ÉXITO EN SPOTIFY

PROYECTO FINAL DATA SCIENCE
CODER HOUSE - COMISIÓN 42390

MELISA CAPUTO

AGENDA

1. MOTIVACIÓN Y AUDIENCIA
2. RESUMEN DE METADATA
3. PROBLEMA A RESOLVER
4. EDA + VISUALIZACIONES
5. FEATURE ENGINEERING
6. SELECCIÓN DE MODELOS

MOTIVACIÓN Y AUDIENCIA

En la era digital, la música se volvió más accesible que nunca gracias a plataformas como Spotify. Detrás de la aparente simplicidad de escuchar una canción en línea, se esconde un mundo de datos y características musicales que influyen en la cantidad de reproducciones.

El objetivo de este proyecto es elaborar un modelo de Machine Learning que permita predecir si una canción será exitosa o no a partir de un análisis detallado de sus características musicales.

Con este proyecto, se espera ayudar a la industria musical a tomar decisiones más informadas y precisas en cuanto a la producción y promoción de nuevos lanzamientos, lo que en última instancia beneficiará a plataformas de streaming, discográficas, artistas y oyentes.

RESUMEN DE METADATA

La plataforma Spotify proporciona un dataset que contiene información muy útil para el proyecto. El dataset original contaba con una cantidad muy numerosa de registros ya que contenía el ranking diario de los años 2017 y 2018 para cada región del mundo. Para simplificar el análisis y optimizar el procesamiento de los datos, se trabajó con un nuevo dataset filtrado por la región estadounidense como muestra de la población total.

Luego de la limpieza y transformación del dataset, se cuenta con una cantidad de **72692 registros** y **20 columnas**. Las variables que componen al dataset son: posición en el ranking, nombre de la canción, artista, cantidad de reproducciones **(variable target)**, año, mes y día del ranking. El resto de las variables corresponden a características musicales medidas en diferentes rangos numéricos: bailabilidad, energía, clave musical, ruido, modo (mayor o menor), presencia de palabras habladas, acústica, volumen de instrumentos, probabilidad de que se haya grabado en vivo, positivismo, tempo, duración, compás medio.

PROBLEMA A RESOLVER

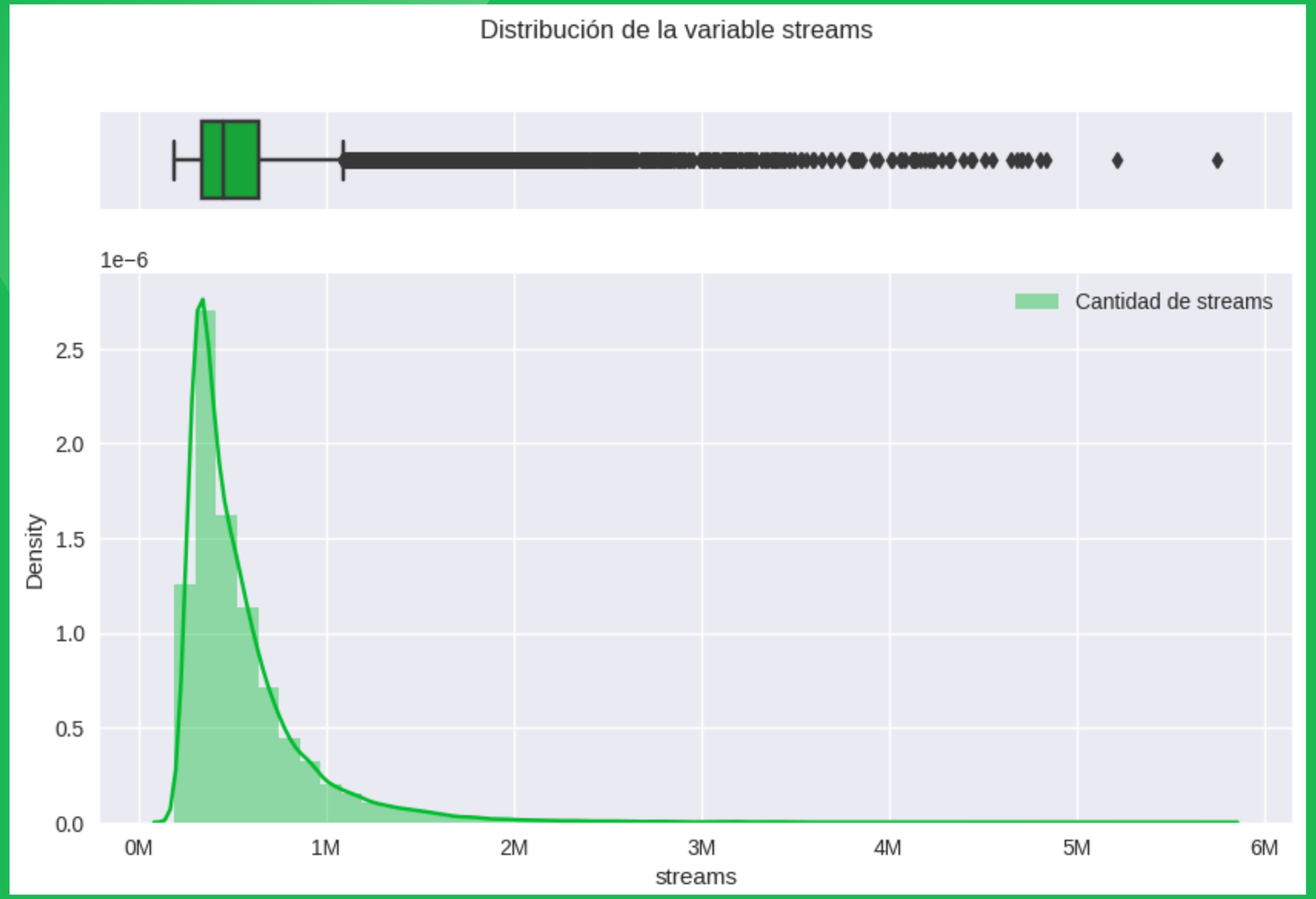
La pregunta objetivo principal de este proyecto es: ¿Qué características debe tener una canción para ser exitosa en cuanto a su cantidad de reproducciones?

Se emplearán herramientas y algoritmos de aprendizaje automático para desarrollar modelos que busquen predecir el éxito de una canción en función de sus características musicales. Además, se crearán visualizaciones y se usarán técnicas de análisis exploratorio de datos para comprender mejor la relación entre las variables.

Para poder aplicar todos los conocimientos adquiridos en el curso, en la etapa de entrenamiento de modelos el problema fue adaptado a los tres tipos de algoritmos vistos: **regresión, clasificación y clustering**.

EDA + VISUALIZACIONES

DISTRIBUCIÓN DE LA VARIABLE OBJETIVO

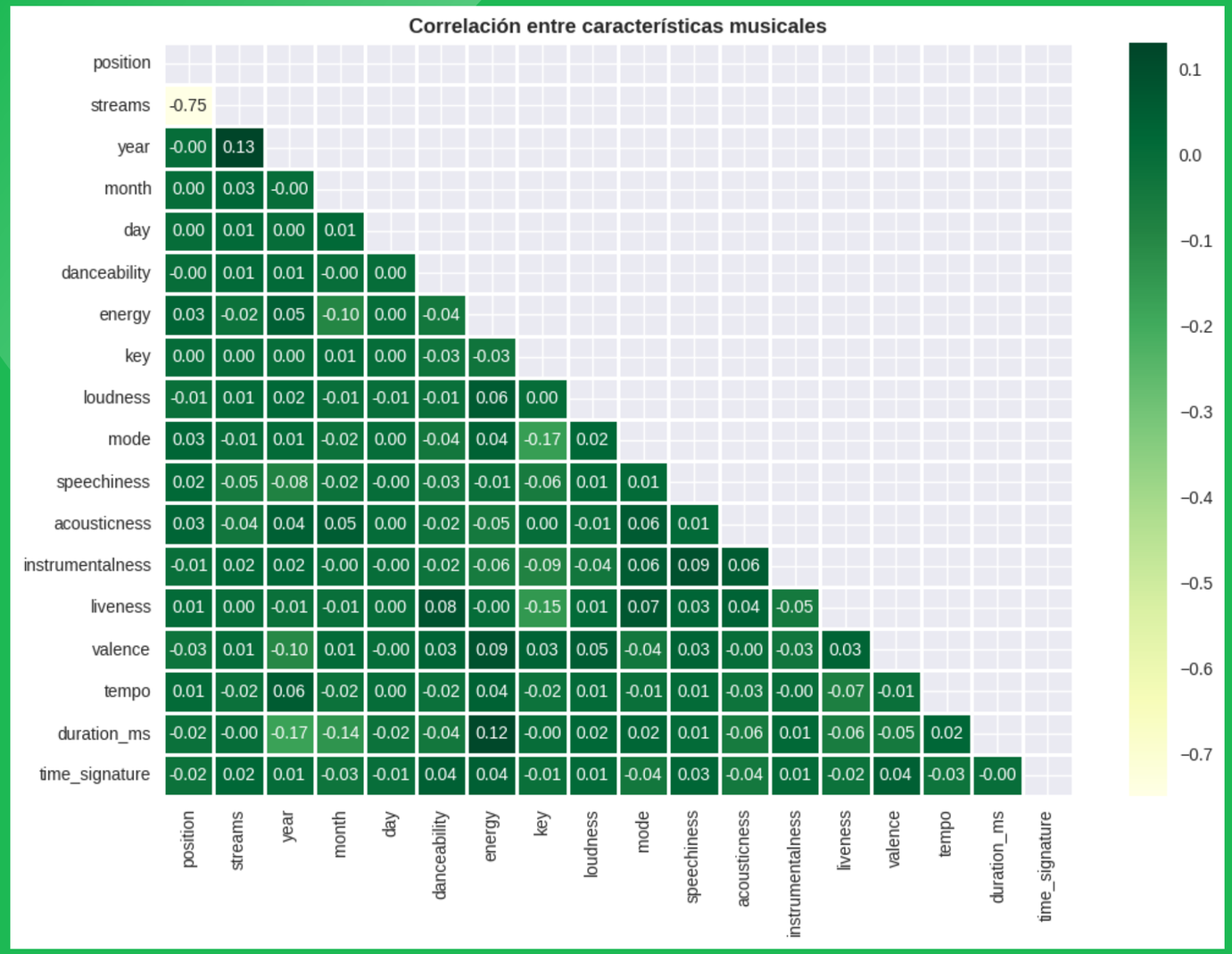


La varianza de los datos es alta, con una gran concentración de datos alrededor de la media y valores extremos que contribuyen a la asimetría y curtosis de la distribución.

El mínimo de reproducciones es de **~192K**, mientras que el máximo es **~5M**. El promedio gira en torno a los **500K** y solo el **8.23%** de las canciones logran superar el millón de reproducciones diarias.

Se cuenta con presencia de valores extremos por encima de los **5M**, aunque luego de analizarlos se encontró que corresponden a lanzamientos realmente exitosos en su día de estreno.

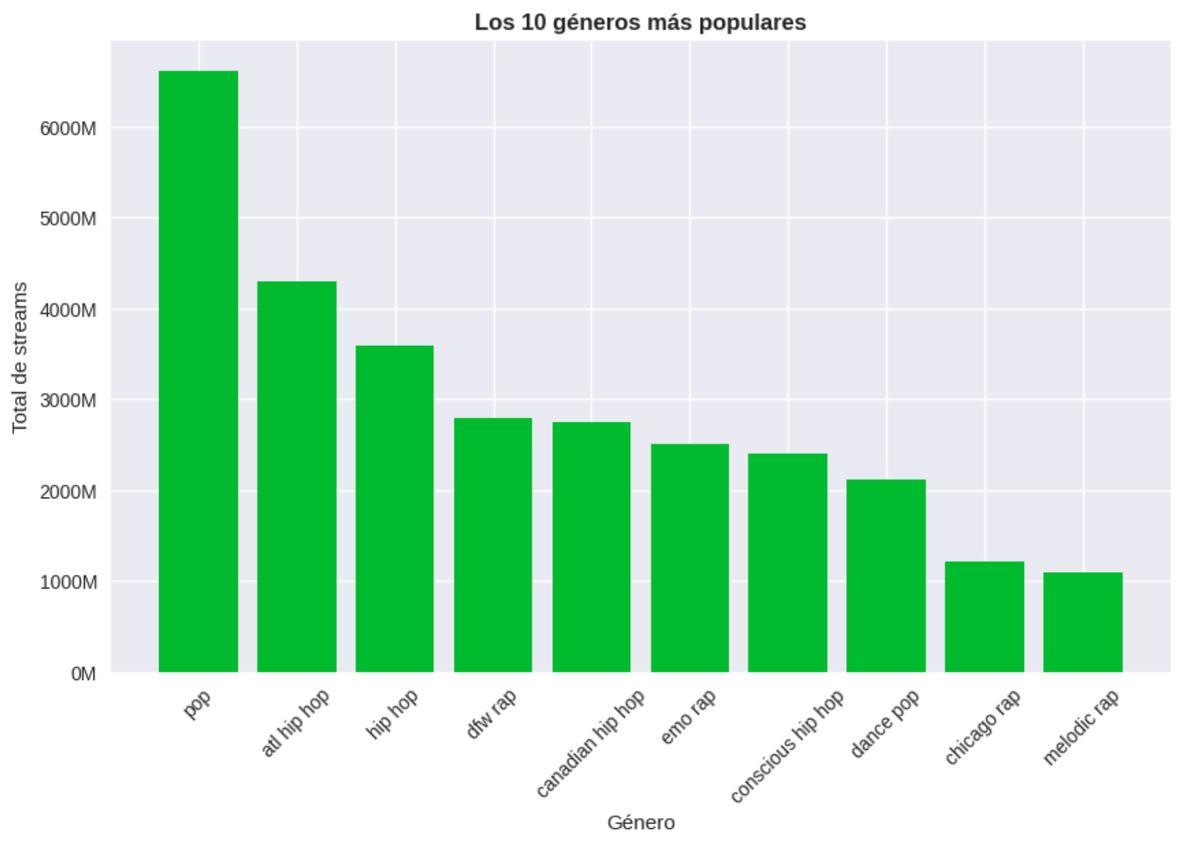
RELACIÓN ENTRE CARACTERÍSTICAS MUSICALES

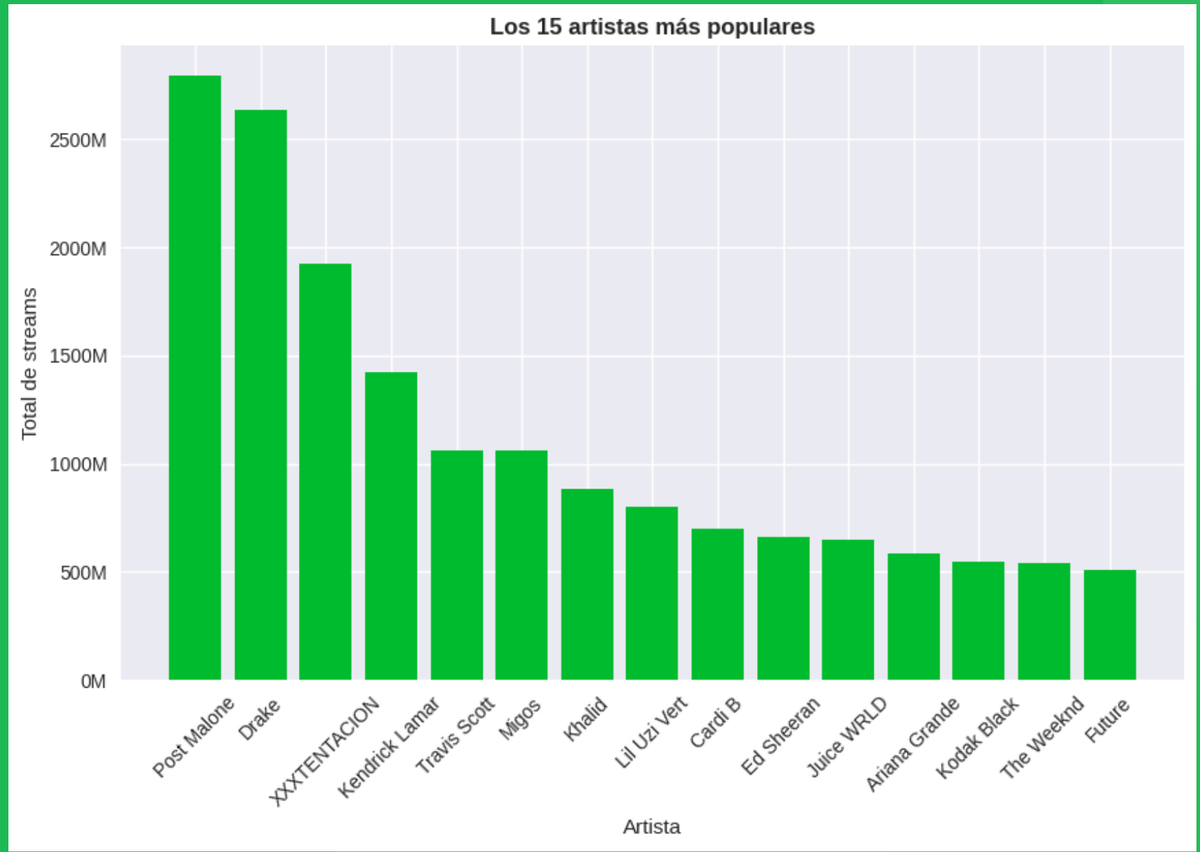


No existen correlaciones lineales **fuertes** entre las variables y tampoco al compararlas con la cantidad de streams salvo por la posición del ranking, cosa que tiene sentido ya que es una variable que depende directamente de la cantidad de reproducciones.

Se realizó un análisis más detallado con foco en las correlaciones significativas **débiles**, tomando como referencia las que se encuentran por encima de un valor absoluto de 0.1 de correlación. En los tres casos analizados, se encontró que existe una relación estadística significativa.

ARTISTAS Y GÉNEROS

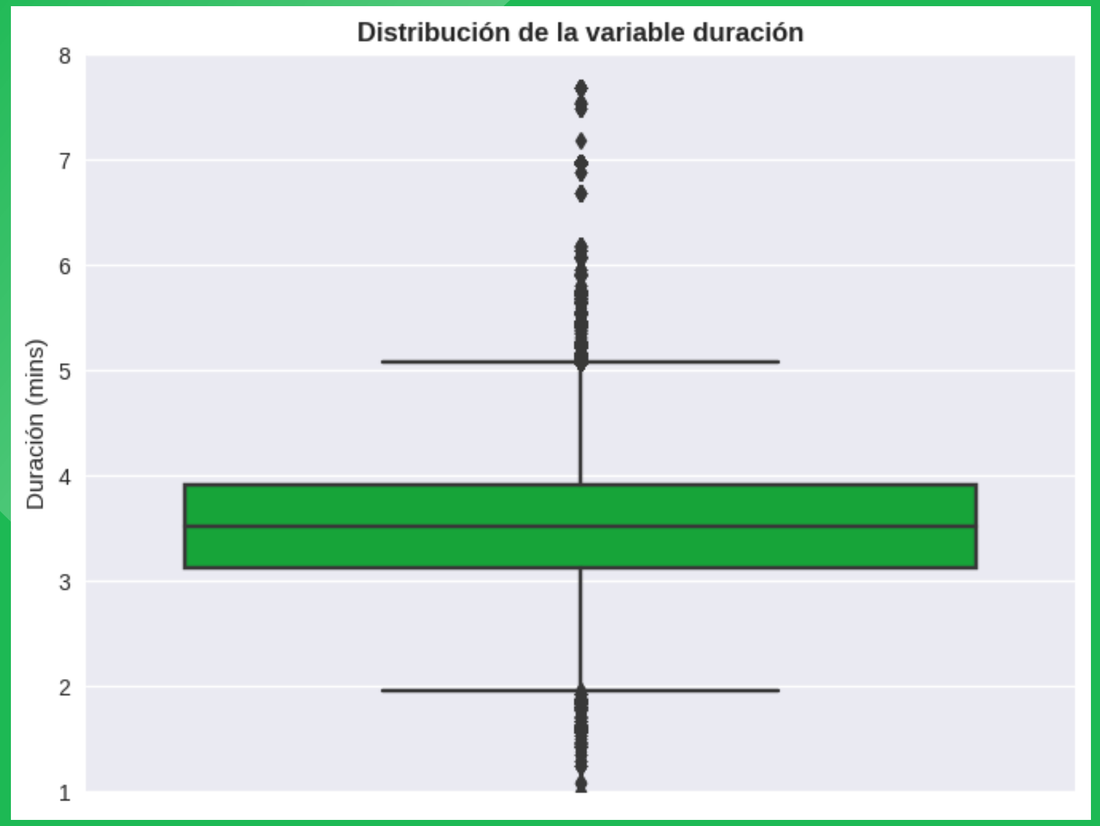


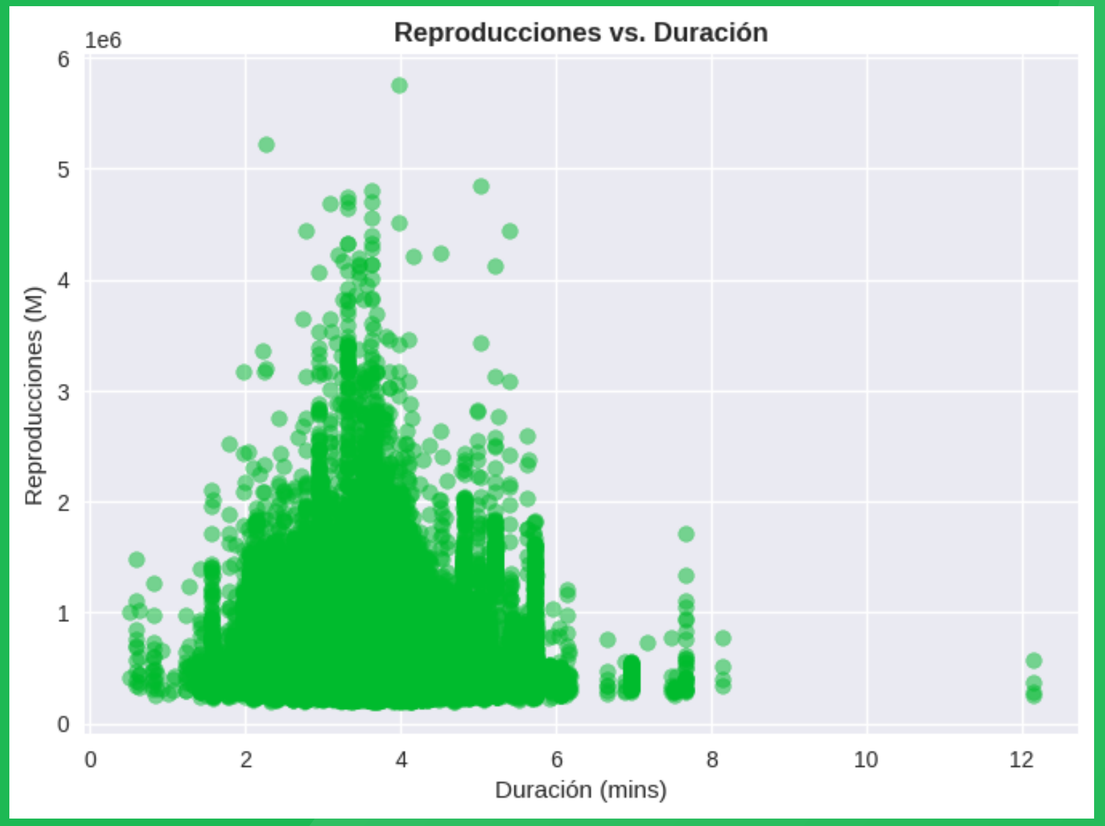


Las diferentes variantes de **Hip Hop** y **Rap** se encuentran entre los géneros más populares junto con el **Pop**.

La mayoría de artistas ubicados en el top 15 pertenecen al género **Hip Hop / Rap**, confirmando la popularidad que veiamos en el gráfico anterior.

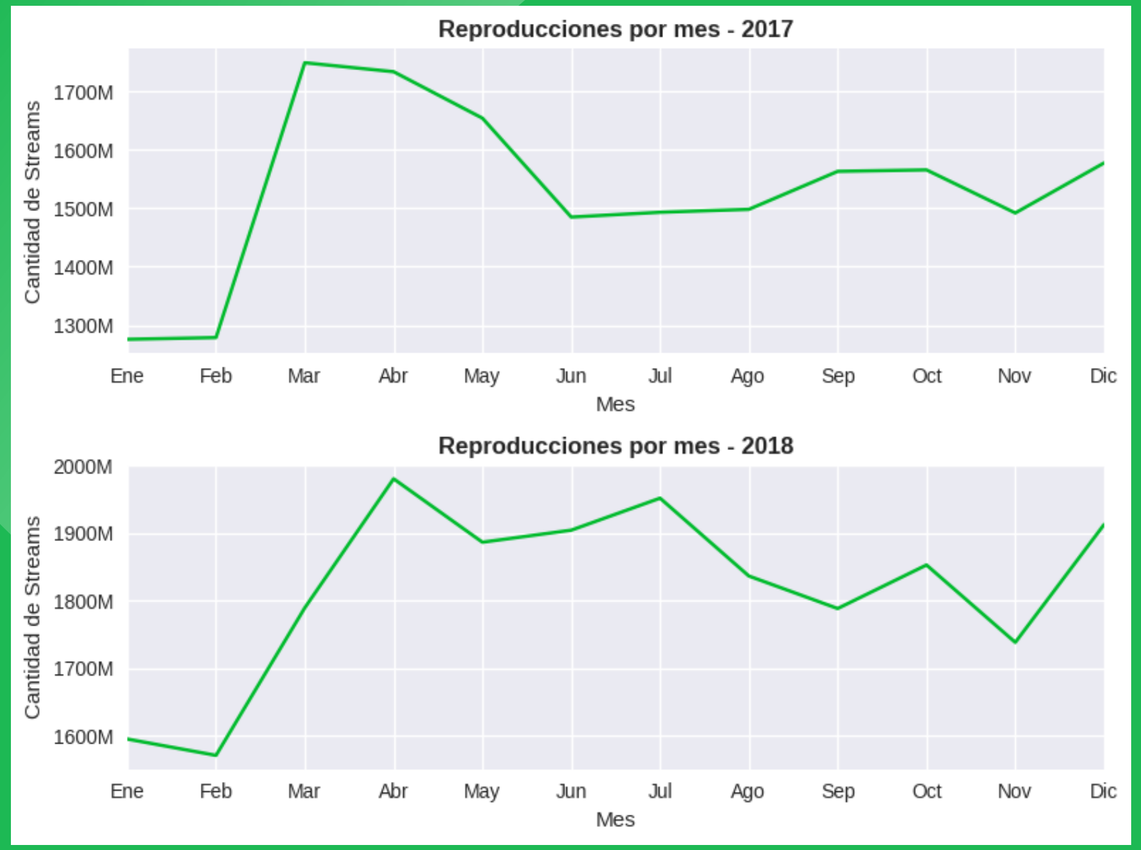
DURACIÓN DE LAS CANCIONES



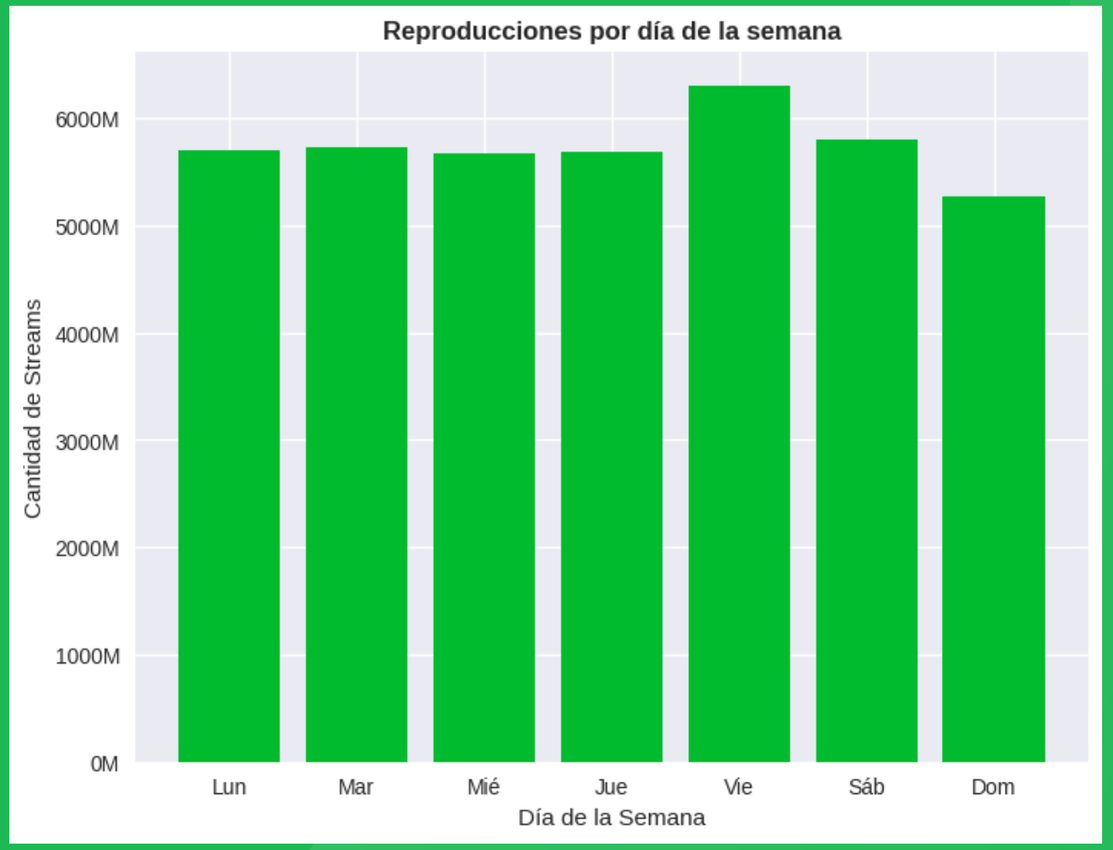


El promedio de duración de las canciones es de entre **3 a 4 minutos.**
Las canciones con una duración en torno al promedio tienen más posibilidades de ser exitosas, mientras que las canciones con corta o larga duración no consiguen más de 2M reproducciones.

PATRONES TEMPORALES



Ambos años muestran un patrón similar en la evolución de la cantidad de streams durante los **primeros meses del año**.



Los **viernes** son los días con mayor cantidad de reproducciones. El resto de los días tiene una distribución equitativa, salvo por los **domingos** que parecen ser los días de menor actividad.

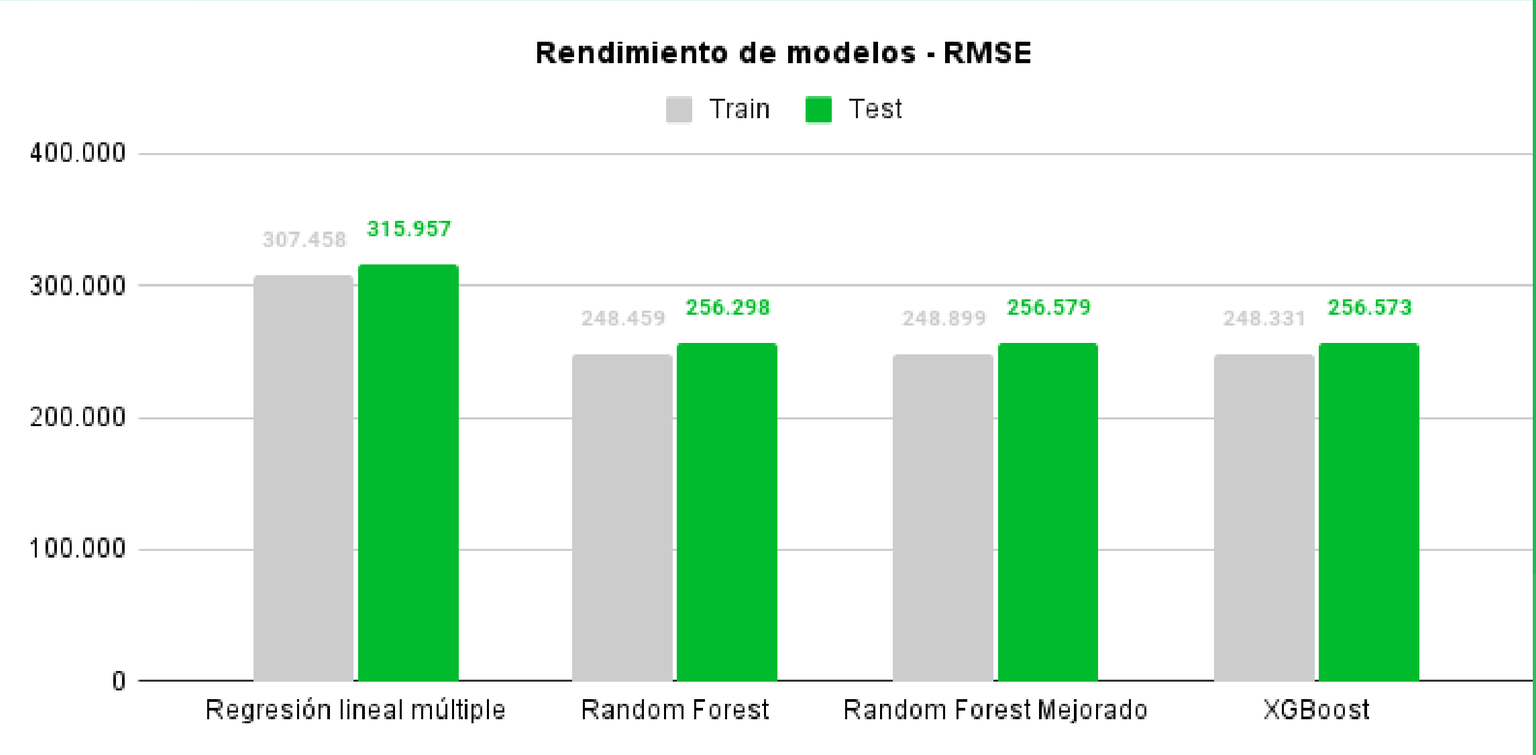
FEATURE ENGINEERING

Se realizaron las siguientes modificaciones sobre el conjunto de datos para prepararlos para el entrenamiento de modelos:

- Se eliminó la variable “url” por no tener relevancia para el análisis y la variable “region” ya que previamente se filtró por una sola región para achicar el volumen del dataset.
- Se corrigió el tipo de dato de algunas variables de acuerdo al diccionario correspondiente.
- Se eliminaron los registros con datos faltantes dado que el volumen del dataset es lo suficientemente grande y eran variables difíciles de reemplazar por ser categóricas.
- Se agrego la variable “genre” al dataset, variable muy importante para el análisis que se obtuvo desde la API oficial de Spotify.
- Se decidió conservar todos los outliers luego de realizar el EDA, ya que se observó que no se trataba de errores y que tienen un gran aporte para el entrenamiento de modelos.

SELECCIÓN DE MODELOS

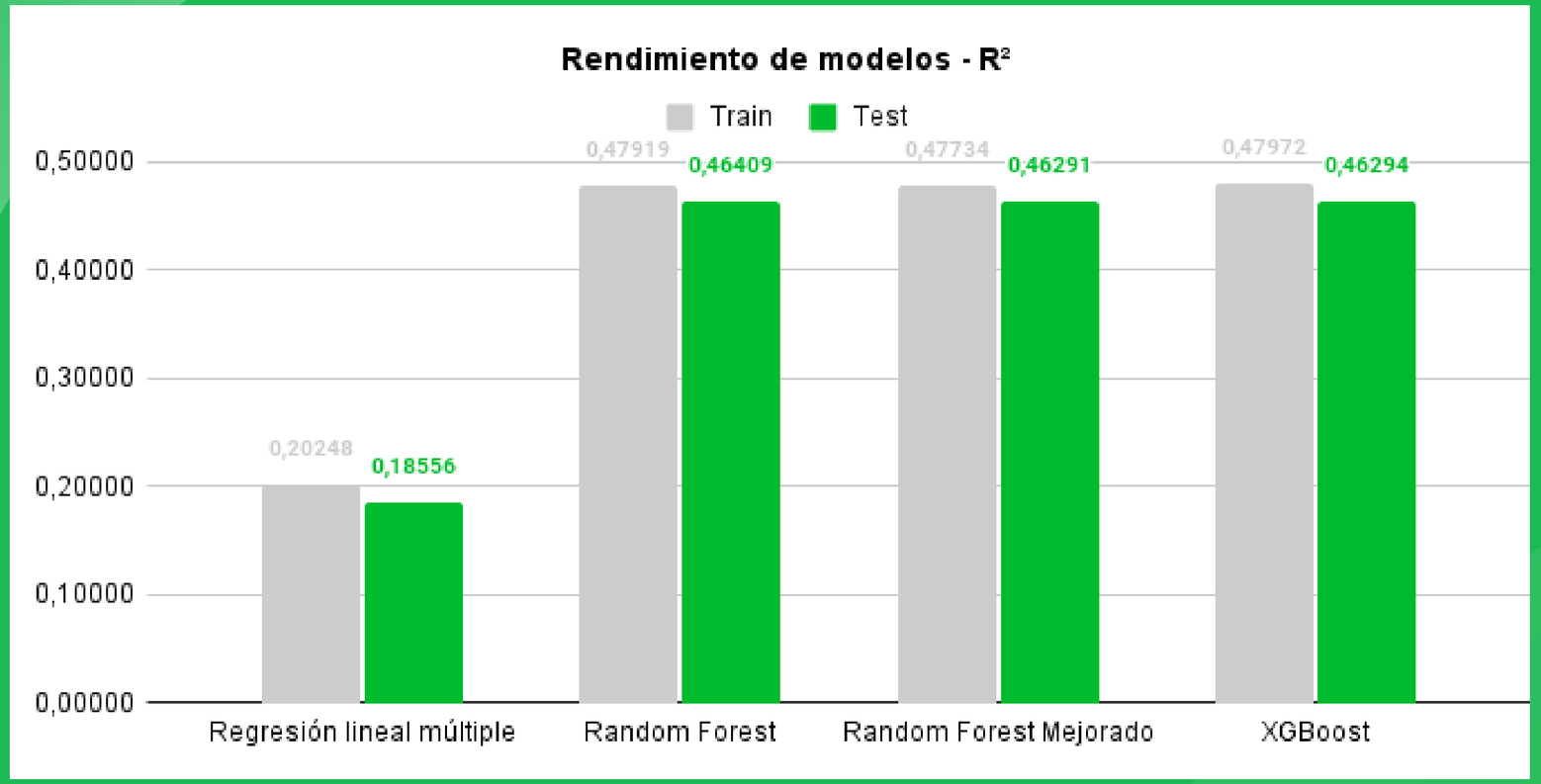
MODELOS DE REGRESIÓN



Para el entrenamiento de todos los modelos se uso la técnica de validación cruzada **K-fold** y se alternaron los métodos de búsqueda en grilla para optimización de hiperparámetros.

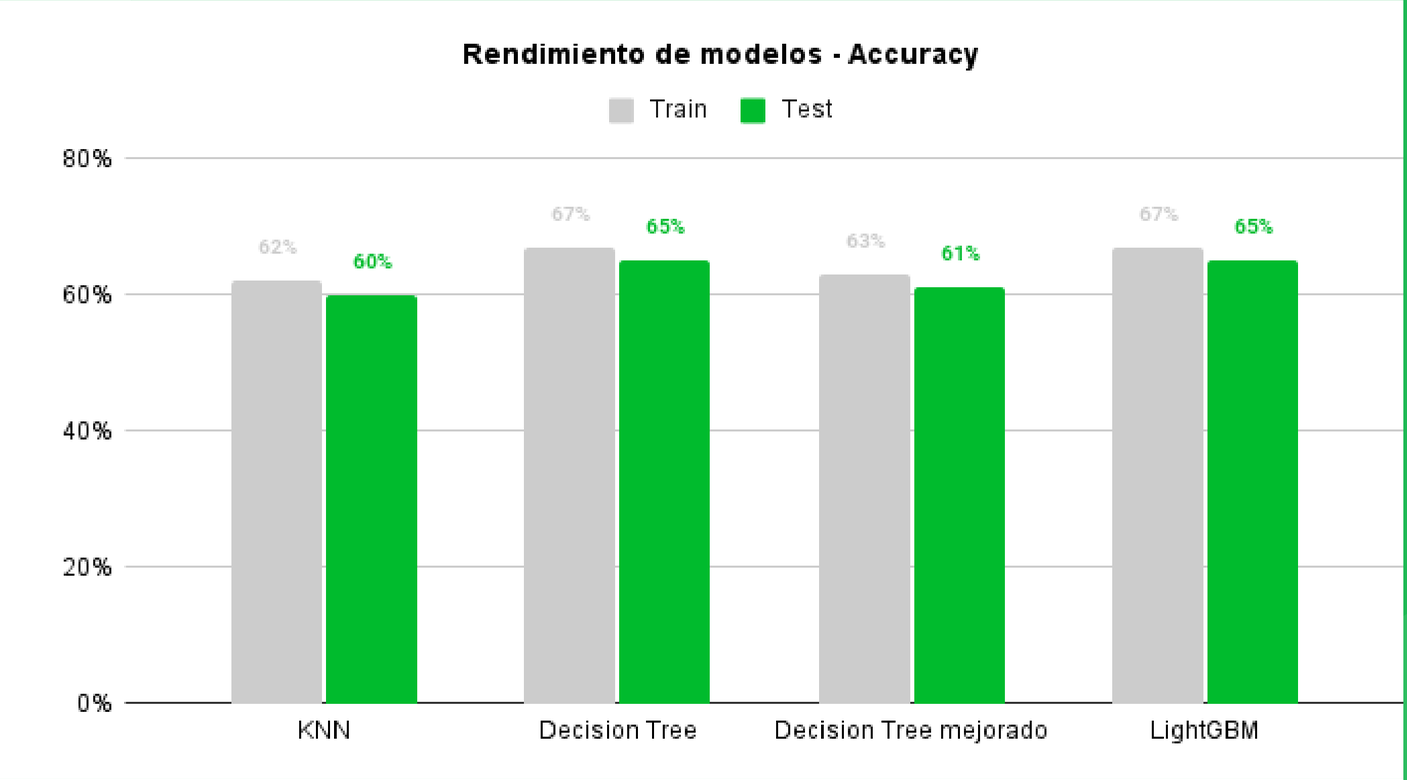
En términos generales, los modelos basados en árboles tienen resultados de rendimiento muy similares y muestran una menor tendencia al overfitting en comparación con la Regresión Lineal Múltiple. En cuanto al Random Forest mejorado, si bien los resultados obtenidos no lograron mejorar al modelo original, su tiempo de entrenamiento fue menor alcanzando un rendimiento bastante similar.

MODELOS DE REGRESIÓN



Al incluir el coeficiente de determinación en la evaluación de modelos, los resultados respaldan la conclusión de que los modelos basados en arboles tienen mejor rendimiento, destacando entre ellos el modelo **XGBoost**. Su capacidad para minimizar el error y explicar la variabilidad en los datos sugiere un rendimiento superior en comparación con los demás modelos evaluados.

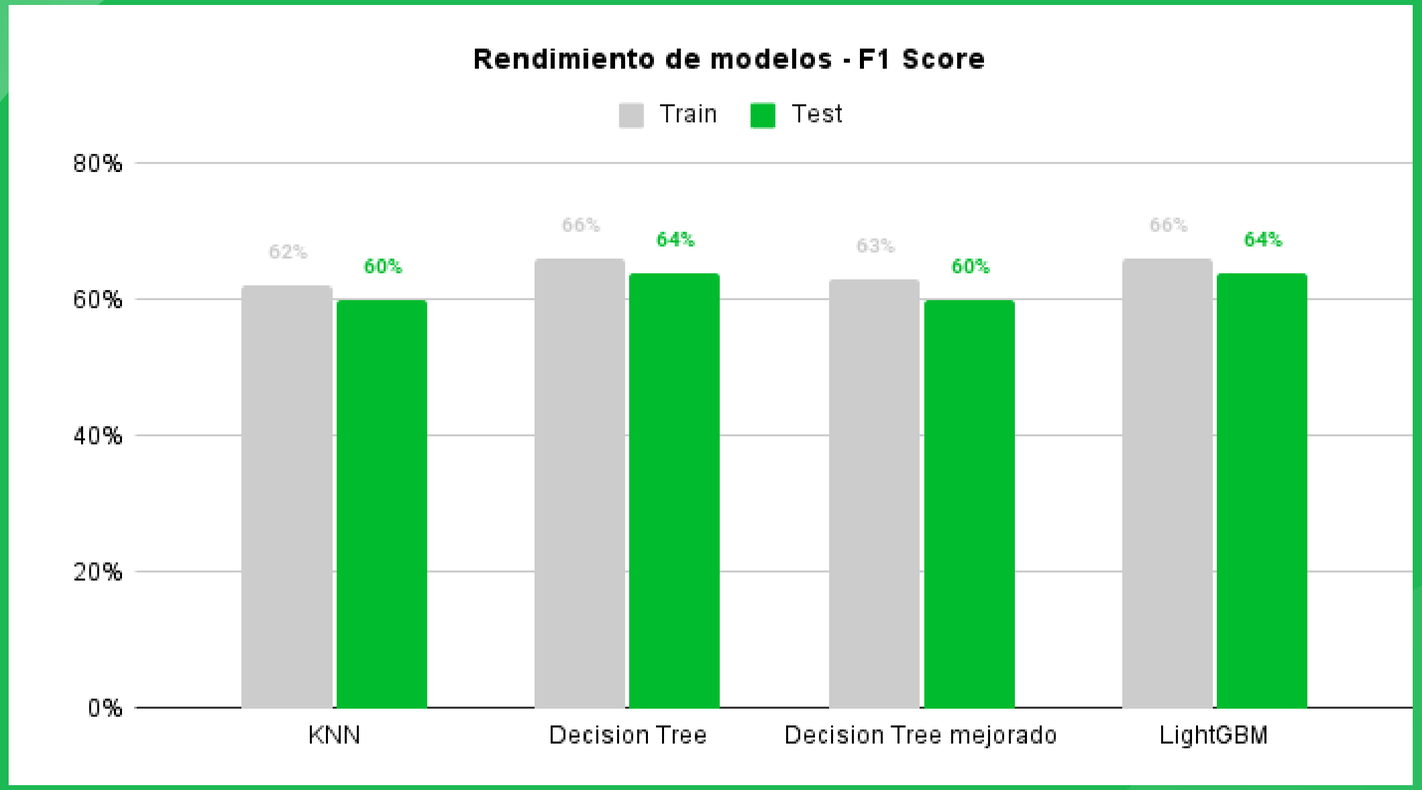
MODELOS DE CLASIFICACIÓN



Para el entrenamiento de todos los modelos se uso la técnica de validación cruzada **StratifiedKFold**, salvo en el caso del Decision Tree mejorado en donde se balancearon previamente las clases. Además, se alternaron los métodos de búsqueda en grilla para optimización de hiperparámetros.

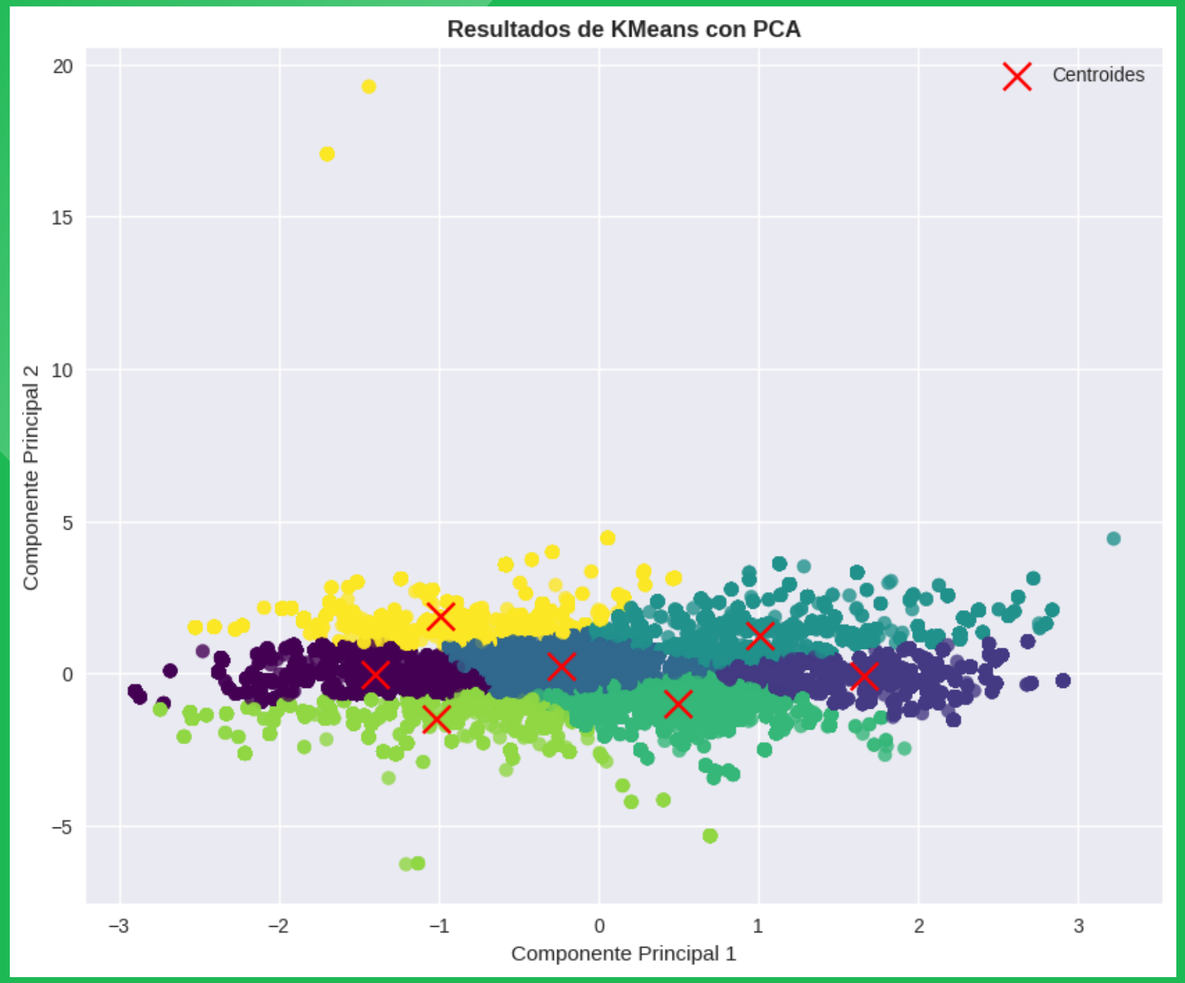
En términos generales, los modelos **Decision Tree** y **LightGBM** presentan el mejor rendimiento. En cuanto al Decision Tree mejorado, si bien los resultados obtenidos no lograron mejorar al modelo original, su tiempo de entrenamiento fue menor alcanzando un rendimiento bastante similar.

MODELOS DE CLASIFICACIÓN



Al incluir el F1 Score en la evaluación de modelos, los resultados respaldan la conclusión de que tanto **Decision Tree** como **LightGBM** son los mejores modelos, logrado un equilibrio eficaz entre precisión y recall en ambos conjuntos de datos.

MODELO DE CLUSTERING



Al convertir el problema en uno de clustering con el algoritmo **K-means**, se identificaron patrones entre las canciones y se las agrupó según sus similitudes.

Esto permitirá entender mejor el mercado y la competencia que tendrá una nueva canción.

Se estableció una cantidad óptima de 7 clusters y se aplicó PCA para poder graficar en 2D los clusters obtenidos.

¡GRACIAS!