Proyecto de Machine Learning para Predicción de Popularidad en Spotify

Autor: Javier Felipe Suarez
Carrera :Ciencia de Datos
Institucion: CoderHouse
Comision: 32795

01 Contexto y Audiencia

02 Hipótesis/Preguntas de Interés

03 Metadata

04 Análisis exploratorio

05 Insights y Recomendaciones

CONTEXTO, AUDIENCIA Y LIMITACIONES

Contexto

Una pregunta qué probablemente la mayoría de los músicos y productores musicales se hacen en los últimos años es qué hace que una pista logre muchas reproducciones en Spotify.
Con unos 158 millones de usuarios en 178 países, Spotify es el líder mundial en el sector de la música en streaming.

Análisis

Mi análisis se basará en 586 mil registros propios de Spotify, los datos se dividen en unas 20 columnas, cada una de las cuales describe la pista y sus características.
Procederé a analizar los datos obtenidos de Kaagle enfocándome en las relaciones entre las pistas de mayor popularidad y sus características musicales, para intentar encontrar posibles elementos que ayuden a incrementar la popularidad de un track en Spotify.

Audiencia

Este análisis intenta contestar, con evidencia, las preguntas del párrafo anterior por lo cuál puede ser de utilidad tanto para músicos independientes que quieran mejorar su posicionamiento en la app, como para productores musicales.

Limitaciones

Para medir la popularidad, Spotify mantiene una relación entre cantidad de reproducciones totales y actuales, lo cual genera un sesgo temporal. Por otro lado, medir exactamente qué da popularidad a la música, cómo en toda expresion artística, no es absolutamente posible. Por tanto, los datos aquí obtenidos son orientativos.

PREGUNTAS DE **INTERÉS**

Preguntas principales:

- ¿Qué características tienen los tracks más populares que encontramos de Spotify?
- ¿Existe relación entre la estructura musical de una pista y su popularidad?

Preguntas secundarias:

- ¿Qué relación hay entre la popularidad y las notaciones musicales?
- ¿El lenguaje explícito es más o menos popular?
- ¿La bailabilidad influye en la popularidad?
- ¿Un track popular qué relación tiene con la energía?
- ¿El uso de lenguaje hablado y la instrumentalidad influyen en la popularidad?

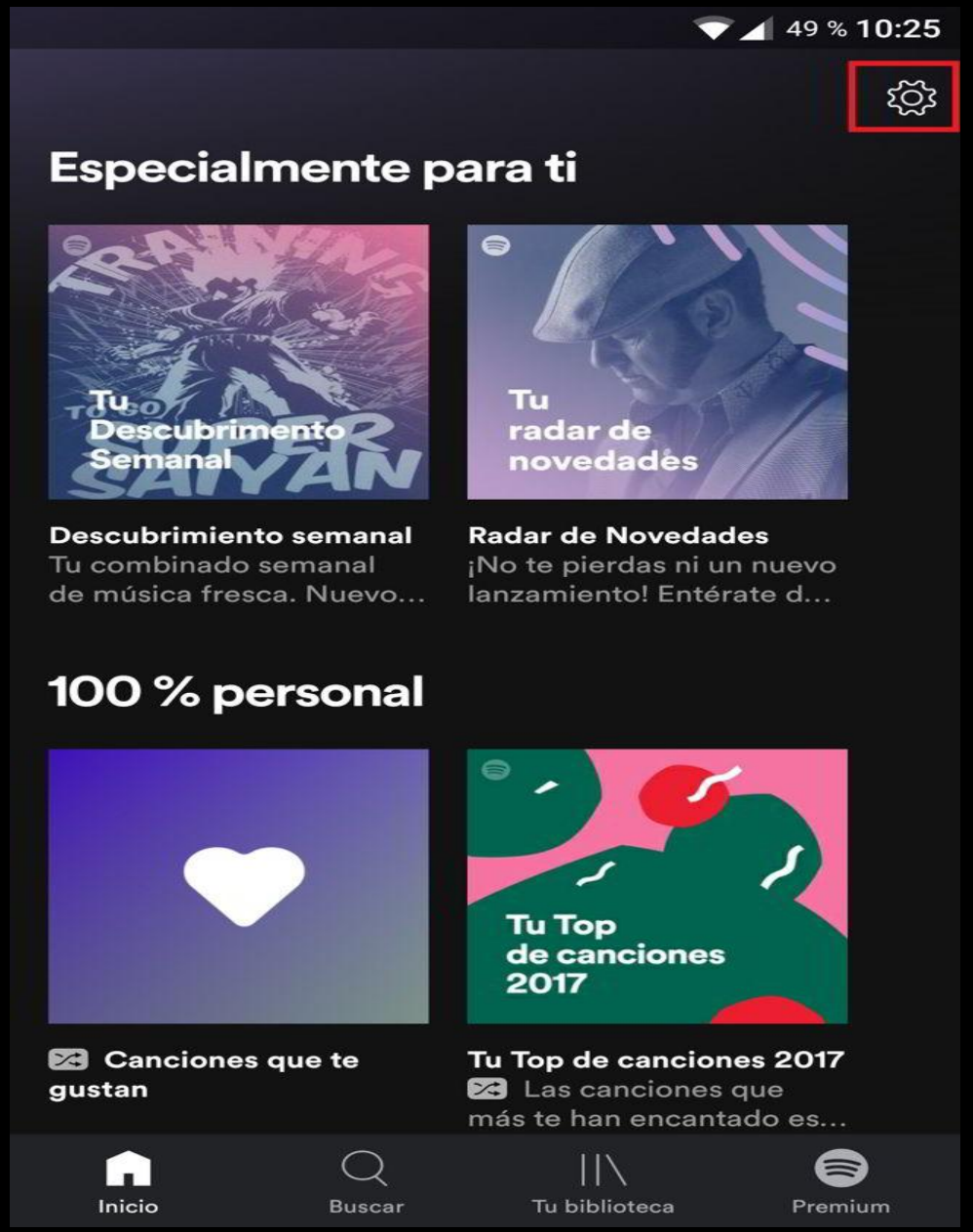
RESUMEN METADATA

114K
ARTISTS

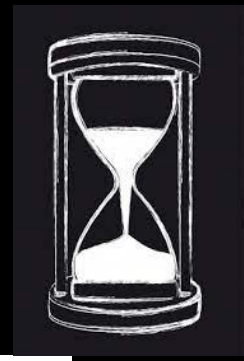
101 AÑOS
EN
CANCIONES

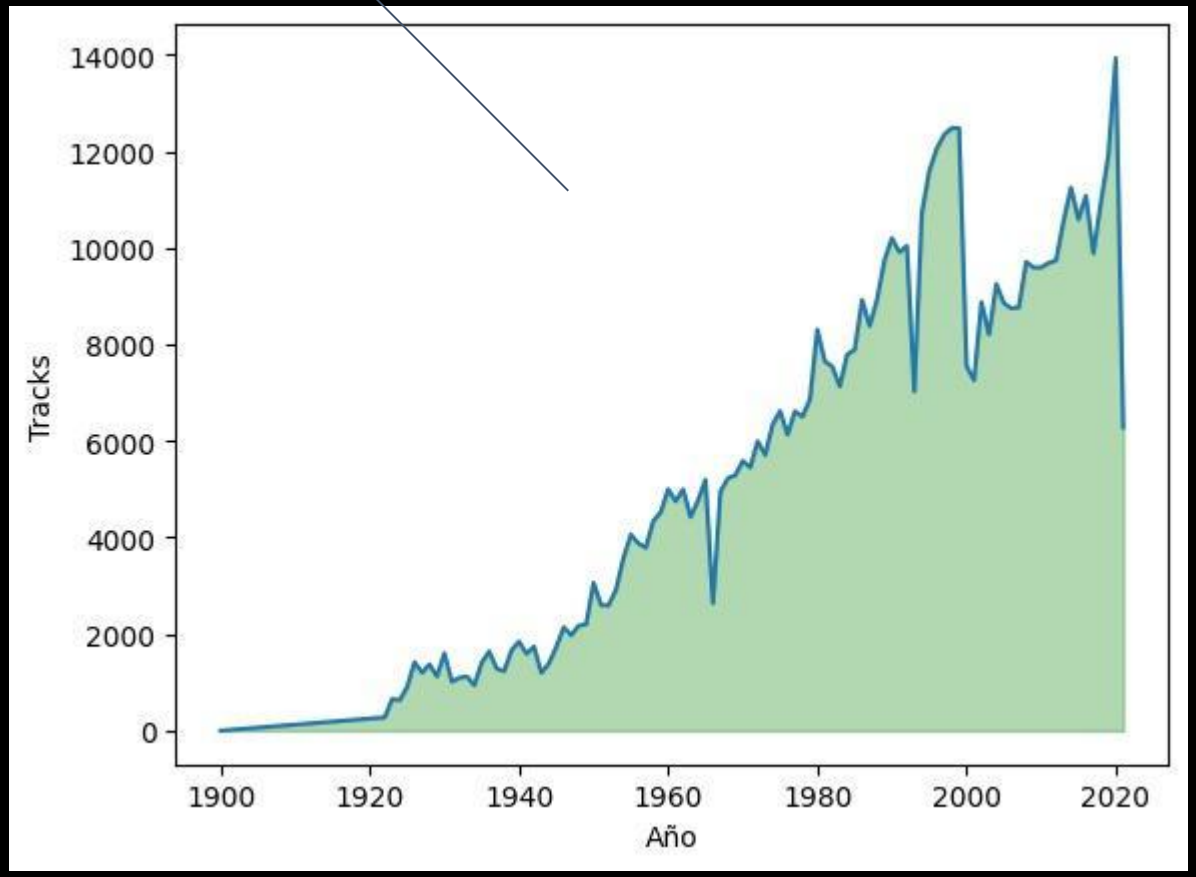
586k
TRACKS

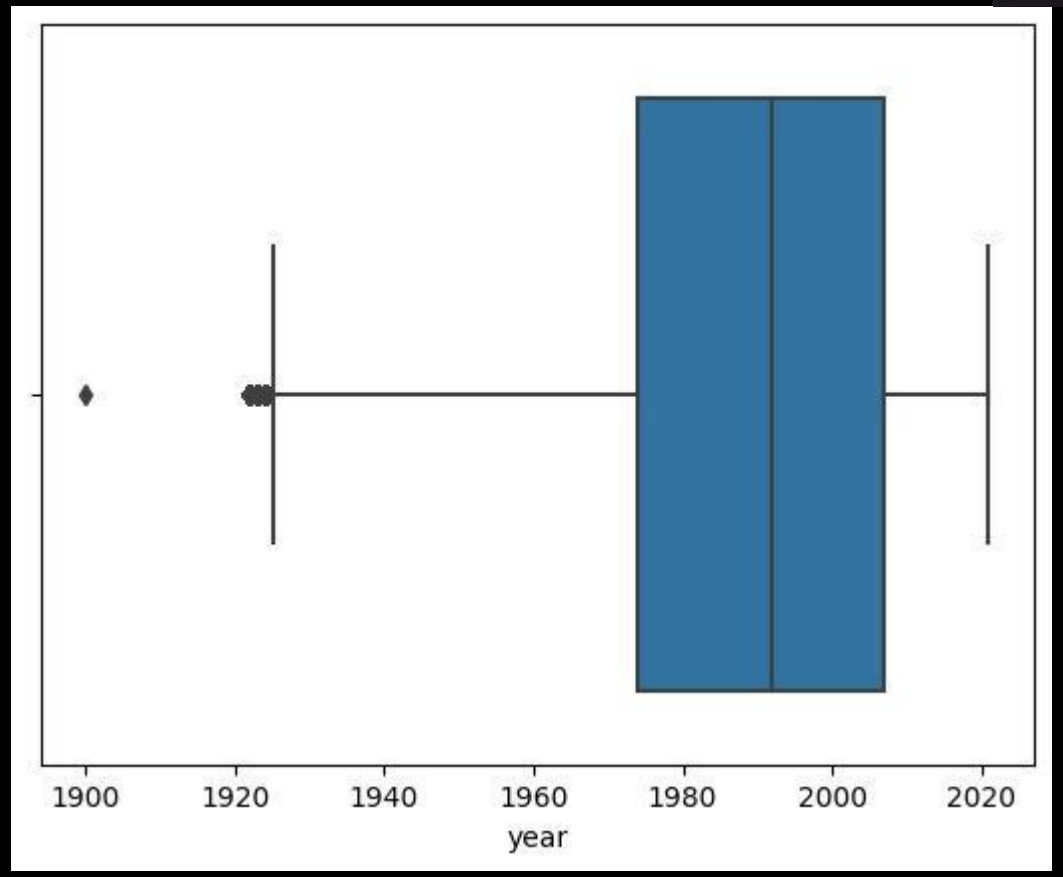
ANÁLISIS EXPLORATORIO



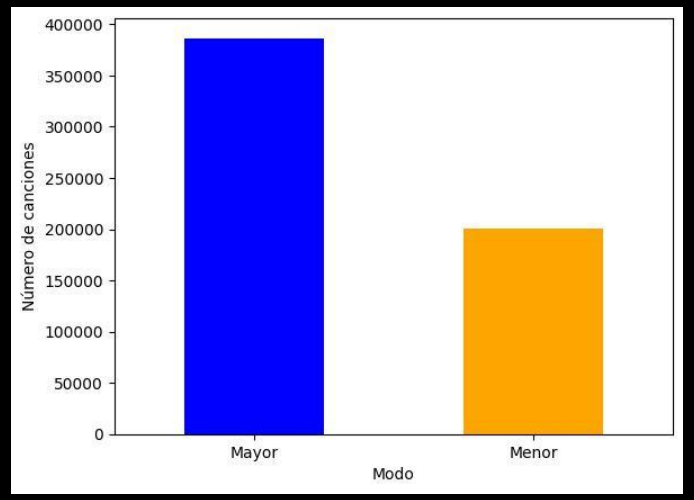
DISTRIBUCIÓN DE LOS TRACKS EN EL TIEMPO (La mayoría se concentran en los últimos 30 años)

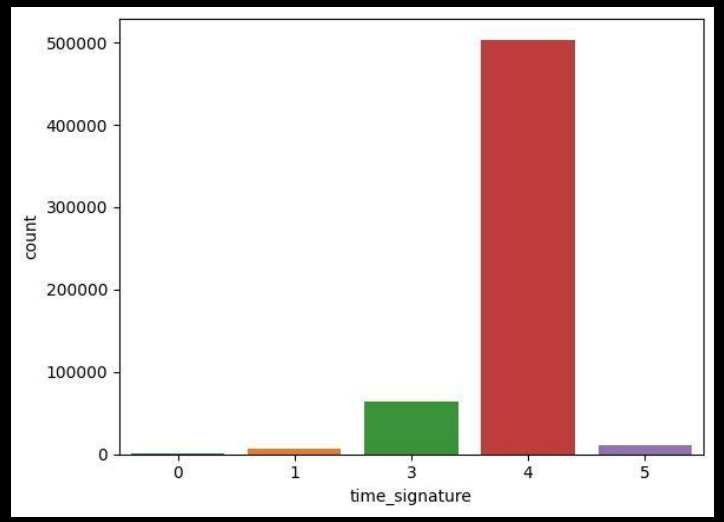


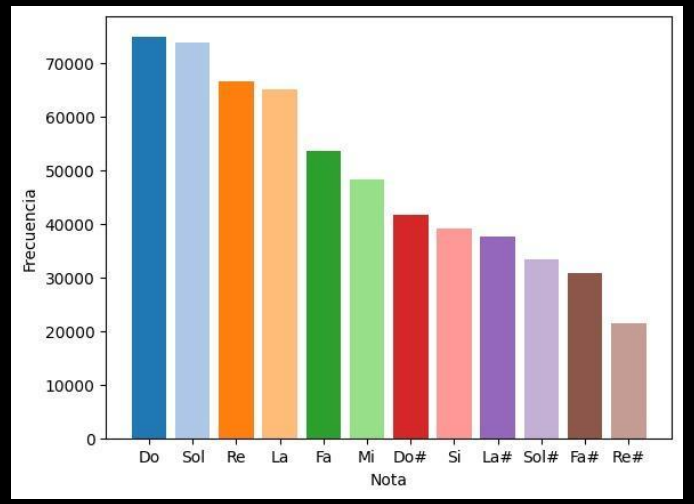


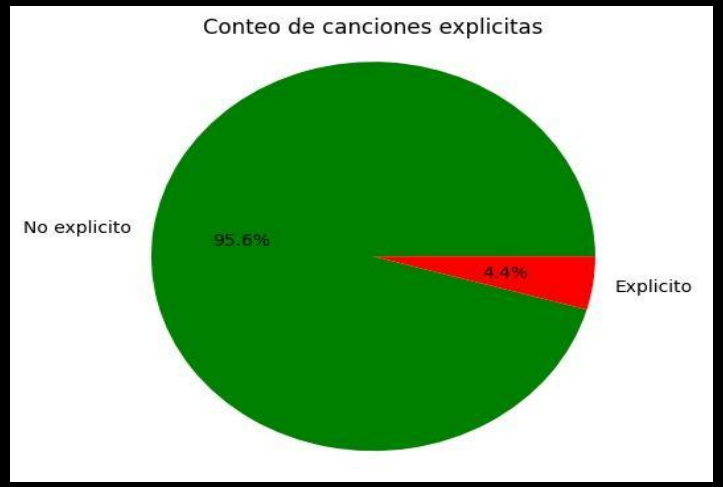


¿CÓMO ES LA ESTRUCTURA MUSICAL DE LOS TRACKS DE SPOTYFY? 586K TRACKS



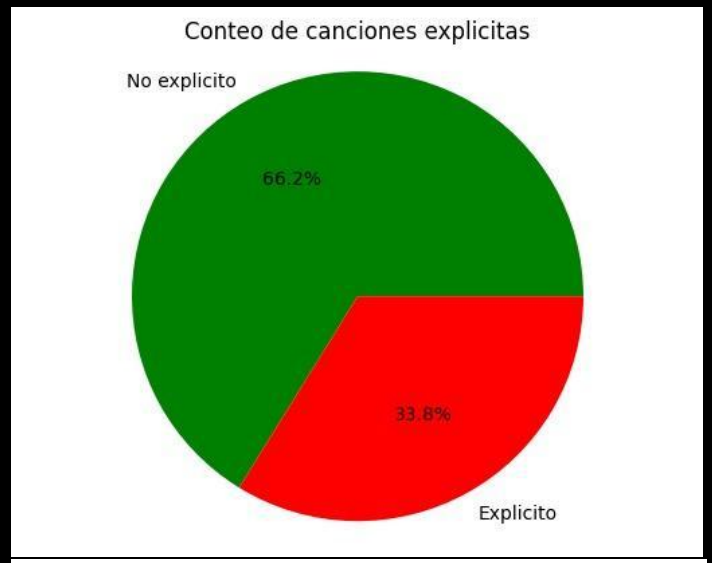


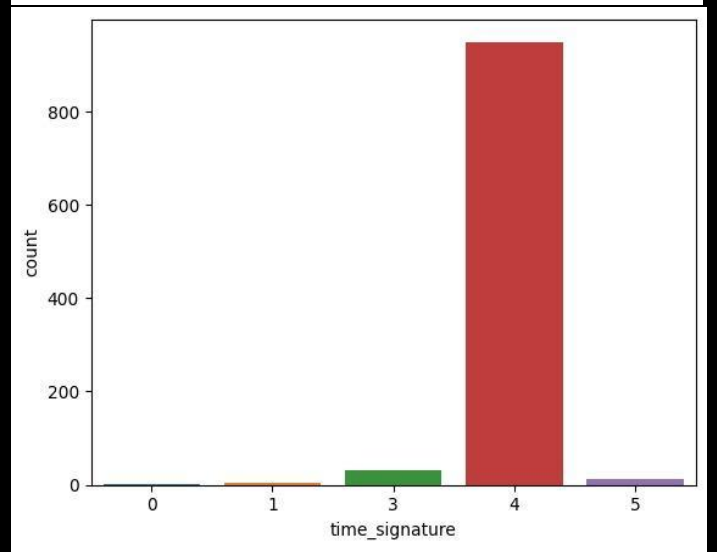


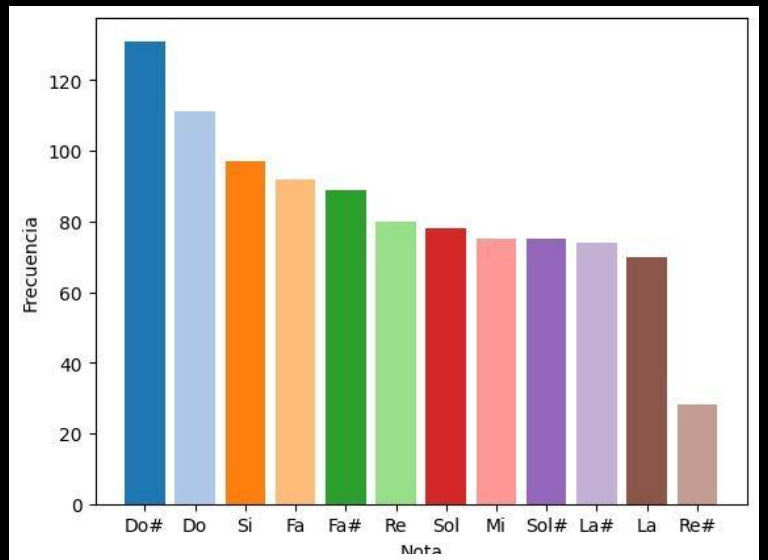


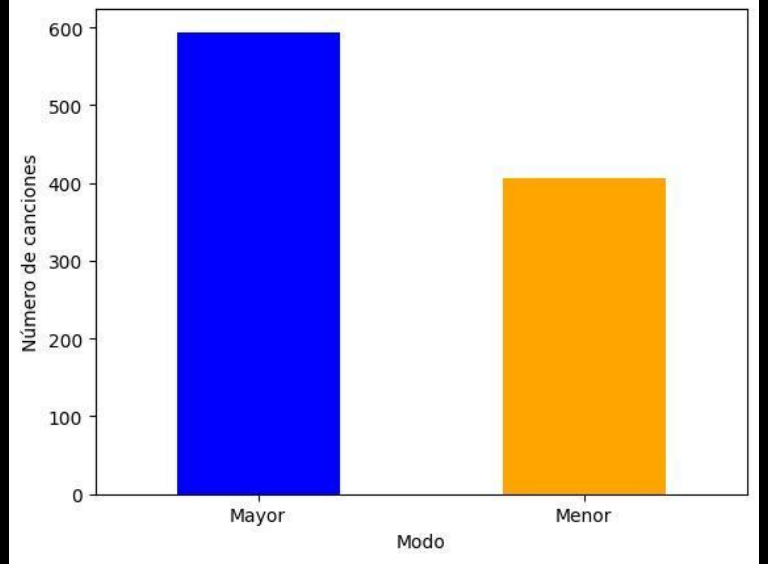
- <u>VALORES MEDIOS</u>
- ENERGÍA: 54
- POPULARIDAD: 27
- BAILABILIDAD: 56
- ESTADO DE ÁNIMO (VALENCE): 55

¿CÓMO SON LOS 1000 TRACKS MÁS POPULARES?

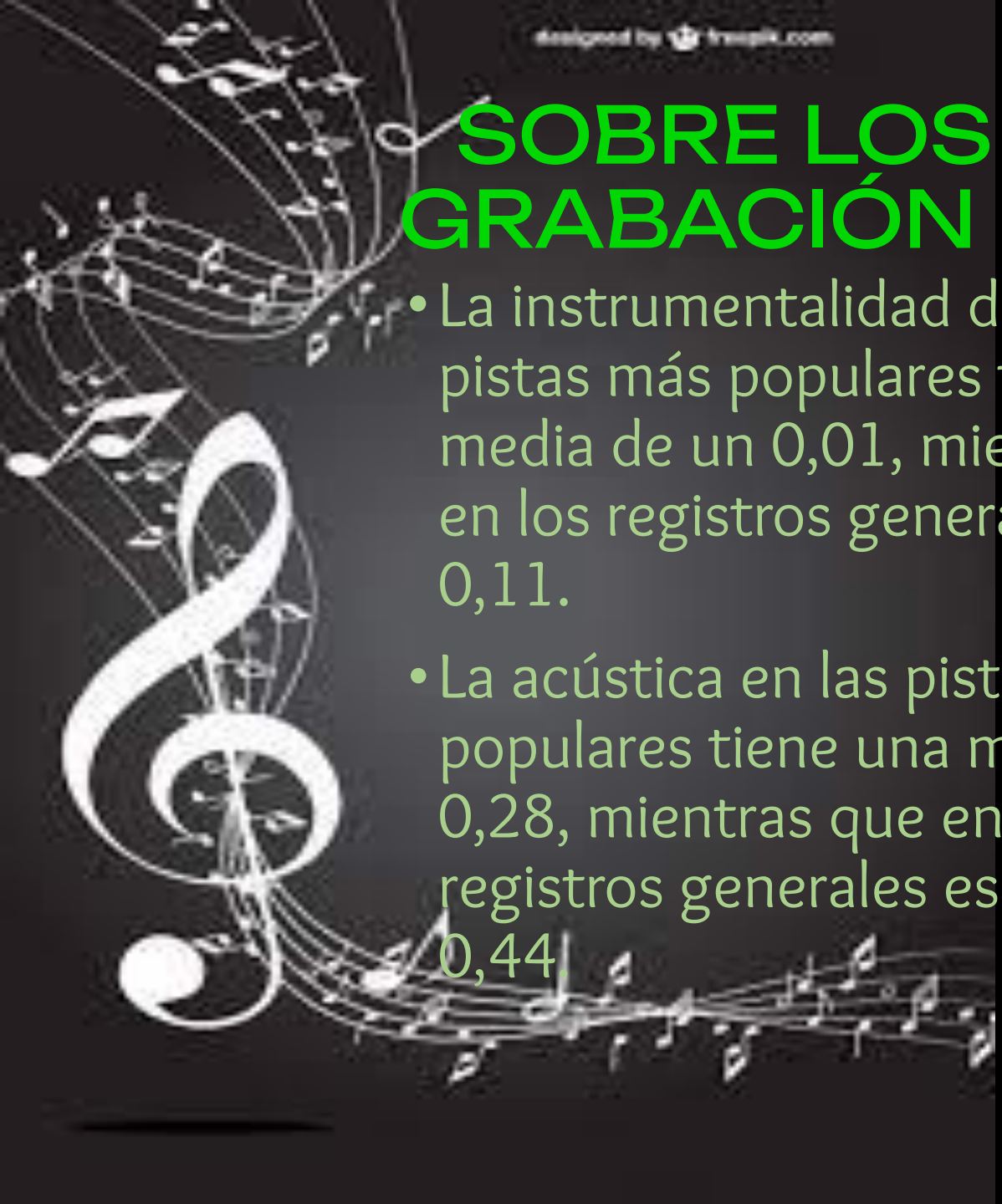








- <u>VALORES MEDIOS</u>
- ENERGÍA: 65
- POPULARIDAD: 82
- BAILABILIDAD: 66
- ESTADO DE ÁNIMO (VALENCE): 51



SOBRE LOS RECURSOS DE GRABACIÓN DE LOS TRACKS

- La instrumentalidad de las pistas más populares tiene una media de un 0,01, mientras que en los registros generales es de 0,11.
- La acústica en las pistas más populares tiene una media de 0,28, mientras que en los registros generales es de un 0,44.
- Las grabaciones en vivo de las pistas más populares tienen una media de 0,21, mientras que en los registros generales es de 0,18
- La cantidad de lenguaje hablado(speechness) es de 0,10 en ambos casos.

INSIGHTS & RECOMENDACIONES

INSIGHTS & RECOMENDACIONES

Insights

*El lenguaje explícito hace que la popularidad, la danzabilidad y la energía aumenten.

*En la estructura musical, las primeras cuatro notas mantienen una dispersión acumulada, pero en las más populares las notas cambian en tres de cuatro. Siendo, DO# la más utilizada.

*La Valencia, o estado de ánimo no parece mantener una corelación directa con la popularidad, ya que baja de 0,55 a 0,44 en las más populares.

*El Time Signature se mantiene en general y los modos también, prevaleciendo los mayores.

*Parecería haber una correlación inversa entre la acústica y la instrumentalidad con la popularidad.

*La media de duración de las canciones populares es de 200' y se mantiene cercana a la media general que es de 230.

Recomendaciones:

*Si lo que se busca es posicionar mejor una pista en Spotify, se sugiere trabajar en composiciones en 4 tiempos, utilizando las notas DO#, SI, DO y FA#, que contengan lenguaje explícito y una alta energía y bailabilidad.

*Se sugiere reducir el uso de instrumentales y acústicos, mantener una duración de no más de 200' y no abusar del uso de lenguaje hablado.

*Tener en cuenta que Spotify mide la popularidad entre las reproducciones totales y las actuales, por tanto, mantener actualizado el catálogo de un artista tiene una correlación directa con su popularidad en la app.

¿Podemos predecir la popularidad de una nueva pista en Spotify antes de su lanzamiento?

¿Qué características musicales influyen más en la popularidad de una pista en Spotify?

Esta pregunta busca identificar qué características específicas de las pistas musicales (como danceability, energy, loudness, valence, etc.) tienen una mayor influencia en su popularidad. Utilizamos técnicas de análisis de correlación y visualizaciones para descubrir estas relaciones. ¿Podemos predecir la popularidad de una nueva pista en Spotify antes de su lanzamiento?

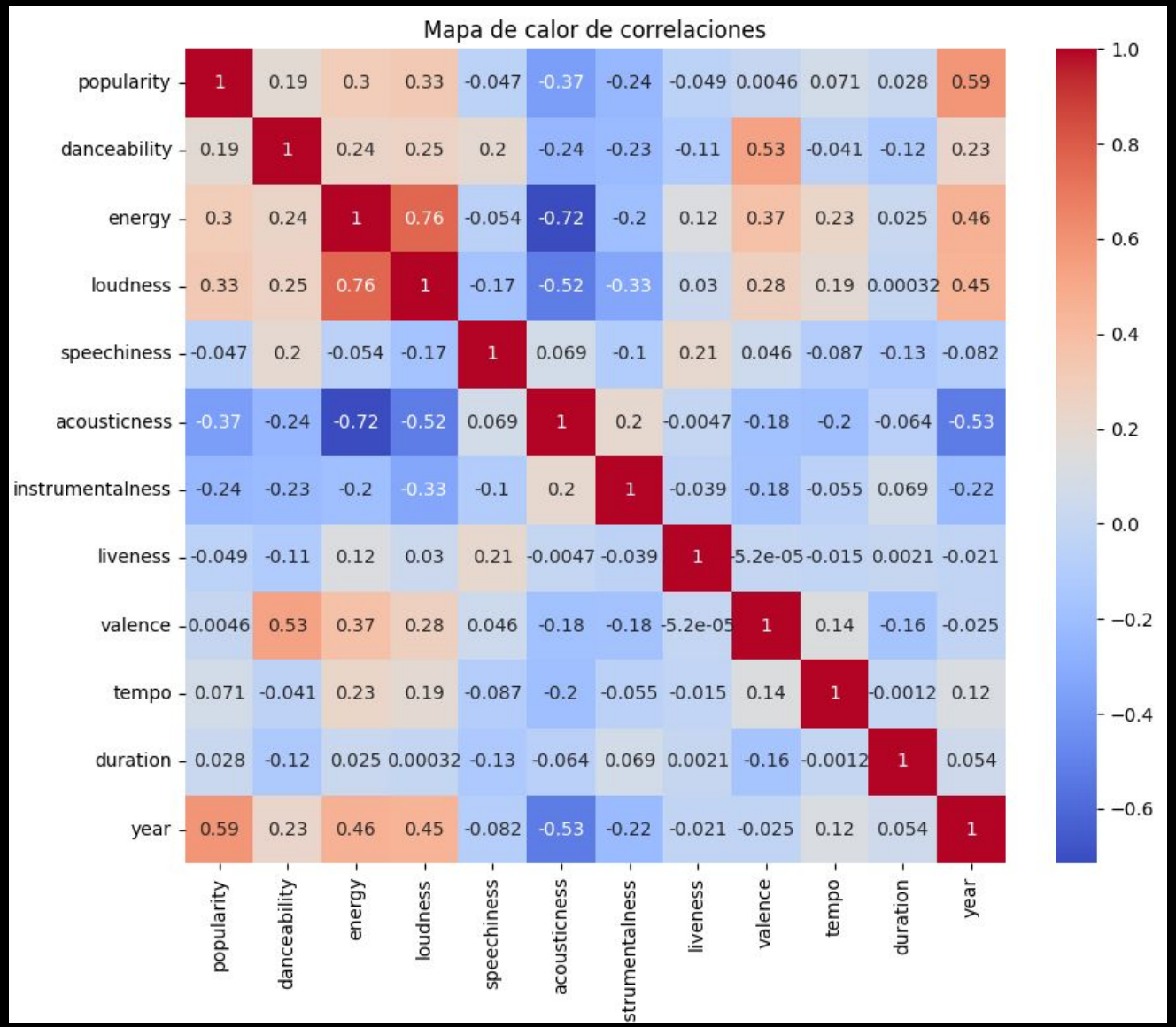
Esta pregunta implica la construcción de un modelo de machine learning capaz de predecir la popularidad de una pista musical en base a sus características antes de que sea lanzada. Utilizamos algoritmos de regresión o clasificación para crear el modelo y evaluar su rendimiento.

Presentación del desafío: Nuestra misión es predecir el ranking de popularidad de canciones en Spotify para tomar decisiones informadas.

Elección del Modelo de Clasificación: En nuestro análisis de la popularidad de las canciones en Spotify, nos encontramos con la tarea de predecir el nivel de popularidad de una canción en tres categorías: **baja, media y alta.** Para abordar esta tarea, exploramos y evaluamos varios modelos de clasificación, incluyendo Regresión Logística, Vecinos Más Cercanos, Árbol de Decisión, Análisis Discriminante Lineal, Naive Bayes y Bosques Aleatorios. A medida que avanzamos en la evaluación de estos modelos, buscamos uno que se ajustara mejor a nuestros objetivos y al problema de negocio en cuestión.

Usamos los rangos interquartiles para generar una categoria Ranking: Alta, Media, Baja.

Variables Determinantes: Al realizar la selección de características, identificamos las variables más determinantes para nuestro modelo. Entre estas variables, se destacaron la energía, el loudness y el tempo. Estos atributos reflejan aspectos importantes de una canción que podrían influir en su popularidad, como su energía general, su nivel de ruido y su ritmo. Al enfocarnos en estas características clave, logramos construir un modelo más efectivo y ajustado a nuestro problema de negocio.



Analisis de Modelos de Machine Learning

Regresión Logística:

- Precisión alta en la categoría "1", pero menor en "0" y "2".
- Recall alto en la categoría "2", pero menor en "0" y "1".
- F1-score alto en la categoría "2", pero menor en "0" y "1".
- Precisión y recall bajos en general.

Vecinos más Cercanos:

- Precisión y recall similares en todas las categorías, un poco más bajos.
- F1-score balanceado en todas las categorías, también un poco más bajo.

Árbol de Decisión:

- Precisión, recall y F1-score perfectos en todas las categorías.
- Posible indicio de sobreajuste en el conjunto de entrenamiento.

Análisis Discriminante Lineal (LDA):

- Buen rendimiento en todas las métricas y categorías.

Precisión y recall muy altos.

Naive Bayes:

- Precisión alta en todas las categorías, especialmente en "1".
- Recall más bajo en la categoría "1", balanceado en otras.
- F1-score alto en todas las categorías, ligeramente más bajo en "1".

Bosques Aleatorios:

- Precisión, recall y F1-score perfectos en todas las categorías.
- Posible sobreajuste en el conjunto de entrenamiento.

Dado que buscamos un equilibrio entre precisión y recall, especialmente en la categoría "1" que representa el ranking de popularidad medio, el modelo de Análisis Discriminante Lineal (LDA) se destaca al mostrar un buen rendimiento en todas las métricas y categorías. Esto lo convierte en una elección sólida para predecir el ranking de popularidad en las categorías alta, media y baja.

Basándonos en la búsqueda de un equilibrio entre precisión y recall, especialmente en la categoría "1" que representa el ranking de popularidad medio, hemos seleccionado el modelo de Análisis Discriminante Lineal (LDA). Este modelo demuestra un buen rendimiento en todas las métricas y categorías, lo que lo convierte en una elección sólida para predecir el ranking de popularidad en las categorías alta, media y baja.

```
Reporte de clasificación (Análisis Discriminante Lineal):
              precision    recall  f1-score   support

           0       0.98      0.97      0.97     27292
           1       0.98      0.97      0.97     30813
           2       0.97      0.98      0.97     59216

    accuracy                           0.97    117321
   macro avg       0.98      0.97      0.97    117321
weighted avg       0.97      0.97      0.97    117321
```

Conclusiones

Tras entrenar y evaluar nuestro modelo LDA, obtuvimos resultados prometedores en términos de precisión, recall y F1-score en las tres categorías de popularidad.

Logramos una precisión general del 97%, lo que indica que nuestro modelo es capaz de predecir con alta exactitud el ranking de popularidad de una canción.

Además, el análisis de las métricas de cada categoría mostró que el modelo tiene un buen equilibrio entre precisión y recall, lo que es esencial para nuestro objetivo de identificar canciones que tengan un ranking de popularidad medio.

En resumen, mediante la elección de un enfoque de clasificación adecuado y la identificación de variables clave, pudimos desarrollar un modelo que puede proporcionar información valiosa para la toma de decisiones en la industria musical.