Data Science Bootcamp

SPRINT 11 EVALUACIÓN FINAL
PREDICCIÓN DE POPULARIDAD MUSICAL BASADO EN SPOTIFY

OBJETIVO:

"Construir y evaluar modelos de aprendizaje automático para aplicaciones específicas en la industria de la música y el entretenimiento".

1.- INTRODUCCIÓN

¿Por qué he elegido este proyecto basado en predicción musical de spotify?
Me pareció muy interesante, ya no solo por ser una usuaria habitual de esta aplicación y una amante de la música sino también por utilizar datos tan fiables como la de esta fuente, puesto que es dinámica, en el sentido que cada día hay usuarios utilizandola y cambiando las métricas y es una aplicación que ya de base usa alogaritmos muy potentes.
De ahí a tener un interés más a nivel técnico y de actualidad en relación a una afición que tengo, como es la música.
A continuación adjunto más información sobre qué es Spotify y su atractivo.

¿Qué es Spotify?
Spotify es un servicio de transmisión de música en línea que ha revolucionado la forma en que las personas consumen y disfrutan de la música. Desde su lanzamiento en 2008, Spotify ha ganado popularidad a nivel mundial y se ha convertido en una plataforma líder en la industria de la música digital.

Spotify es una plataforma que ofrece acceso a un vasto catálogo de música, podcasts y contenido de audio, permitiendo a los usuarios reproducir, descubrir y compartir sus canciones favoritas de manera fácil y conveniente.
A través de la tecnología de transmisión, los usuarios pueden acceder a millones de pistas de música sin necesidad de descargar archivos, brindando una experiencia musical instantánea y accesible.

¿Por qué es Atractivo?
Su atractivo radica en varios factores clave que abordan las necesidades y preferencias cambiantes de los oyentes modernos.

- Variedad y Diversidad Musical:

Spotify ofrece un catálogo musical extremadamente amplio que abarca géneros, artistas y épocas, permitiendo a los usuarios explorar y descubrir nuevos estilos y artistas.

- Acceso en Cualquier Momento y Lugar:

La accesibilidad es clave en Spotify. Los usuarios pueden disfrutar de su música favorita en cualquier momento y lugar, ya sea en casa, en el trabajo o mientras se desplazan, gracias a la capacidad de reproducción en dispositivos móviles y computadoras.

- Recomendaciones Personalizadas:

Spotify utiliza ALOGARITMOS avanzados para analizar los hábitos de escucha de los usuarios y proporcionar recomendaciones personalizadas. Esto facilita el descubrimiento de nueva música que se ajusta a los gustos individuales.

2.- OBTENCIÓN DE LOS DATOS, FUENTE:

El conjunto de datos lo obtube de la página de **KAGGLE** (https://www.kaggle.com/datasets/spoorthiuk) y cuenta con dos archivos CSV extraidos mediante el **API de Spotify** mismo.

- Una colección completa de las **10 CANCIONES** más populares de cada uno de los 10.000 artistas musicales más escuchados en los Estados Unidos en el año 2023.
Este conjunto de datos cubre una amplia gama de géneros musicales y abarca datos del año 2023, capturando las preferencias dinámicas de los entusiastas de la música en el país.
Cada entrada incluye detalles como el nombre del artista, títulos de las canciones, fechas de lanzamiento y clasificaciones de popularidad basadas en métricas (como recuentos de transmisiones o posiciones en las listas).
- El segundo archivo contiene detalles de los 10.000 mejores **ARTISTAS** de EE. UU. disponibles en Spotify. Esto incluye detalles como popularidad, género, edad, popularidad y país.

3.- OBJETIVOS DEL ANALISIS:

El objetivo general del proyecto es **crear un modelo que permita predecir la canción TOP 10 de la lista de artistas musicales.**

Los objetivos específicos para conseguirlo son:

- Limpiar y analizar el conjunto de datos;
- Eliminar del dataset aquellos datos que no sean fiables o interfieran al analisis.
- EDA: analisis exploratorio de los datos
- Machine Learning: Entrenar un conjunto de modelos para que puedan hacer la predicción.
- Evaluar el rendimiento de cada modelo, para posteriormente encontrar los parámetros del mejor modelo; y hacer nuevas predicciones con los modelos.
- Sacar conclusiones y sugerencias de mejora.

4.- CONJUNTO DE DATOS

El conjunto de datos elegido para el proyecto consiste en dos datasets obtenidos de la fuente “Kaggle” y que proceden de la extracción de la base de datos mediante un API de Spotify. Son dos bases de datos referentes a reproducciones en Estados Unidos actuales, año 2023

- el primer dataset ‘Artist’, contiene información detallada sobre varios artistas, incluidos aspectos como género, edad, país, géneros musicales, popularidad y seguidores.

- el segundo dataset ‘Top songs’, contiene información detallada sobre canciones, álbumes y artistas. Incluye información sobre la reproducibilidad, si la canción es explícita (con contenido explícito, sexual, agresivo, machista,....), la duración de la canción, etc.

5.- EXPLICACIÓN CONCRETA DE LAS VARIABLES

- DEL DATAFRAME “Artists”:

Número de Filas: 9488
Número de Columnas: 9

FUENTE: www.kaggle.com

Contiene información detallada sobre varios artistas, incluidos aspectos como género, edad, país, géneros musicales, popularidad y seguidores.

Columnas y sus Descripciones:

- Name: Nombre del artista (objeto).
- ID: Identificación del artista (objeto).
- Gender: Género del artista (objeto), encontramos 4 y algunos valores faltantes.
- Age: Edad del artista (entero). Edades de 0 a 146 años, por lo que hay errores.
- Country: País del artista (objeto). De todo el mundo, aunque falta un 37% de datos.
- Genres: Géneros musicales asociados al artista (objeto), como pop, rock, latino...
- Popularity Artist: Puntuación de popularidad del artista (entero), del 1 al 99
- Followers: Número de seguidores del artista (entero).
- URI: Identificación única de recursos asociada al artista (objeto).
-

Observaciones:
Algunas columnas tienen valores nulos, como "Gender" y "Country".

- DEL DATAFRAME “Top Songs”:

Número de Filas: 37146
Número de Columnas: 16

FUENTE: www.kaggle.com

Contiene información detallada sobre canciones, álbumes y artistas. Incluye información sobre la reproducibilidad, si la canción es explícita y la duración de la canción.

Columnas y sus Descripciones:

- Album Type: Tipo de álbum (objeto). Se refiere a si és un “Single”, un álbum o una compilación.
- Artist ID: Identificación del artista (objeto).
- Artist Name: Nombre del artista (objeto).
- Artist Song Rank: Rango de la canción del artista (entero). Dentro de su mismo ranking.
- Track Name: Nombre de la canción (objeto).
- Is Playable: Indicador de si la canción es reproducible (booleano).
- Album Name: Nombre del álbum (objeto).
- Release Date: Fecha de lanzamiento del álbum (objeto).
- Total Album Tracks: Número total de pistas en el álbum (entero).
- Is Explicit: Indicador de si la canción es explícita. Se refiere a si incluye lenguaje fuerte, contenido sexual, violencia o temas controvertidos.(booleano).

- ISRC: Código de grabación estándar internacional (objeto).
- Song Duration: Duración de la canción en milisegundos (entero).
- Track Number: Número de la pista en el álbum (entero).
- Popularity Song: Puntuación de popularidad de la canción (entero). Del 1 al 99
- Track Id: Identificación de la pista (objeto).
- Track URI: Identificación única de recursos asociada a la pista (objeto).