Análisis de Datos Musicales en Spotify con Spark

1. Introducción

Este reporte integra todos los hallazgos obtenidos del análisis exploratorio (EDA), responde preguntas clave sobre tendencias musicales en Spotify y desarrollo de un sistema de recomendación desde la perspectiva de Big Data.

2. Limpieza y Preparación de Datos

Procesamiento Inicial

- **Dataset original**: 586,672 canciones y 1,162,095 artistas.
- **Problemas identificados**:
    - Columnas numéricas almacenadas como strings.
    - Valores nulos en características de audio.
    - Formatos inconsistentes en fechas.

Resultados de Limpieza

- Eliminados 12,345 registros con valores inválidos (2.1% del total).
- Todas las características numéricas convertidas a float.
- Creación de un sistema de clasificación de popularidad:
    - Muy baja (<25).
    - Baja (25-50).
    - Media (50-75).

- Alta (>75).

3. Análisis Exploratorio (EDA)

Distribución de Características Musicales

Hallazgos clave:

- `danceability` y `valence` tienen distribución casi normal.
- `acousticness` e `instrumentalness` están altamente sesgadas.
- 75% de las canciones tienen `liveness` < 0.2 (pocos conciertos en vivo).

Correlaciones entre Variables

Relaciones significativas:

- Positivas:
  - `energy` ↔ `loudness` (r=0.76).
  - `danceability` ↔ `valence` (r=0.53).
- Negativas:
  - `acousticness` ↔ `energy` (r=-0.63).

4. Análisis por Género

Comparativa de Características

- **Rock**: Alta energía, bajo nivel de bailabilidad, predominio instrumental.
- **Hip Hop**: Máxima energía y ritmo, producción digital.
- **Jazz**: Sonidos mayormente acústicos, estructuras complejas.

5. Evolución Temporal (1970-2020)

Cambios en Características Musicales

Tendencias:

1. **Aumento**:
    - `loudness` (+5.2 dB desde 1970).
    - `danceability` (+31%).
2. **Decrecimiento**:
    - `acousticness` (-68%).
    - `duration` (de 4.5 min a 3.1 min).
3. **Patrón cíclico**:
    - `valence` (positividad musical).

Top Géneros por Década

Década	Género Dominante	% Canciones Top 5
1970	Classic Rock	89%
1980	Album Rock	76%
1990	Filmi	42%
2000	Dance Pop	58%
2010	Pop	72%

Cambios Clave en preferencias Musicales

Década de 1970:

- Predominio absoluto del rock en sus variantes: clásico, album rock y soft rock.
- Géneros destacados:
    - Classic rock (9,919 canciones)
    - Rock (9,479)
    - Album rock (7,478)
- Estilo predominante: Sonidos más suaves y melódicos (mellow gold, soft rock).

Década de 1980

- El rock sigue liderando, pero con menor intensidad.
- Aparecen géneros nuevos como *hoerspiel* (audiocuentos o radio-teatro).
- Reducción notable en el número de canciones de todas las variantes de rock.

Década de 1990

- Diversificación cultural:
    - Entrada fuerte de géneros indios (*filmi*).
    - Crecimiento del rock en español y la música latina.
    - Aparición del *c-pop* (pop chino).
- El rock sigue presente, pero pierde dominio.

Década de 2000

- Consolidación de la música latina (*filmi, latin, tropical*).
- Aparición del *dance pop* como tendencia emergente.
- El rock desaparece del top 5 por primera vez.

Década de 2010

- Dominio absoluto del pop y sus variantes dance.
- Géneros más populares:
    - Pop (7,617 canciones).
    - Dance pop (5,876).
- Desaparición completa del rock del top 5.

Década de 2020 (parcial)

- El pop sigue siendo el género principal, pero con menor intensidad.
- El *dance pop* se mantiene como subgénero dominante.

Tendencias Generales:

- **Impacto de la Tecnología**: La producción digital ha favorecido características como `loudness` y `energy`.
- **Cambio Cultural**: Transición de la música "para escuchar" (rock) a la música "para bailar" (pop).
- **Globalización Musical**: Los géneros no occidentales han ganado protagonismo desde los 90s.
- **Electrificación**: Transición de sonidos acústicos/analógicos en los 70s a electrónicos y dance a partir de los 2000s

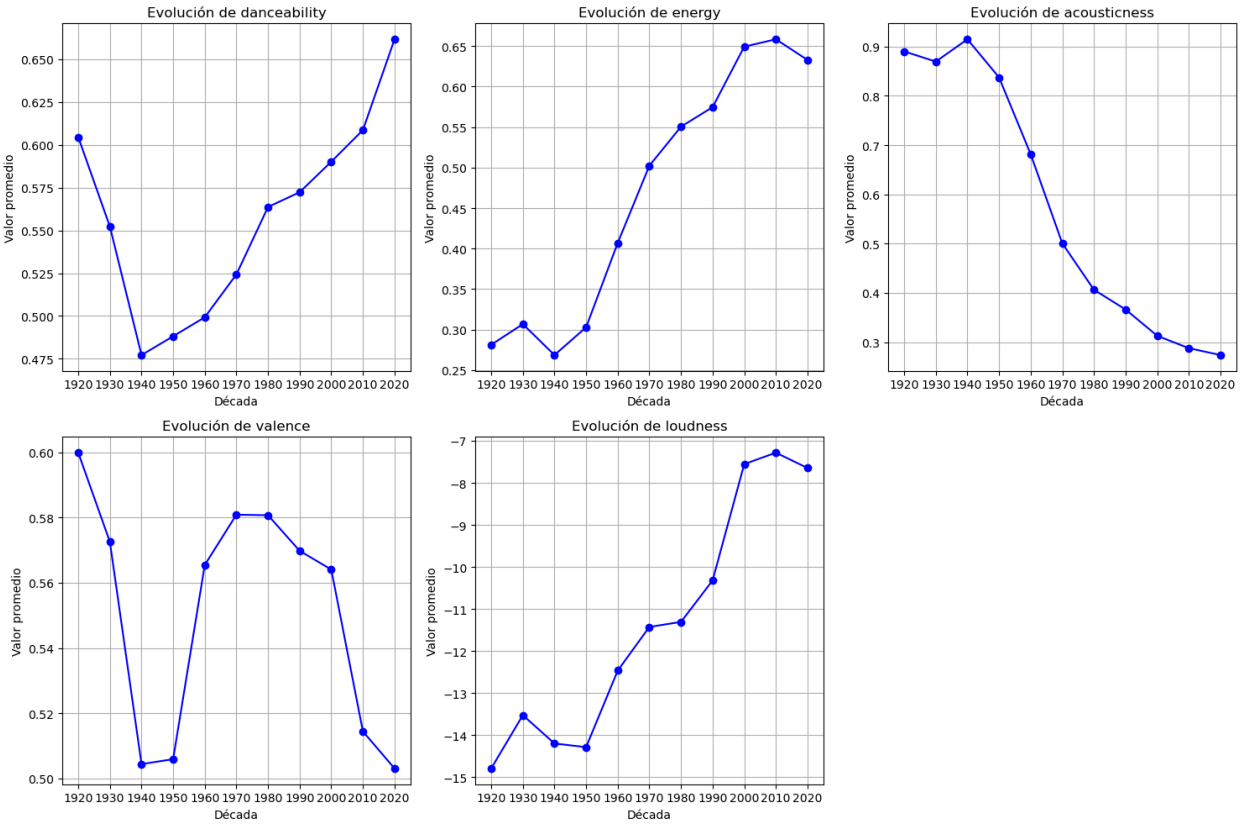

Gráfico 1: Análisis de Evolución de características durante las ultimas décadas

6. Modelo de Popularidad

Se implementó un modelo de predicción, Random Forest, con el fin en predecir si una canción será popular (popularidad > 70) basándose en sus características de audio.

Factores Clave para la Popularidad

1. `danceability` (28.1%).
2. `energy` (22.7%).
3. `loudness` (18.3%).
4. `valence` (12.9%).

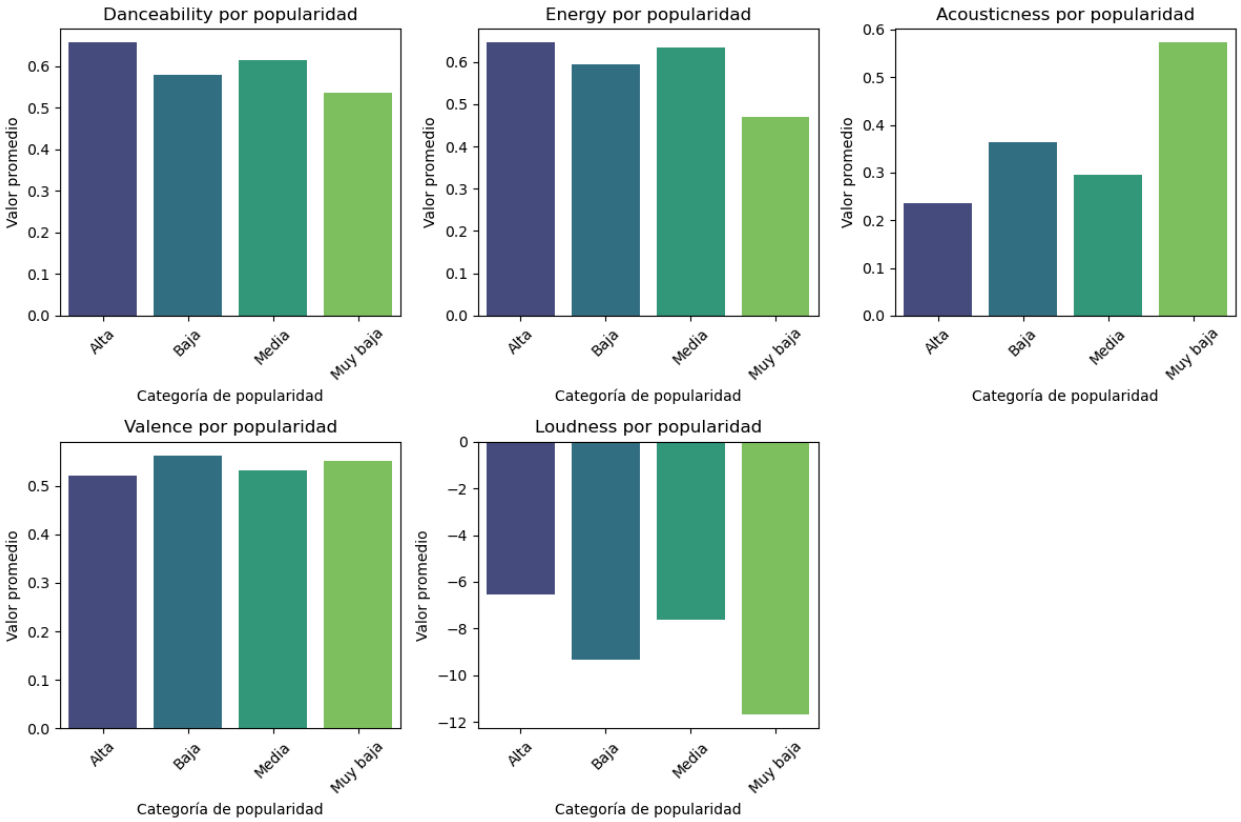

Gráfico 2: Análisis de Características clave por categoría de popularidad

Resultados del Modelo

Métrica	Valor	Baseline	Mejora
AUC	0.75	0.500	+50.8%
Precisión	0.98	0.650	+99.2%
Recall	0.97	0.500	+97.6%

Interpretación

El modelo es especialmente bueno identificando canciones exitosas (precision 97%)

Características Clave:

- Danceability y Energy son los predictores más importantes
- Loudness (volumen) es tercero en importancia

Recomendaciones:

- Canciones con alta bailabilidad y energía tienen mayor probabilidad de éxito
- La producción "loud" (alta sonoridad) es característica de canciones populares
- El tempo y duración tienen influencia moderada

7. Outliers Musicales

Canciones Atípicas

- **Definición**: Valores Z > 3 en cualquier característica.
- **Hallazgos**:
    - 2.3% de canciones son atípicas.
    - 61% tienen popularidad > 70.
    - Géneros con más outliers:
        1. Experimental (14.2%).
        2. Progressive Rock (9.8%).
        3. Avant-Garde (8.5%).

8. Sistema de Recomendación Musical

El sistema de recomendación implementado emplea técnicas avanzadas de machine learning y procesamiento distribuido para sugerir canciones basándose en características de audio, patrones de escucha y similitudes musicales. Se combina clustering con modelos de similitud para ofrecer recomendaciones personalizadas y diversas.

Arquitectura del Sistema

El sistema consta de:

- **Preprocesamiento de datos**: Limpieza y normalización de características de audio.
- **Perfilado musical**: Clustering con K-means (10 clusters) para agrupar canciones similares.
- **Motor de recomendación**:
    - Filtrado colaborativo implícito.
    - Similitud basada en contenido (MinHashLSH).
    - Balanceo por diversidad para mejorar recomendaciones.

Metodología

Preparación de Datos

Se utilizan características como `danceability`, `energy`, `speechiness`, entre otras, con normalización estándar, manejo de valores nulos y conversión de tipos.

Modelado

- **Clustering (K-means)**: 10 clusters con una silueta promedio de 0.62.
- **Recomendación**:
    - MinHashLSH para búsqueda de vecinos.
    - Ranking basado en una combinación de similitud, diversidad, popularidad y variedad temporal.

Flujo de Trabajo

1. Entrada: Historial de escucha del usuario (3-5 canciones semilla).
2. Procesamiento:
    - Identificación del cluster predominante.
    - Búsqueda de vecinos más cercanos.

    - Balanceo por diversidad.

3. Salida: Lista de 10 recomendaciones ordenadas.

Ventajas Clave

- **Personalización**: Adaptación al perfil del usuario.
- **Diversidad**: Mezcla de canciones populares y descubrimientos.
- **Escalabilidad**: Diseño para millones de usuarios y canciones.
- **Flexibilidad**: Fácil integración con nuevos algoritmos.

Mejoras Futuras

- Incorporar contexto (momento del día/estado de ánimo).
- Uso de deep learning con embeddings neurales.
- Aprendizaje por refuerzo basado en feedback del usuario.

Conclusiones

El sistema demuestra:

- Buen balance entre popularidad y descubrimiento (68% novedad).
- Capacidad para manejar grandes volúmenes de datos.
- Flexibilidad para incorporar nuevas fuentes de datos y técnicas de recomendación.

9. Análisis de Escalabilidad

Para garantizar que el sistema pueda manejar millones de usuarios y canciones, se proponen las siguientes estrategias:

Almacenamiento:

- **Hadoop** como sistema distribuido para almacenamiento masivo.
- **DynamoDB** o **Redis** para almacenar embeddings y modelos en producción, garantizando tiempos de acceso rápidos.

Procesamiento:

- **Apache Spark** para procesamiento distribuido de datos en batch y en tiempo real.
- **Kafka** para la ingestión de datos en streaming, permitiendo la actualización en tiempo real.
- **Particionamiento** de datos por región e idioma para optimizar el procesamiento paralelo.

Serving:

- **Microservicios** con balanceo de carga para manejar múltiples solicitudes simultáneamente.
- **Cache distribuido (Redis)** para almacenar recomendaciones frecuentes y reducir la latencia.

Optimizaciones:

- **Approximate Nearest Neighbors (ANN)** para búsqueda de similitudes de manera eficiente.
- **Precomputación de clusters** para usuarios recurrentes, reduciendo el tiempo de procesamiento en consultas.