa en dividir de forma recursiva el espacio de los datos mediante reglas de decisión simples, **generando una estructura en forma de árbol**. En cada nodo interno del árbol se evalúa una condición sobre una de las variables predictoras, mientras que en las hojas se asigna una categoría como resultado final.

Para construir el árbol, el algoritmo selecciona en cada paso la variable y el umbral de partición que mejor separen las clases según algún criterio de impureza, como el índice de Gini.

Estos dos tipos de modelos son especialmente útiles por su capacidad para capturar relaciones no lineales entre las variables, su interpretación sencilla y la facilidad para visualizar el proceso de decisión.

4.3.3 Random Forest

Random Forest es un algoritmo de aprendizaje supervisado que se basa en la **construcción de múltiples árboles de decisión**, combinando sus predicciones para mejorar la precisión del modelo y reducir el riesgo de sobreajuste. Esta técnica puede aplicarse tanto a problemas de regresión como de clasificación y se caracteriza por su capacidad para manejar **grandes volúmenes de datos** y relaciones complejas entre variables.

El algoritmo genera múltiples árboles a partir de muestras aleatorias con reemplazo (**bootstrap**) del conjunto de entrenamiento. En cada nodo de cada árbol, se selecciona aleatoriamente un subconjunto de variables predictoras para determinar la mejor partición, lo que introduce diversidad entre los árboles y reduce el riesgo de multicolinealidad. La agregación de resultados permite obtener un modelo más robusto que el basado en un único árbol.

➔ Optimización del modelo

Para evitar que los árboles individuales crezcan en exceso y se ajusten demasiado a los datos (**overfitting**), es necesario regular su complejidad mediante la optimización de varios hiperparámetros. Entre los más importantes se encuentran:

- **ntree**: número de árboles en el bosque.
- **maxdepth**: profundidad máxima del árbol. Limita el número de niveles jerárquicos para evitar sobreajuste y mejorar la interpretabilidad.
- **minsplit**: número mínimo de observaciones requerido para dividir un nodo.
- **minbucket**: número mínimo de observaciones en los nodos terminales.
- **mtry**: número de predictores considerados en cada división.

El ajuste adecuado de estos hiperparámetros permite encontrar un equilibrio entre complejidad del modelo y capacidad predictiva, optimizando su rendimiento.

➔ Importancia de las variables

La importancia de las variables predictoras se estima mediante los siguientes indicadores:

- **MeanDecreaseAccuracy:** evalúa cuánto disminuye la precisión del modelo (medida a través del error OOB, *out-of-bag*) cuando se desordena aleatoriamente una variable. Si al hacer este desorden el rendimiento del modelo empeora considerablemente, se interpreta que esa variable tiene una influencia significativa en la predicción.

- **MeanDecreaseGini**: mide cuánto contribuye una variable a **reducir la impureza** de los nodos en los árboles. La impureza representa la heterogeneidad dentro de un nodo: lo ideal es que los nodos sean internamente homogéneos (contengan observaciones similares) y estén bien diferenciados entre sí. Cuanto mayor es la

reducción de impureza asociada a una variable, mayor es su importancia para generar divisiones efectivas en el árbol.

Bosque de regresión

En el contexto de regresión, Random Forest se emplea para predecir una variable numérica continua, en este caso, el nivel de éxito de una canción. La predicción final del modelo se obtiene como el promedio de las predicciones generadas por todos los árboles del bosque.

Para evaluar el rendimiento del modelo, se utilizan las siguientes métricas:

- **Error cuadrático medio (Mean of Squared Residuals)**: indica el promedio de los residuos al cuadrado, es decir, la diferencia entre los valores predichos y los reales. Cuanto menor es este valor, mejor es el ajuste del modelo.
- **Porcentaje de varianza explicada (% Var Explained)**: mide qué proporción de la variabilidad de la variable dependiente es explicada por el modelo. Se calcula como:

$$\%Var\ explained\ = 1 - \frac{MSE\ del\ modelo}{Varianza\ total\ de\ la\ variable\ respuesta} \times 100$$

  Un valor más alto de esta métrica indica que el modelo captura mejor la estructura de los datos y tiene mayor capacidad explicativa.

- También se calcula el **Root Mean Squared Error (RMSE)**, ya descrito en el apartado del modelo lineal, como medida complementaria del ajuste del modelo.

Estas medidas permiten **evaluar tanto la precisión de las predicciones como la calidad del ajuste del modelo**, ofreciendo una visión completa de su rendimiento en el análisis del éxito musical.

Bosque de clasificación

Para predecir el **nivel de éxito categórico** (bajo, moderado o alto), se utiliza Random Forest como modelo de clasificación. La clase predicha es aquella que más veces aparece como resultado en los árboles del bosque. Para evaluar el rendimiento del modelo, se utilizan varias métricas:

- ***Accuracy*:** proporción total de predicciones correctas realizadas por el modelo.
- ***Sensitivity* (Sensibilidad o Recall):** capacidad del modelo para identificar correctamente los casos positivos de cada clase.
- ***Specificity* (Especificidad):** capacidad del modelo para identificar correctamente los casos negativos (es decir, no pertenecientes a una clase dada).

Para evitar problemas de ajuste asociados al desequilibrio o diferencia de tamaños muestrales en las clases de éxito, se utiliza la **estratificación por éxito** para crear las muestras de entrenamiento y test, así como las submuestras de los árboles. Este permite realizar un muestreo estratificado, garantizando que cada submuestra respete las proporciones originales de las clases presentes en el conjunto de entrenamiento. Se evita así el sesgo hacia una clase mayoritaria y se mejora la estimación en las clases minoritarias.

Además, se emplea la **matriz de confusión** como herramienta para visualizar el desempeño del modelo, mostrando el número de observaciones correctamente clasificadas y los errores cometidos por clase.

En resumen, el método Random Forest representa una herramienta versátil y eficaz para abordar tanto problemas de regresión como de clasificación dentro del análisis musical. Su capacidad para manejar grandes volúmenes de datos, reducir el riesgo de sobreajuste y ofrecer medidas interpretables de importancia de las variables lo convierte en una técnica especialmente útil en contextos complejos como el presente estudio. Además, la posibilidad de ajustar sus hiperparámetros permite adaptar el modelo a las características particulares de los datos, optimizando así su rendimiento predictivo.

Para llevar a cabo todo el análisis estadístico y computacional del presente estudio, se ha utilizado software especializado que permite implementar de forma eficiente las distintas técnicas metodológicas empleadas.

4.4 Software

El software utilizado en este trabajo ha sido ***R*** (R Core Team, 2023), un lenguaje de programación y entorno de s**oftware libre** ampliamente utilizado para el análisis estadístico, la visualización de datos y el desarrollo de modelos predictivos. Para facilitar

su uso, se empleó ***RStudio*** (Posit, PBC, 2023), un **entorno de desarrollo integrado** (IDE) que mejora la organización del código, la ejecución de scripts y la interpretación de resultados, ofreciendo una interfaz para trabajar con R.

Para el desarrollo del análisis estadístico y los modelos predictivos, se utilizaron diversas **librerías de R**, cada una con funcionalidades específicas que facilitaron la exploración, transformación, visualización y modelado de los datos:

- ***ggplot2*** (Wickham, 2023): utilizada para la creación de gráficos estadísticos de alta calidad. Permite representar de forma clara y personalizada la información mediante la gramática de gráficos.

- ***gridExtra*** (Auguie, 2017): empleada para organizar múltiples gráficos de *ggplot2* en una misma figura, lo que facilita la comparación visual de distintos resultados.

- ***patchwork*** (Pedersen, 2023): usada para combinar varios gráficos generados con *ggplot2* de forma sencilla y flexible, permitiendo crear visualizaciones más completas y ordenadas.

- ***dplyr*** (Wickham et al., 2023): utilizada para la manipulación eficiente de datos, incluyendo tareas como filtrado, selección de columnas, creación de nuevas variables o agrupación de observaciones.

- ***corrplot*** (Wei & Simko, 2021): permite visualizar de forma gráfica las matrices de correlación entre variables, facilitando la interpretación de relaciones lineales.

- ***rpart.plot*** (Milborrow, 2022): facilita la representación visual de los árboles de decisión ajustados con rpart, mostrando las reglas de decisión y los nodos de forma clara.

- ***caTools*** (Tuszynski, 2021): se utiliza para dividir la base de datos en conjuntos de entrenamiento y test, manteniendo las proporciones originales de las clases.

- ***rpart*** (Therneau & Atkinson, 2023): empleada para construir árboles de decisión simples, tanto para regresión como para clasificación.

- ***randomForest*** (Liaw & Wiener, 2002): utilizada para ajustar modelos de Random Forest, tanto de regresión como de clasificación, permitiendo además obtener métricas de importancia de las variables y evaluar el rendimiento del modelo.

- ***caret*** (Kuhn, 2023).: sirve como framework para entrenar y evaluar modelos predictivos. Es un conjunto de funciones que buscan optimizar el proceso de creación de modelos predictivos.

A continuación, se exponen los resultados obtenidos tras aplicar las distintas técnicas estadísticas y de machine learning, valorando su capacidad para explicar y predecir el éxito musical a partir de las variables disponibles.

5. Resultados

En esta sección presentamos todos los resultados obtenidos de los análisis exploratorios y predictivos realizados, y descritos en la sección de Metodología.

5.1 Análisis exploratorio y preprocesado

Abordamos esta sección presentando en primer lugar los resultados del análisis exploratorio univariado, para describir la información disponible, y luego del análisis bivariado, para identificar asociaciones entre las variables disponibles.

5.1.1 Análisis Univariado

Se presentan a continuación los resultados del análisis exploratorio y las decisiones tomadas sobre el preprocesado, en base a este análisis. En primer lugar identificamos los valores faltantes en la base de datos y realizamos la imputación correspondiente. A continuación, presentamos el análisis exploratorio de las variables de éxito relativas a la interacción del público. Después presentamos el descriptivo de las variables predictoras, diferenciadas por su escala de variación: **escala entre 0-1**, **escalas diversas** y **categórica**.

Observamos que algunas variables numéricas, como *Views, Likes, Comments, Stream y Duration*, presentan distribuciones muy asimétricas, con valores muy altos. Por ello, se optó por dividir los valores de *Views* y *Stream* entre 10^9 y los de *Likes* y *Comments* entre 10^7 para facilitar su visualización en los boxplots de la Figura 1, sin alterar la forma general de la distribución.

Cabe destacar que no ha sido necesario escalar las variables numéricas, ya que muchas de ellas ya se encontraban normalizadas en un rango entre 0 y 1, lo que facilita su comparación y análisis conjunto.

Valores faltantes

Durante la revisión inicial de la calidad del conjunto de datos, se identificó la presencia de **2.178 valores faltantes**, distribuidos tanto entre las variables predictoras como en la respuesta. En concreto, las variables afectadas fueron: *Danceability, Energy, Key, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo, Duration, Views, Likes, Comments y Stream.*

Del total de variables analizadas, se observó que las **variables de respuesta** —Views, Likes, Comments y Stream— son las que presentan un volumen mayor de valores faltantes, con 470, 541, 569 y 576 valores, respectivamente.

En cambio, las **variables predictoras** *Danceability, Energy, Key, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo y Duration* presentan únicamente 2 valores faltantes cada una, lo que equivale a una proporción muy baja con respecto al tamaño total del conjunto de datos (20718).

Destacar que la variable ***Key*** contiene la categoría "ND" que indica tonalidades no determinadas. Aunque no se considera un valor faltante formalmente, representa ausencia de información y no ha sido imputada, manteniéndose en el análisis por su posible relevancia.

Dado que se trata de un número reducido de valores perdidos y que no se identificaron patrones sistemáticos de ausencia, se optó por aplicar una **técnica de imputación** simple pero efectiva: reemplazar los valores faltantes por la media de cada variable correspondiente.

Variables de éxito

En la Figura 1 se muestra la distribución de las 4 variables de éxito en la base de datos. Se aprecia en todas ellas, una distribución no negativa y asimétrica a la derecha. El rango de variación es mayor en las variables *Views* y *Stream*.

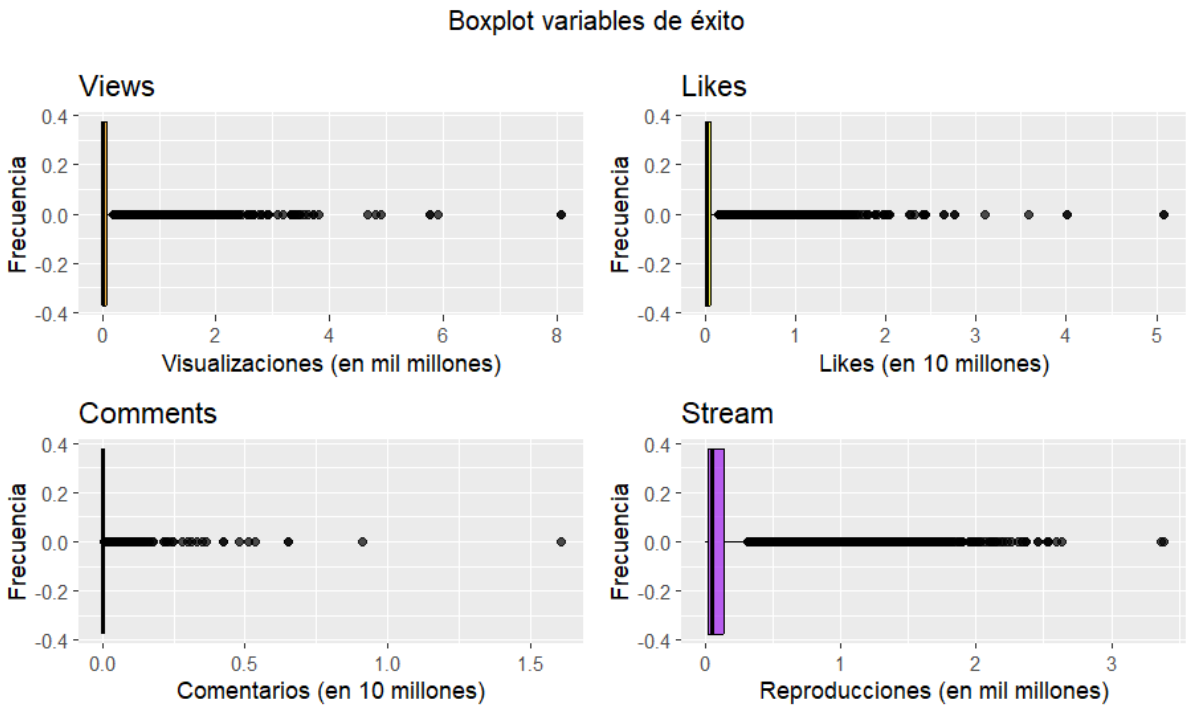

Figura 1. Boxplots de las 4 variables de éxito: Views, Likes, Comments y Stream.

Los diagramas de caja de las variables relacionadas con el éxito de las canciones (*Views, Likes, Comments y Stream*) muestran una distribución altamente asimétrica y concentrada. En todos los casos, aproximadamente el 70 % de los datos se encuentra agrupado en valores muy bajos, cercanos a cero.

Sin embargo, se observa una gran dispersión hacia la derecha del gráfico en forma de numerosos valores atípicos (*outliers*), lo cual indica que existe una minoría de canciones que alcanzan cifras extremadamente altas en estas métricas, y que sin duda dificultará el análisis (al ser pocas y extremas).

Variables predictoras

En la Figura 2 se presentan los boxplots de las variables en escala 0/1.

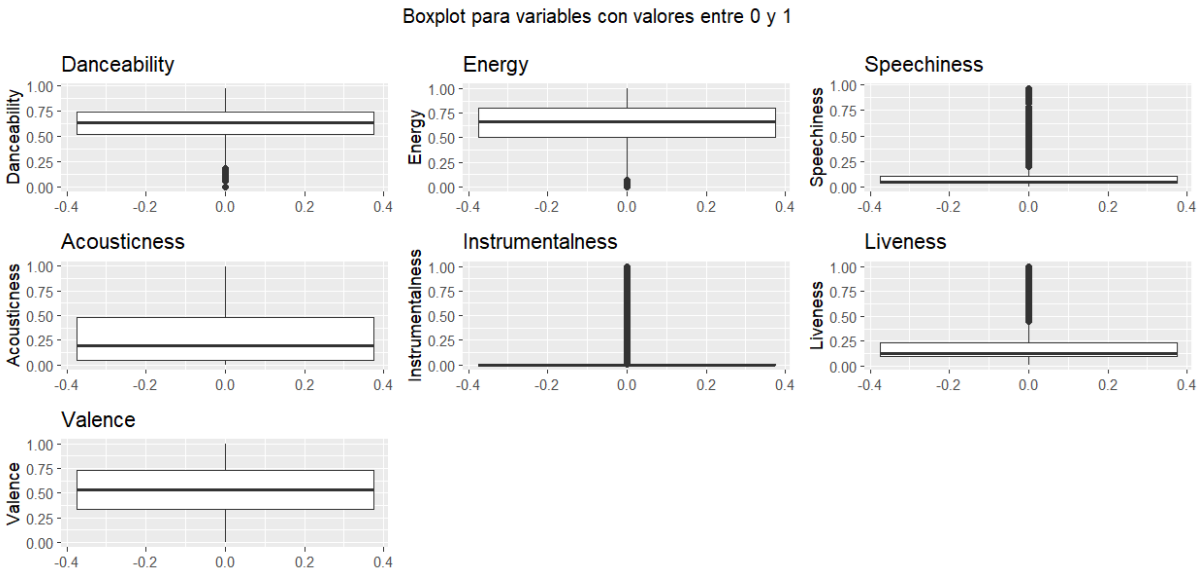

Figura 2. Boxplot variables numéricas entre 0 y 1

Encontramos que *Danceability, Energy y Valence* tienen prácticamente el 75% de los datos por debajo de 0.75, lo que habla de datos distribuídos a lo largo de todo el rango de variación. Además, para estas variables aproximadamente el 50% central de los datos se encuentra en valores centrales de la escala 0-1, lo que implica simetría en sus distribuciones.

Sin embargo, encontramos que las variables *Speechiness, Instrumentalness y Liveness*, tienen su percentil 75 por debajo de 0.25, lo que habla de una gran concentración de registros con valores muy pequeños. Como se observa en la *Figura 2*, el 25% de datos restantes, se distribuyen por valores entre 0.25 y 1, pero no se podría hablar de catalogación como valores atípicos, pues realmente son un volumen importante, y por lo tanto describen la asimetría severa y la gran dispersión de su distribución.

A medio camino encontramos la variable *Accousticness*, que tiene una mediana cerca de 0.3, y un rango intercuartílico entre 0.1 y 0.5, lo que pone de manifiesto también, la asimetría de los datos. El hecho de que el percentil 75 sea 0.48 y el 25% de los datos más altos se repartan hasta el máximo 1, tampoco permite reconocer valores anómalos relevantes.

En resumen, variables como *Danceability, Energy y Valence* presentan distribuciones simétricas y centradas, mientras que variables como *Speechiness, Acousticness e Instrumentalness* presentan asimetría alta, con una proporción importante de

observaciones en el cuartil superior. Esta diversidad sugiere que las canciones del conjunto abarcan una gran variedad de estilos y estructuras acústicas.

La *Figura 3* presenta los histogramas para las variables continuas *Tempo*, *Loudness* y *Key* (representada de forma categórica). Además, se muestra un diagrama de caja (*boxplot*) para *Duration*.

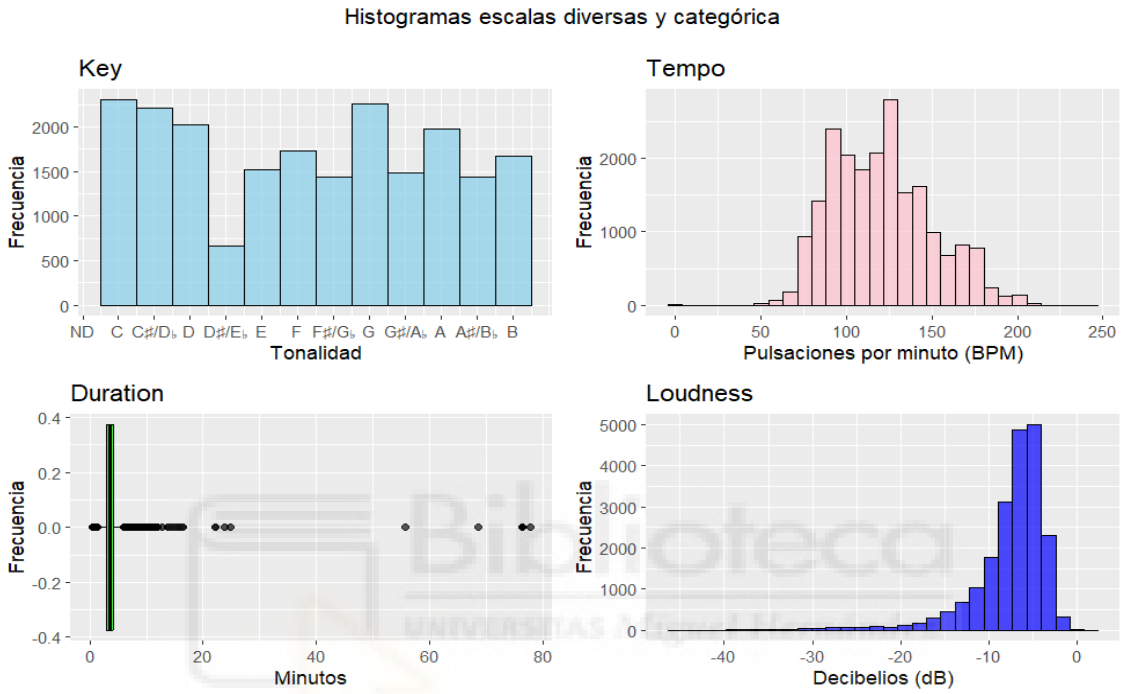

Figura 3. Histogramas y boxplot de varias características registradas: Tonalidad, Pulsaciones por minuto, Minutos (duración) y Decibelios de las canciones en la base de datos.s

Los **histogramas** presentados en la *Figura 3* ofrecen una visión general de la distribución de las variables seleccionadas, a continuación, se presenta una interpretación general de cada variable:

La variable ***Key*** muestra una distribución relativamente uniforme entre las distintas tonalidades musicales, aunque se aprecian ligeros picos de frecuencia en notas como C, F♯ y G. Además, se incluye la categoría "ND", correspondiente a aquellas canciones cuya tonalidad no pudo ser determinada.

Se ha decidido no incluir esta variable en el análisis predictivo, debido principalmente a que ésta es una variable categórica y el enfoque de este estudio se va a centrar en estudiar

la asociación entre el éxito y las variables numéricas disponibles que describen las características de las canciones.

En cuanto a ***Tempo***, la distribución presenta una forma aproximadamente simétrica, con una leve asimetría positiva. La mayoría de las canciones se concentran en un rango de 80 a 140 BPM, lo que sugiere una prevalencia de ritmos moderados a enérgicos.

Respecto a ***Duration***, se observa en el boxplot una distribución claramente asimétrica hacia la derecha, indicando la presencia de una cola larga. Esto revela que, si bien la mayoría de los temas tienen duraciones comprendidas entre los 150 y 500 segundos, existen canciones que exceden considerablemente este rango.

Por último, la variable ***Loudness*** presenta una asimetría negativa. La mayoría de las canciones se sitúan en un rango entre -10 y -5 dB, lo que refleja un volumen medio-alto característico de producciones musicales modernas.

5.1.2 Análisis bivariado

Con el fin de investigar las relaciones entre las variables numéricas predictoras y con las variables de éxito, representamos un correlograma en la *Figura 4*, que presenta las correlaciones en una escala de color en la que los colores fríos (azules) intensos representan correlaciones próximas a 1 y colores cálidos (rojos) intensos las próximas a -1, y por supuesto una gradación en color e intensidad para correlaciones intermedias.

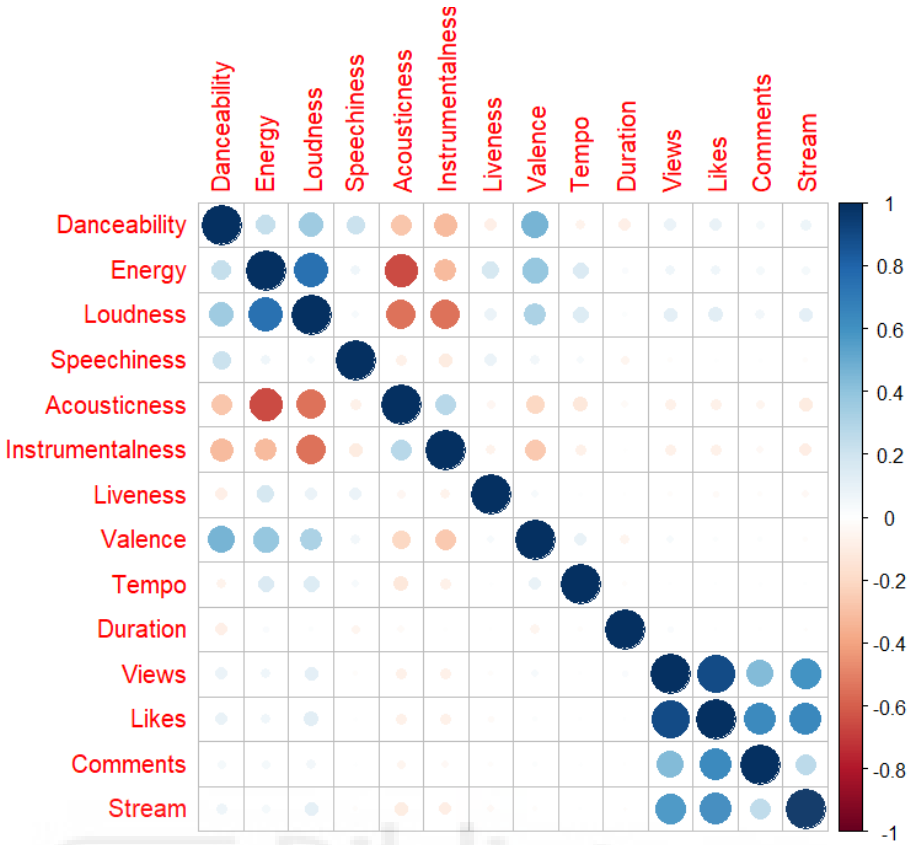

Figura 4. Correlograma de las variables numéricas.

Tras representar la **matriz de correlaciones** del conjunto de variables numéricas *(Figura 4)*, examinamos el coeficiente de correlación entre dos variables y observamos que varias parejas de nuestros datos tienen **correlaciones positivas fuertes**, como son:

- *Views*, *Likes*, *Comments*, y *Streams*: Existe una fuerte correlación positiva entre estas variables, lo que indica que las canciones con más reproducciones tienden a tener más "me gusta", comentarios y, por lo tanto, son más populares. Cabe destacar que la variable *Likes* es la que más correlacionada está con el resto de estas cuatro variables. Estas correlaciones tan altas justifican la búsqueda de un índice, construido con estas cuatro variables, para definir, si es posible, un único 'indicador de éxito'.

- *Loudness* y *Energy*: Las canciones más fuertes suelen ser más enérgicas.

- *Danceability* y *Valence*: Existe una ligera correlación positiva entre la bailabilidad y la valencia (positividad emocional de la canción). Esto podría indicar que las canciones más bailables suelen ser las más positivas.

- *Valence* y *Energy*: Ligera correlación positiva entre la energía y la valencia (positividad) de una canción. Esto nos indica que las canciones con más energía suelen ser las más positivas.

También podemos visualizar que tenemos parejas de variables con **correlaciones negativas**, estas son:

- *Acousticness* y *Loudness*: Las canciones acústicas tienden a ser menos fuertes y viceversa.
- *Acousticness* y *energy*: correlación fuerte negativa, esto implica que una canción con instrumentalidad acústica no tiene porque tener energía.
- *Instumentalness* y *Loudness*: Las canciones sin contenido vocal tienden a ser menos fuertes y viceversa.

En conclusión, las correlaciones entre las variables sonoras y las de éxito son, en general, bastante bajas. Esto sugiere que la capacidad explicativa de las características musicales para predecir el éxito de una canción podría ser limitada, lo que puede dificultar la construcción de modelos predictivos eficientes basados únicamente en estas variables.

5.2 Aprendizaje no supervisado para la reducción de la dimensión

Habiendo explorado diversas técnicas de reducción de dimensionalidad, pasamos ahora a presentar los resultados obtenidos al aplicar el **Análisis de Componentes Principales**.

Realizamos el análisis de componentes principales para calcular un índice de éxito que aglutine la información en las variables de éxito *Views*, *Comments*, *Likes* y *Stream*, y encontramos, como se aprecia en la Tabla 2, que sólo con la primera componente ya explicamos un 69.43% de la variabilidad en los datos, como se aproxima en gran medida al 70% la consideraremos como válida.

	PC1	PC2	PC3	PC4
Standard deviation	1,6665	0,8679	0,6322	0,26423
Proportion of Variance	0,6943	0,1883	0,0999	0,01745
Cumulative Proportion	0,6943	0,8826	0,9826	1,00000

Tabla 2. Resumen de la Varianza Explicada por los Componentes Principales.

Así pues, es razonable asumir que estas variables numéricas de éxito podemos resumirlas en **una única dimensión**, sin perder información relevante, que es la primera componente resultante. A esta nueva variable la denominaremos "**Éxito**" en adelante, que será nuestra v**ariable respuesta a predecir en función del resto**.

Tabla 3. Pesos de las variables de éxito en la primera componente principal.

Variable	Pesos en la PC1
Views	-0.5413415
Likes	-0.5804891
Comments	-0.4053259
Stream	-0.4535336

La nueva variable de éxito se define como una **combinación lineal** de las variables:

$$Exito = -0.54 \times Views - 0.58 \times Likes - 0.41 \times Comments - 0.45 \times Stream$$

En base a los pesos que mostramos en la *Tabla 3*, observamos que los valores más bajos en PC1 (negativos), corresponden a valores más altos en las variables originales, lo que

implica **mayor éxito**. Por tanto, esta componente puede interpretarse como un indicador inverso de éxito: a menor puntuación en PC1, mayor popularidad de la canción.

Entre las variables consideradas, *Likes* y *Views* son las que más influyen en esta dimensión, lo que sugiere que estas métricas son especialmente relevantes para caracterizar el **éxito**.

Una vez obtenido el nuevo indicador de éxito, se analiza su relación con las distintas variables disponibles en la base de datos, con el objetivo de identificar aquellas que podrían actuar como **potenciales predictores**.

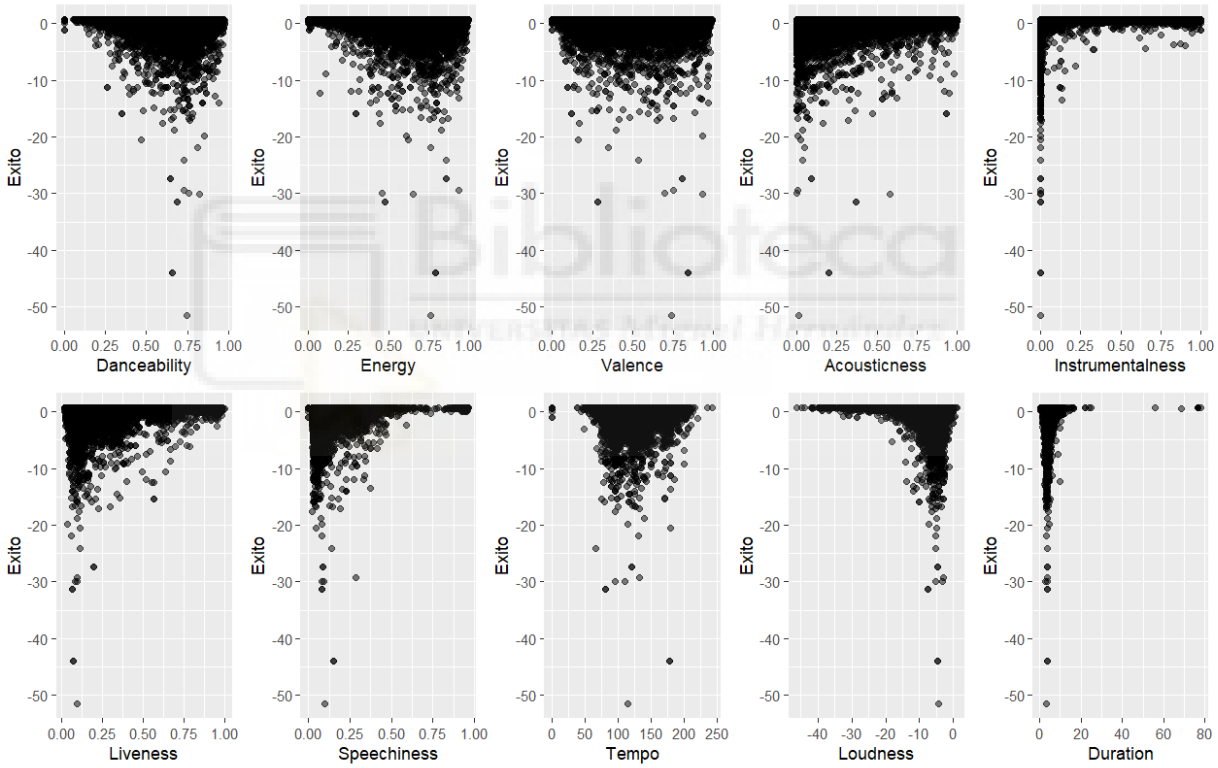

Figura 5. Gráficos de dispersión entre la variable Éxito y las variables predictoras numéricas.

Las observaciones con mayores valores de variables como *Danceability, Energy, Valence y Loudness* tienden a presentar una **mayor dispersión en los niveles de éxito**. Es decir, hay canciones con características muy "altas" en estos atributos que tienen tanto altos como bajos niveles de éxito, reflejando mayor variabilidad.

En cambio, cuando estas variables tienen valores bajos, el éxito tiende a concentrarse en niveles más reducidos (menor dispersión).

Por otro lado, variables como *Instrumentalness*, *Speechiness*, *Acousticness* y *Duration* muestran que la mayoría de los datos están concentrados en los extremos bajos del éxito, lo que indica que estas características, en su mayoría, no están asociadas a canciones particularmente exitosas.

En particular, se evidencia una forma de "triángulo invertido" en muchas variables, indicando que existe un rango de valores centrales en la característica, asociados a un mayor éxito, mientras que valores extremos a la izquierda y derecha suelen estar asociados a éxitos más leves.

El objetivo a continuación es ajustar diversos modelos para predecir esta variable de éxito.

5.3 Aprendizaje supervisado para la predicción del éxito

Con el objetivo de estudiar la capacidad predictiva de las variables disponibles en la base de datos sobre el nivel de éxito de las canciones, se recurre a técnicas de aprendizaje supervisado.

Para facilitar la interpretación del indicador de éxito derivado de la **primera componente principal**, se optó por trabajar con su versión simétrica, de modo que los pesos pasan a ser positivos y, por tanto, los valores más altos se asocian a un mayor nivel de éxito. Con el objetivo de posibilitar una transformación logarítmica (que escala a magnitudes menores la variable indicadora), se procedió a reescalar la variable ***Exito*** para que tomara valores estrictamente positivos, con la transformación $LogExito = log(max(Exito) - Exito + 1)$, que cumple:

$$min(Exito) \Leftarrow Exito \Leftarrow max\,(Exito) \rightarrow 1 \Leftarrow max(Exito) - Exito + 1 \Leftarrow max(Exito) - min(Exito) + 1$$

Esta nueva variable ***LogExito*** será utilizada como **variable respuesta en el análisis.**

Para el ajuste de estos modelos, las variables predictoras *Tempo*, *Duration* y *Loudness* fueron previamente **estandarizadas** (restando media y dividiendo por la desviación típica), ya que se trata de variables numéricas que pueden presentar escalas diferentes.

A continuación, se presentan los distintos modelos supervisados que se aplicaron, comenzando por el modelo lineal.

5.3.1 Modelo lineal

Inicialmente se ajustó un **modelo lineal basal** que incluía todas las variables disponibles como predictoras del éxito. Este modelo completo sirvió como punto de partida para identificar qué variables aportan significativamente a la explicación de la variabilidad en la respuesta LogExito.

Tras aplicar un procedimiento de **selección automática de variables** mediante la función *step* en R, se obtuvo un modelo lineal final que incluye las siguientes variables predictoras del índice de éxito: *Danceability*, *Energy*, *Valence*, *Loudness*, *Speechiness*, *Acousticness*, *Instrumentalness*, *Liveness* y *Duration*.

El modelo resultante es **estadísticamente significativo** (*F = 121.8, gl = 9 y 20.708, pvalor< 2.2e-16*), sin embargo, su capacidad explicativa es muy limitada: el coeficiente de determinación ajustado (**R^2 ajustado**) es de apenas 0.04987, lo que indica que el modelo sólo explica alrededor del 5% de la variabilidad observada en la variable de éxito. Además, el error cuadrático medio de la raíz (RMSE), que cuantifica la diferencia promedio entre los valores predichos y observados, se ha estimado en **0.4516**, lo que refuerza la limitada capacidad predictiva del modelo.

Este bajo poder predictivo era esperable, ya que los análisis exploratorios previos ya mostraban que no existía una relación lineal clara entre las variables independientes y el éxito. Además, se han detectado violaciones de supuestos fundamentales (*Figura 6*) del modelo lineal, como la **linealidad y la normalidad de los residuos**, lo que invalida a este modelo como herramienta predictiva.

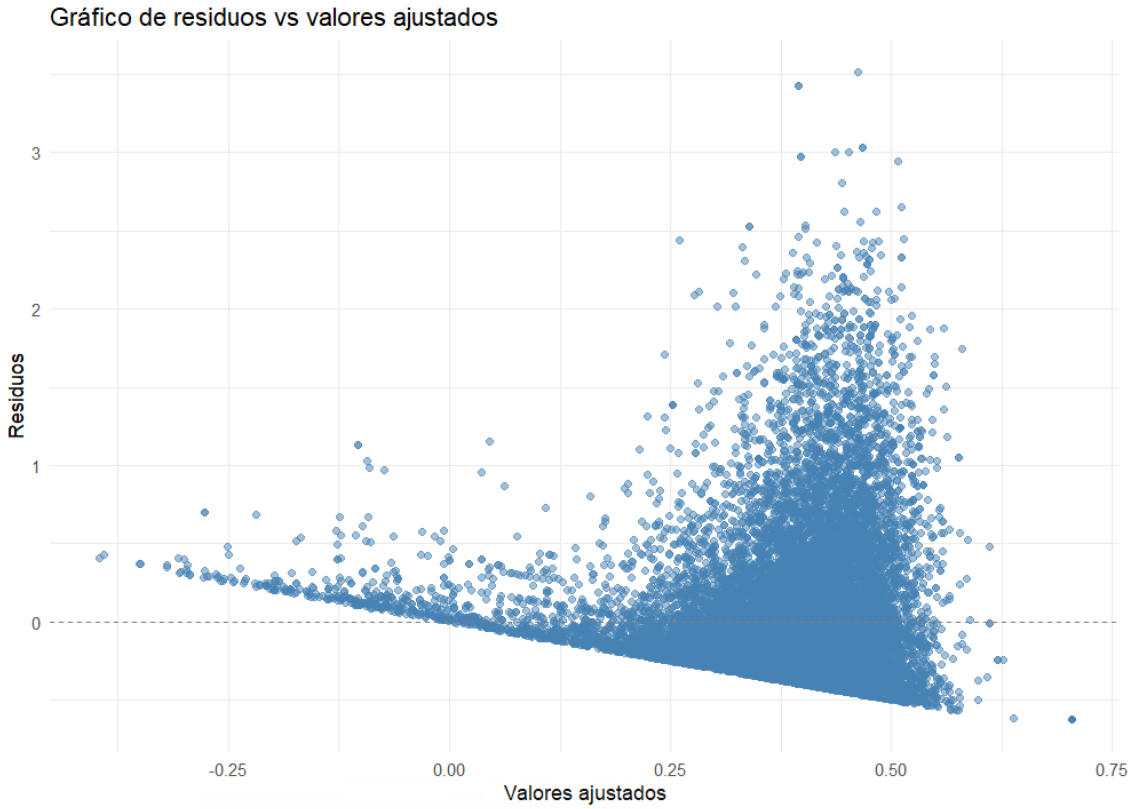

Figura 6. Gráfico de residuos frente a valores ajustados del modelo lineal

En este caso, se observa una clara forma de abanico (patrón de dispersión creciente), lo que indica **heterocedasticidad**: los residuos aumentan a medida que lo hacen los valores ajustados. Esto sugiere que el modelo no se ajusta de forma uniforme a lo largo del rango de valores predichos y, por tanto, viola uno de los supuestos fundamentales del modelo lineal.

La *Figura 7* muestra los **coeficientes estimados** del modelo lineal final obtenido. En el eje horizontal se representan los valores estimados de los coeficientes, mientras que en el eje vertical se listan las variables incluidas en el modelo.

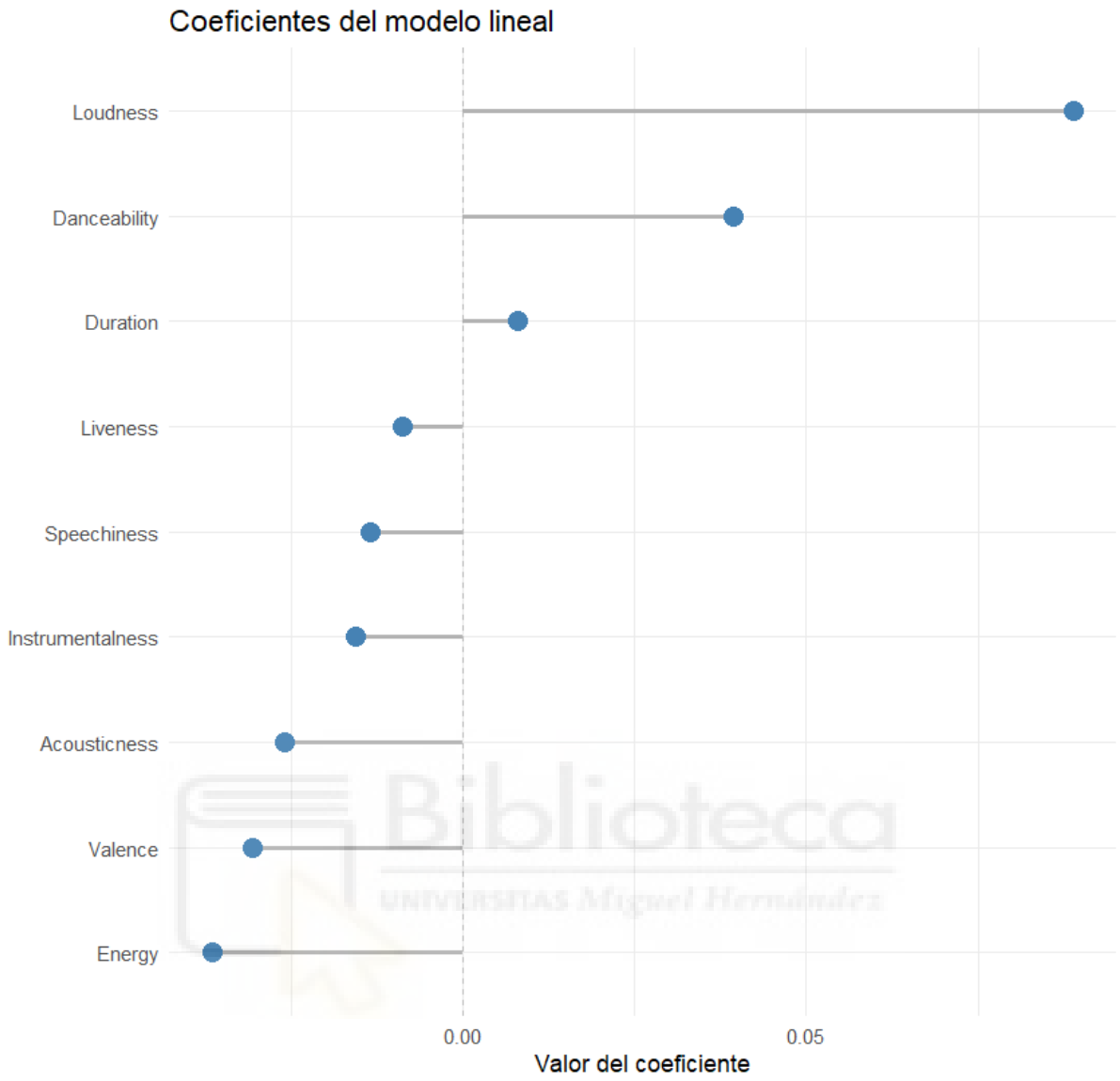

Figura 7. Coeficientes del modelo lineal ajustado con las variables seleccionadas por el procedimiento automático adelante-atrás basado en el AIC.

El gráfico de coeficientes del modelo lineal permite identificar cómo influyen las distintas variables en el nivel de éxito de una canción. Las variables con **coeficientes positivos**, como *Loudness* (0.08897) y *Danceability* (0.03946), se asocian a un mayor éxito: es decir, canciones más fuertes y bailables tienden a alcanzar mejores resultados.

Por el contrario, variables como *Energy* (–0.03653), *Valence* (–0.03097), *Acousticness* (–0.02579), *Instrumentalness* (–0.01559) y *Speechiness* (–0.01359) presentan **coeficientes negativos**, lo que indica que un mayor valor en estas características se relaciona con un menor nivel de éxito.

Finalmente, variables como *Liveness* (–0.0088) y *Duration* (0.00813), a pesar de estar incluidas en el modelo, muestran un peso reducido, lo que sugiere que su influencia sobre el éxito es menos relevante.

La variable *Tempo* fue descartada durante el proceso de selección, al no aportar una mejora significativa al ajuste del modelo. En conjunto, estos resultados refuerzan la idea de que, si bien ciertos rasgos musicales tienen una relación con el éxito, esta es limitada y no estrictamente lineal.

5.3.2 Árboles de decisión

A continuación, se aplicaron **árboles de decisión** sobre las variables predictoras musicales con el objetivo de analizar el éxito de las canciones. Para ello, se realizó previamente una **partición del conjunto de datos**, tomando como referencia la variable de éxito, **LogExito**, dividiéndolo en un 70% para entrenamiento y un 30% para prueba. Esta división permite ajustar los modelos con los datos de entrenamiento y evaluar su rendimiento sobre datos no vistos, garantizando así una validación adecuada.

En primer lugar, se construyó un árbol de regresión utilizando como variable respuesta LogExito. Posteriormente, dicha variable fue categorizada en tres niveles para ajustar un árbol de clasificación:

- **Éxito bajo**: valores entre 0 y 1
- **Éxito moderado**: valores entre 1 y 2
- **Éxito alto**: valores mayores a 2

Esta estrategia, de utilizar la variable de éxito en su versión numérica y en una versión categórica, permitió comparar ambos enfoques en términos predictivos y, además, identificar las variables musicales que más influyen en el nivel de éxito.

Árbol de regresión

A continuación vamos a ver un árbol de regresión que ofrece una representación visual y simplificada de cómo distintas características musicales influyen en el éxito:

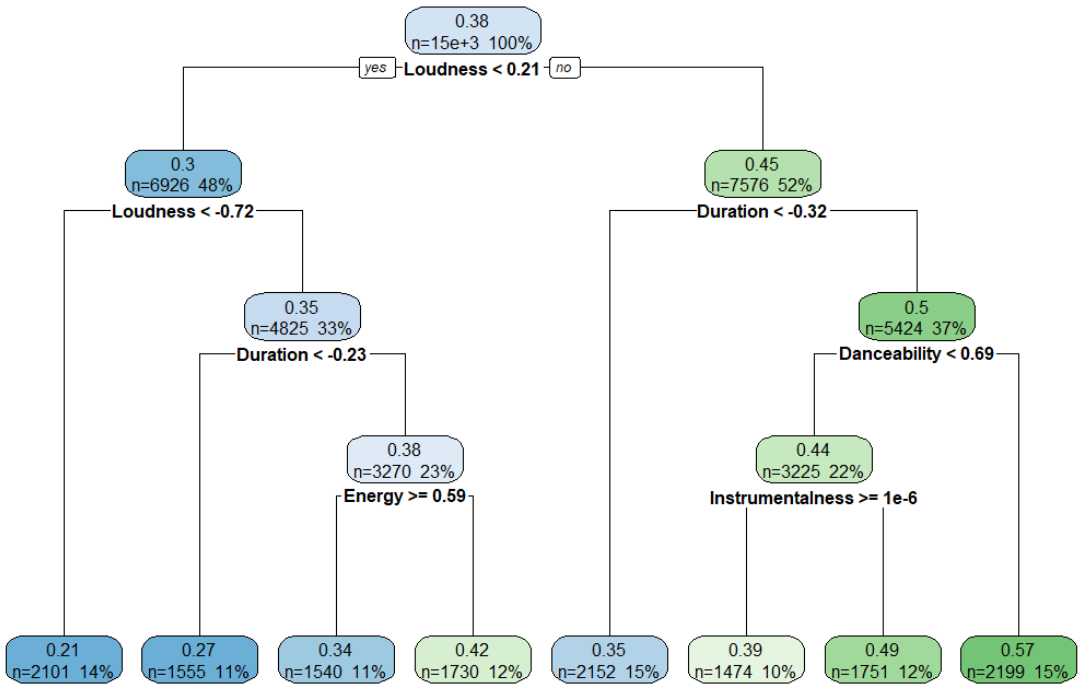

Figura 7. Árbol de regresión para la predicción del éxito musical en función de características acústicas.

Se construyó un árbol de regresión para predecir el LogExito a partir de las variables musicales. Para **evitar sobreajuste y favorecer la interpretabilidad**, se estableció una profundidad máxima (**maxdepth**) de 4, valor seleccionado tras observar que profundidades mayores no reducían significativamente el error y aumentaban la complejidad del árbol dificultando su interpretación. No obstante, se permitió que el proceso de poda considerase hasta profundidad 8. Los parámetros mínimos para dividir un nodo (**minsplit**) y para crear nodos terminales (**minbucket**) se fijaron en el 10% del tamaño total del conjunto de entrenamiento, lo que asegura que las particiones tengan una representatividad adecuada.

En cuanto al parámetro de complejidad (**cp**), se seleccionó el valor más bajo posible (cp = 0.001), ya que fue el que **minimizó el error de validación cruzada** (xerror). Este ajuste reflejó un mejor rendimiento en términos de capacidad predictiva generalizada, es decir, una menor tasa de error esperada en datos nuevos, en comparación con el valor por defecto (cp = 0.01), que resultó en un mayor error de validación. Por tanto, se optó por el cp más bajo sin sacrificar estabilidad del modelo, permitiendo capturar más estructura relevante en los datos.

El modelo resultante tiene un **RMSE** de **0.4507**, lo cual indica un error moderado en la predicción del éxito musical. Este valor es muy similar al obtenido con el modelo lineal (RMSE = 0.4516), lo que sugiere que, pese a utilizar enfoques distintos, ambos modelos presentan un rendimiento comparable en términos de precisión. No obstante, el árbol de regresión presenta una ligera mejora, aunque insuficiente como para considerarlo significativamente superior.

El árbol comienza dividiendo según *Loudness*, seguido por variables como *Duration*, *Danceability*, *Energy* e *Instrumentalness*. Estas divisiones reflejan los umbrales que permiten segmentar las canciones según su nivel de éxito predicho. Se observa que los nodos terminales con valores más altos de éxito están asociados a canciones con mayor *Danceability* e *Instrumentalness*, mientras que los valores más bajos se concentran en ramas con niveles reducidos de *Loudness* y *Duration*. Esto sugiere que ciertos patrones musicales pueden tener relación con la popularidad, aunque de forma no lineal.

Árbol de clasificación

Se construyó a continuación un árbol de clasificación con el objetivo de predecir el **nivel de éxito categorizado** de las canciones a partir de sus características medidas. La *Figura 8* muestra la estructura del árbol que proporciona el modelo, basada en los valores de las distintas variables musicales.

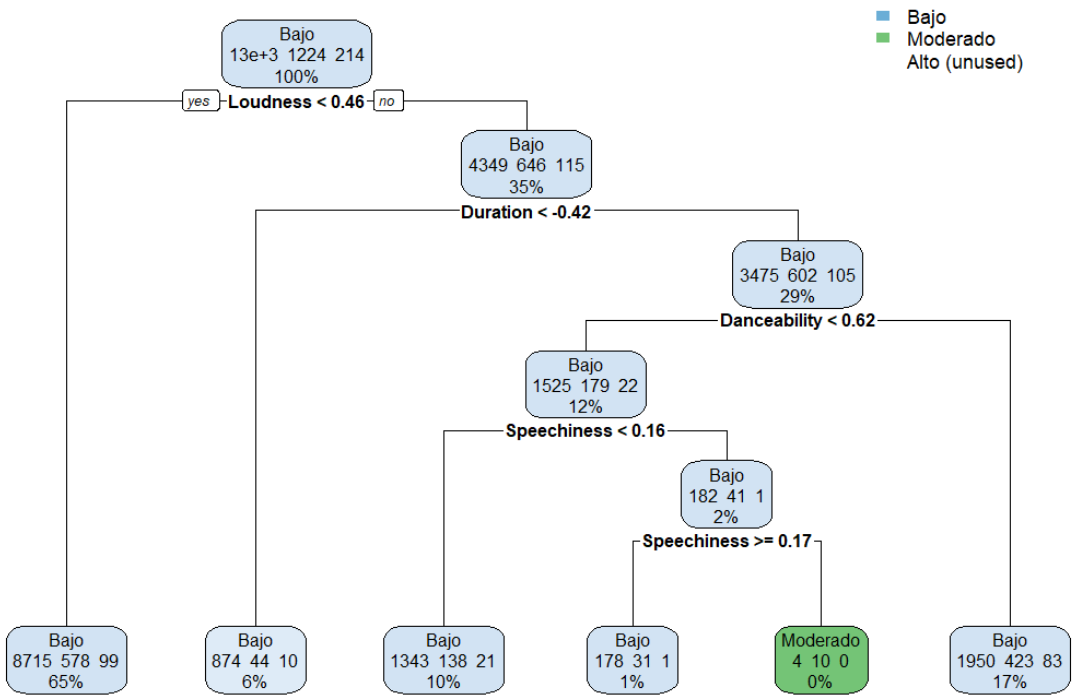

Figura 8. Árbol de clasificación para la predicción del nivel de éxito categórico de las canciones a partir de variables musicales.

Para la construcción del árbol se fijaron los siguientes parámetros: una **profundidad máxima** (maxdepth) de 5, un **mínimo de observaciones** para dividir un nodo (minsplit) igual a 20, y un **parámetro de complejidad** (cp) de 0.008. Este valor fue elegido porque minimiza el error, mostrando mejores resultados que otros valores como cp = 0.001.

En el gráfico, cada nodo terminal presenta el **valor de predicción dominante** (éxito bajo, moderado o alto), junto con la proporción de observaciones que pertenecen a cada categoría.

El árbol parte de la variable *Loudness*, que resulta ser la primera condición de partición. Si el valor de *Loudness* es menor que 0.46, el modelo predice directamente la clase “bajo” con un alto grado de certeza (65% de los datos se encuentran en esta rama). En el caso contrario, el modelo realiza particiones adicionales utilizando las variables *Duration*, *Danceability* y *Speechiness*, identificadas como relevantes para diferenciar entre niveles de éxito.

Aunque la mayoría de las ramas terminan en la categoría **"bajo"**, se observa que en algunas rutas específicas, como aquellas con valores intermedios de *Speechiness*, se consigue una cierta discriminación hacia la clase **"moderado"**, aunque sigue siendo poco frecuente. Por otro lado, la clase **"alto"** no aparece representada en los nodos terminales del árbol podado, lo que indica que no existen suficientes patrones consistentes en los datos que permitan identificarla con un mínimo de fiabilidad dentro de la estructura del modelo.

Esto refleja el **desequilibrio existente entre clases**, siendo "bajo" la categoría claramente mayoritaria en los datos, lo cual limita la capacidad del árbol para detectar patrones en las clases menos representadas.

5.3.3 Random Forest

Con el objetivo de evaluar la capacidad predictiva de los modelos Random Forest, se construyeron dos enfoques: uno de regresión, orientado a predecir una variable continua

de éxito, y otro de clasificación, que utiliza una versión categórica del éxito. Ambos modelos se ajustaron sobre los mismos predictores musicales y emplearon el conjunto de entrenamiento derivado de la partición de los datos (70% entrenamiento, 30% prueba).

Modelo regresión

Con el objetivo de optimizar el rendimiento del modelo de regresión mediante Random Forest, se evaluaron tres configuraciones distintas del parámetro ***mtry***, correspondiente al número de variables seleccionadas aleatoriamente en cada división del árbol. Se probaron los valores de ***mtry* = 3, *mtry* = 4 y *mtry* = 5**, manteniendo constantes el resto de parámetros (ntree = 500).

Tabla 4. Comparación del rendimiento del modelo Random Forest de regresión con distintos valores de mtry (3, 4 y 5). Se muestran el error cuadrático medio (Mean of Squared Residuals) y el porcentaje de varianza explicada.

	Modelo *mtry* = 3	Modelo *mtry* = 4	Modelo *mtry* = 5
Mean of squared residuals	0.1707164	0.1720466	0.1721477
% Var explained	20.4	19.78	19.74

Los resultados obtenidos se resumen en la *Tabla 4*, donde se comparan los valores del **error cuadrático medio** (Mean of Squared Residuals) y el **porcentaje de varianza explicada** por cada modelo. El modelo con ***mtry* = 3** mostró el mejor rendimiento, al presentar el menor error (0.1707) y el mayor porcentaje de varianza explicada (20.4 %). Además, se calculó el **RMSE** para este modelo, obteniéndose un valor de **0.4142**, lo que respalda su rendimiento relativamente superior frente a los demás modelos evaluados, aunque con un margen de error aún moderado.

Por tanto, se selecciona este modelo como el óptimo para el análisis. A partir de él se generó el gráfico de importancia de las variables predictoras (*Figura 9*), el cual permite

identificar los atributos musicales que más influyen en la predicción del éxito de una canción.

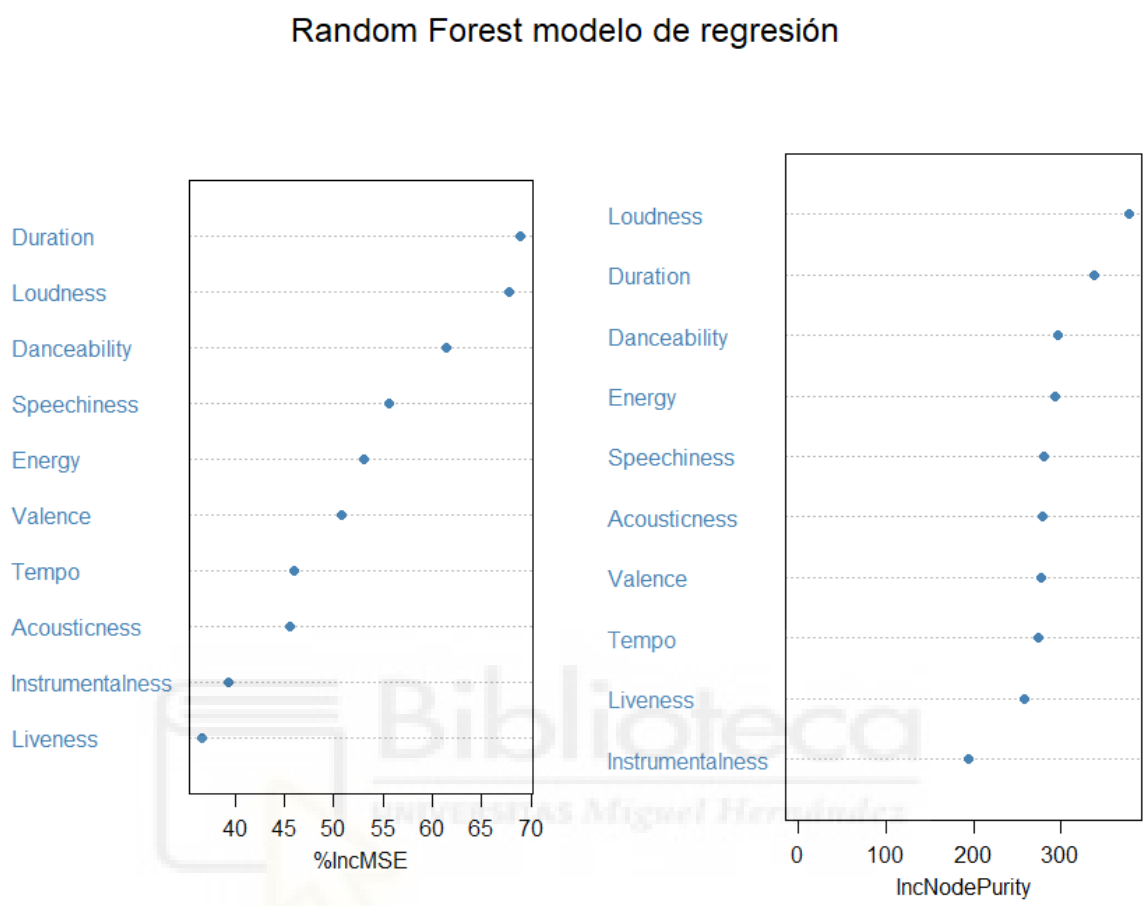

Figura 9. Importancia de las variables en el modelo Random Forest de regresión (Basado en %IncMSE e IncNodePurity)

La Figura 9 muestra la importancia relativa de cada variable predictora en el modelo Random Forest de regresión seleccionado *(mtry = 3)*. En la gráfica de la izquierda, el eje horizontal representa el **incremento porcentual del error de predicción** (%IncMSE) al eliminar cada variable: valores más altos indican mayor influencia en la predicción. En la derecha, la métrica IncNodePurity refleja la **contribución de cada variable a la reducción de impureza** en los nodos del árbol.

Ambos criterios coinciden en señalar que *Duration* y *Loudness* son las variables más importantes para explicar el nivel de éxito de una canción, seguidas por *Danceability* y *Energy*. Por el contrario, variables como *Liveness* e *Instrumentalness* tienen un impacto mucho menor en el modelo.

Modelo clasificación

Tabla 5. Métricas de rendimiento para el modelo Random Forest de clasificación con distintos valores de mtry. Se muestran los valores de sensibilidad, especificidad y precisión global (accuracy) para cada categoría de éxito (bajo, moderado y alto).

	Modelo *mtry* = 2			Modelo *mtry* = 3			Modelo *mtry* = 4			Modelo *mtry* = 5		
	Bajo	Moderado	Alto	Bajo	Moderado	Alto	Bajo	Moderado	Alto	Bajo	Moderado	Alto
Sensitivity	0,9984	0,19439	0,231707	0,998	0,19626	0,231707	0,9979	0,19439	0,231707	0,9977	0,19626	0,231707
Specificity	0,1994	0,99859	0,999837	0,201	0,99824	0,999837	0,1994	0,99806	0,999837	0,201	0,99789	0,999837
Accuracy	0,9191			0,9189			0,9186			0,9186		

Para el análisis del éxito musical categórico (clasificado en bajo, moderado y alto), se aplicó un modelo de Random Forest de clasificación. Se probaron distintos valores del parámetro *mtry* (2, 3, 4 y 5), con el fin de identificar la combinación que proporcionara el mejor rendimiento del modelo. Los resultados se resumen en la *Tabla 5*, donde se comparan las métricas de *accuracy*, *sensitivity* y *specificity* para cada una de las clases.

En general, los modelos con *mtry* = 2 y *mtry* = 3 obtienen el mejor rendimiento, con una *accuracy* del 91.91% y 91.89%, respectivamente. No obstante, observando las métricas de *sensitivity* y *specificity* por clase, el modelo con *mtry* = 2 destaca por lograr valores de sensibilidad ligeramente superiores en las clases minoritarias (moderado y alto), sin sacrificar precisión en la clase mayoritaria (bajo).

Por tanto, se selecciona el modelo con ***mtry* = 2** como el modelo óptimo para la clasificación. A partir de este modelo, se construyó la **matriz de confusión** para visualizar el rendimiento por clase (*Figura 10*), y se generó el **gráfico de importancia de las variables predictoras** (*Figura 11*), con el objetivo de identificar qué características musicales influyen más en la clasificación del éxito.

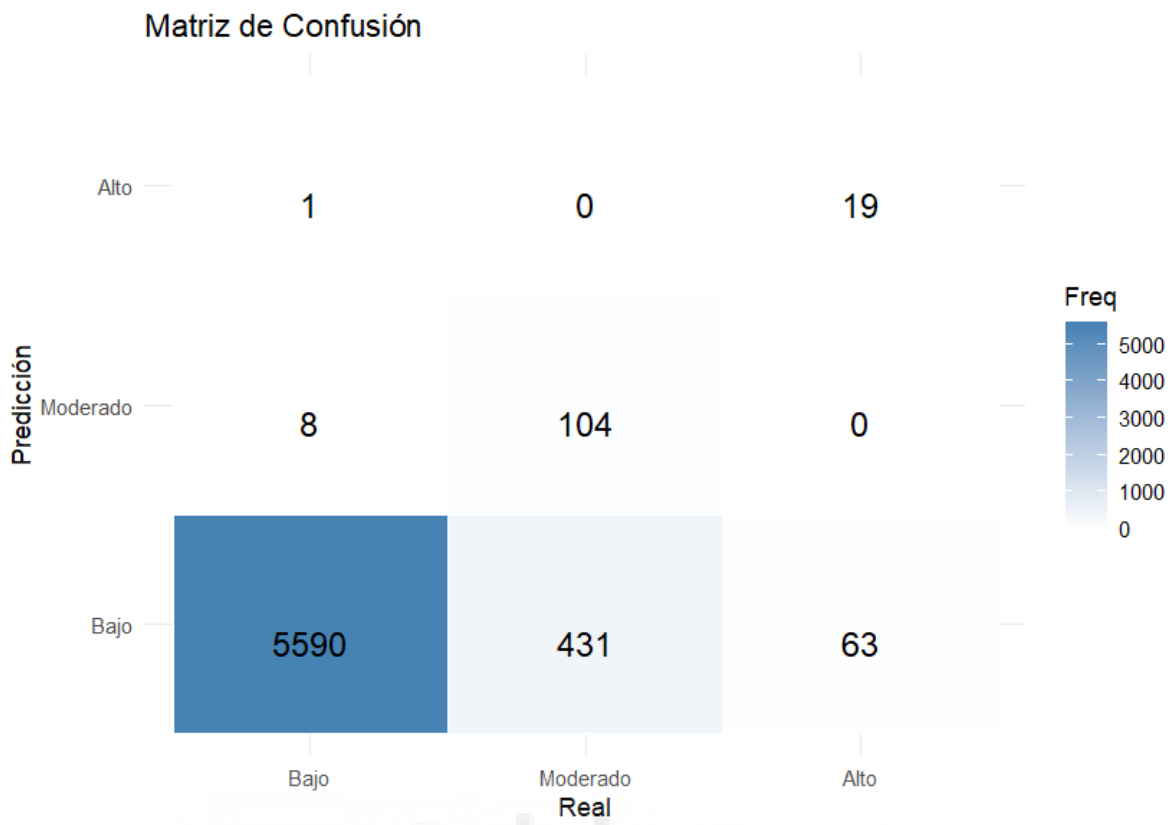

Figura 10. Matriz de confusión del modelo Random Forest de clasificación con mtry = 2

La matriz de confusión muestra que el modelo clasifica muy bien la clase "bajo", con 5590 aciertos, pero tiende a confundir muchas canciones de éxito "moderado" (431) y algunas de "alto" (63) como "bajo". La clase "moderado" presenta menor sensibilidad, con solo 104 aciertos y gran parte de sus casos mal clasificados como "bajo". En cambio, la clase "alto" se predice correctamente en 19 casos, con solo un error. En conjunto, el modelo funciona bien para la clase mayoritaria, pero tiene dificultades para identificar correctamente las clases menos representadas.

A continuación, el gráfico de importancia de variables (*Figura 11*) muestra que, según el *MeanDecreaseAccuracy*, las variables ***Energy***, ***Danceability***, ***Speechiness*** y ***Valence*** **son las que más afectan a la precisión del modelo**. Esta métrica evalúa cuánto disminuye la exactitud del modelo cuando se desordena aleatoriamente una variable. Por otro lado, según el *MeanDecreaseGini*, destacan ***Loudness***, ***Duration*** y ***Tempo*** **como las más relevantes para la partición de los datos**. Esta métrica mide cuánto contribuye cada variable a reducir la impureza de los nodos en el árbol, es decir, busca que los datos dentro de cada nodo sean lo más homogéneos posible. Esto indica que el modelo basa sus

decisiones principalmente en características relacionadas con la **energía**, **ritmo** y **volumen** de las canciones. Sin embargo, como se observó en la matriz de confusión, estas variables no son suficientes para clasificar correctamente las clases menos representadas, especialmente "Moderado" y "Alto".

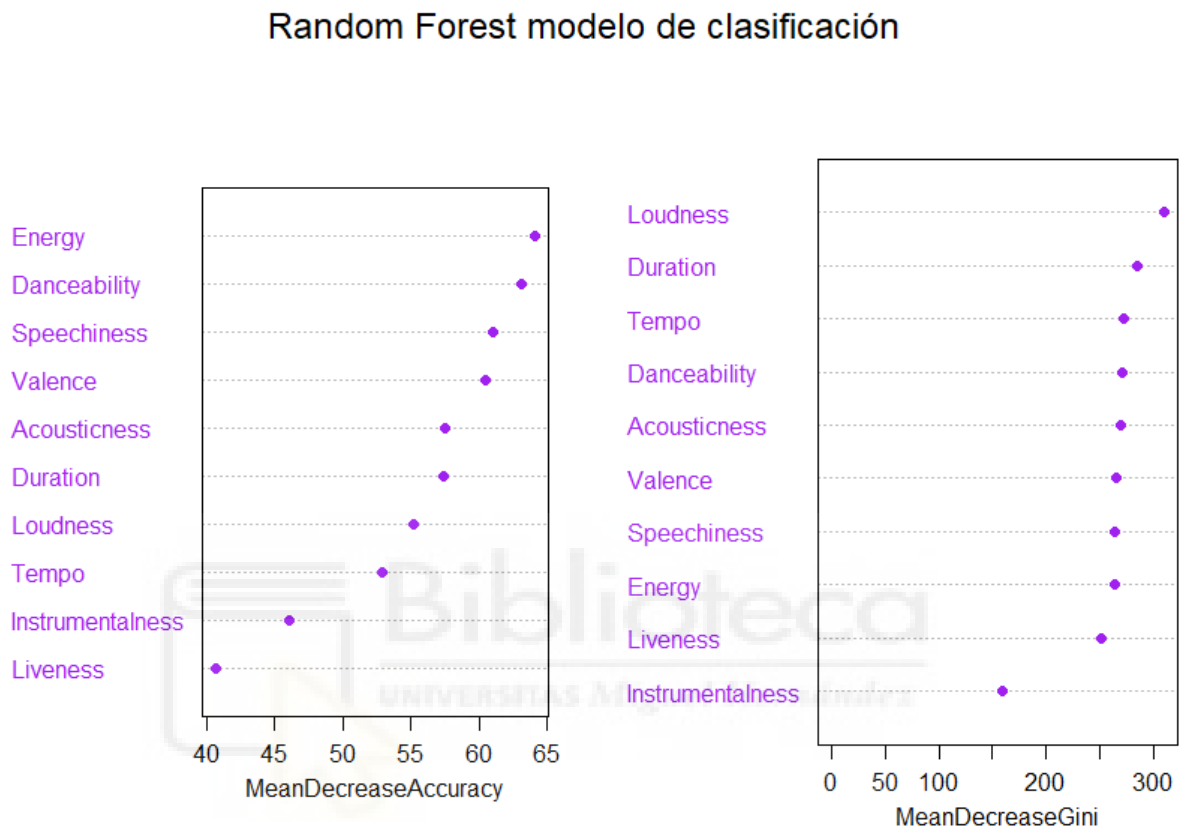

Figura 11. Importancia de las variables en el modelo Random Forest de clasificación (Basado en MeanDecreaseAccuracy y MeanDecreaseGini)

Los resultados obtenidos a partir de los modelos de Random Forest reflejan un **rendimiento moderado** en ambas aproximaciones. El modelo de regresión explicó aproximadamente un 20,4% de la varianza del éxito musical, lo que indica que, aunque el modelo captura parte de la complejidad del fenómeno, aún existe margen de mejora. En cambio, el modelo de clasificación mostró una alta precisión global, pero con limitaciones en la detección de las clases menos representadas (éxito moderado y alto), a consecuencia del grave desequilibrio entre clases. Aun así, ambos modelos identificaron de forma consistente las variables musicales más influyentes, como *Danceability*, *Energy* y *Speechiness*.

6. Conclusiones

Los objetivos planteados al inicio del trabajo se han cumplido de forma progresiva a través del análisis exploratorio, la reducción de dimensión y la aplicación de distintos modelos de aprendizaje supervisado. En primer lugar se construyó un **único indicador de éxito** para cada canción, combinando variables bastante correlacionadas entre ellas mediante la aplicación de PCA (Análisis de Componentes Principales). Este indicador nos ha permitido sintetizar de forma eficiente la información de la interacción del público con cada tema musical.

Posteriormente, se abordó la **predicción del éxito** musical empleando modelos de regresión y clasificación.

Tabla 6. Comparativa de rendimiento e importancia de variables entre los distintos modelos predictivos.

Modelos de regresión		
	RMSE	Variables más influyentes en el éxito
Modelo lineal	0.4515593	*Loudness* y *Danceability*
Árbol de regresión	0.4507146	*Danceability* y *Instrumentalness*
Random Forest	0.4141518	*Duration* , *Loudness* y *Danceability*
Modelos de clasificación		
	Variables más influyentes en el éxito	
Árbol de clasificación	*Duration*, *Danceability* y *Speechiness*	
Random Forest	*Danceability* y *Energy*	

Esta tabla resume los errores (RMSE) obtenidos en los modelos de regresión y de las variables predictoras musicales más influyentes en la predicción del éxito, tanto para modelos con respuesta continua como categórica.

Como puede observarse, el modelo de **Random Forest** aplicado a regresión es el que presenta un **mejor ajuste**, con un RMSE de 0.4142, sensiblemente inferior al de los modelos lineal (0.4516) y de árbol de regresión (0.4507). Esto confirma su mayor capacidad para capturar relaciones no lineales en los datos.

En el **modelo lineal** se observaron asociaciones relevantes entre el éxito y variables como *Loudness* y *Danceability*, mientras que otras como *Liveness* o *Duration* mostraron menor influencia. Además, se descartaron variables sin aportación predictiva significativa, como *Tempo*. No obstante, **este modelo presenta una capacidad predictiva limitada**, por lo que debemos destacar su **escaso poder para predecir el éxito musical,** no supera el diagnóstico.

Los **árboles de decisión** ofrecieron una representación visual clara de las reglas de partición, facilitando la interpretación del proceso predictivo. En estos modelos, se observó que la mayoría de las canciones **tienden a clasificarse como de bajo éxito**, lo que evidencia un desequilibrio en la distribución de la variable respuesta.

Por su parte, los **modelos *Random Forest***, tanto de regresión como de clasificación, aportaron mayor robustez y fiabilidad. En ambos casos, se identificaron consistentemente como variables clave *Loudness* y *Danceability*, lo que sugiere que **canciones con mayor volumen y ritmo bailable presentan más probabilidades de alcanzar éxito.** Los indicadores de importancia de variables respaldaron estos hallazgos de forma sólida.

En conjunto, el análisis muestra que el uso de técnicas de aprendizaje automático aplicadas a características musicales permite construir modelos predictivos con capacidad interpretativa y rendimiento competitivo. Esta aproximación constituye una base sólida para futuros desarrollos, como **sistemas de recomendación musical basados en contenido**.

7. Referencias

- R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

- Posit, PBC. (2023). *RStudio* (versión 2023.06.1) [Entorno de desarrollo de software]. https://posit.co/download/rstudio-desktop/
- Pal, A. (s.f.). *EDA on Spotify and YouTube Dataset* [Notebook]. Kaggle. https://www.kaggle.com/code/apolynomialcurve/eda-on-spotify-and-youtube-dataset
- Zumpano F. (s.f.). *Clustering and regression + PCA* [Notebook]. Kaggle. https://www.kaggle.com/code/chiccoqvc/clustering-and-regression-pca
- Decherisey, H. (s.f.). *RandomForestRegressor - Basic* [Notebook]. Kaggle. https://www.kaggle.com/code/huguesdecherisey/randomforestregressor-basic
- Pilgaonkar, P. (s.f.). *Old Songs vs New* [Notebook]. Kaggle. https://www.kaggle.com/code/prasadpil/old-songs-vs-new/notebook
- Auguie, B. (2017). *gridExtra: Miscellaneous functions for "grid" graphics* (versión 2.3) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=gridExtra
- Kuhn, M. (2023). *caret: Classification and regression training* (versión 6.0-94) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=caret
- Liaw, A., & Wiener, M. (2002). *Classification and regression by randomForest*. *R News, 2*(3), 18–22. https://CRAN.R-project.org/package=randomForest
- Milborrow, S. (2022). *rpart.plot: Plot 'rpart' models* (versión 3.1.1) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=rpart.plot
- Pedersen, T. L. (2023). *patchwork: The composer of plots* (versión 1.1.3) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=patchwork
- Therneau, T., & Atkinson, B. (2023). *rpart: Recursive partitioning and regression trees* (versión 4.1.23) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=rpart
- Tuszynski, J. (2021). *caTools: Tools: Moving window statistics, GIF, Base64, ROC AUC, etc.* (versión 1.18.2) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=caTools
- Wei, T., & Simko, V. (2021). *corrplot: Visualization of a correlation matrix* (versión 0.92) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=corrplot
- Wickham, H. (2023). *ggplot2: Create elegant data visualisations using the grammar of graphics* (versión 3.4.4) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=ggplot2

- Wickham, H., François, R., Henry, L., & Müller, K. (2023). *dplyr: A grammar of data manipulation* (versión 1.1.3) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=dplyr
- Akaike, H. (1974). *A new look at the statistical model identification*. IEEE Transactions on Automatic Control, 19(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2nd ed.). Springer.