UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Music4U

Sistema de Recomendación

Personalizado Solo Para Ti

Equipo #3

1729734 De la Garza Fernández, Carla Mayela

1625654 Zamarrón Medrano, Alan

Grupo 012

Martes y Jueves de 18:30 a 20:00

Minería de Datos

Profra. Mayra Cristina Berrones Reyes

Spotify Dataset 1921-2020, 160k+ Tracks

https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

Descripción de los Datos

Los datos con los que cuenta la base de datos son tablas. Usaremos como tabla principal la tabla “data.csv” que contiene más de 160,000 canciones recopiladas de la API Web de Spotify, pero también usaremos las tablas que están divididas por artista (“data_by_artist.csv”), año (“data_by_year.csv”) y género (“data_by_genres.csv” y “data_w_genres”).

En la tabla “data.csv” encontramos 19 columnas que se dividen en tres categorías: datos numéricos, datos extras y datos categóricos. Las columnas que usaremos son:

Llave primaria

- ***id***: Número de identificación generado por Spotify

Datos numéricos

- ***danceability***: Número flotante. Qué tan bailable es la canción. Puede tener un valor entre 0 y 1.
- ***popularity***: Número entero. Valor de popularidad. Puede tener un valor entre 0 y 100.

Datos extras

- ***explicit***: Valor booleano. 0 = sin contenido expícito, 1 = contenido explícito.

Datos categóricos

- ***artists***: String. Nombre del artista.
- ***release_date***: En formato fecha yyyy-mm-dd. Fecha de lanzamiento de la canción.
- ***name***: String. Nombre de la canción.

La tabla “data_by_artist.csv” cuenta con 15 columnas de las cuales solo usaremos:

- ***artists***
- ***popularity***
- ***danceability***
- ***count***: Número entero. Número total de canciones en la tabla “data.csv” producidas por este artista.

La tabla “data_by_genres.csv” cuenta con 14 columnas de las cuales solo usaremos:

- ***genres***: String. Nombre del género.
- ***danceability***
- ***popularity***

La tabla “data_by_year.csv” cuenta con 14 columnas de las cuales solo usaremos:

- ***year***: Número entero. Cada fila representa un año de 1921 a 2020.
- ***danceability***
- ***popularity***

La tabla “data_w_genres.csv” cuenta con 16 columnas de las cuales solo usaremos:

- ***artists***
- ***danceability***
- ***popularity***
- ***count***
- ***genres***: Lista con valores String. Contiene los géneros de cada artista.

Justificación del Uso de Datos

Decidimos utilizar esta base de datos porque los datos son muy completos y porque pudimos plantearnos varios objetivos al momento de decidir qué objetivo final utilizar. Otro aspecto muy importante para nosotros es que en la página de Kaggle muestran que esta base de datos tiene una calificación de 10.0 de usabilidad. Además de que cuenta con suficientes datos para nuestro proyecto.

También, resulta muy interesante que podamos aplicar la ciencia, las matemáticas y el aprendizaje de máquina para analizar música, un aspecto común y casi indispensable en nuestro día a día.

A diferencia de otras bases de datos también encontradas en Kaggle, esta particularmente cuenta con columnas que son muy útiles, tales como año de lanzamiento, artista, nombre, popularidad, género.

Planteamiento del Problema

Music4U es una empresa de streaming de música. Nuestros usuarios disminuyeron mucho (abandonaron la plataforma) por las nuevas alternativas que ofrecían una plataforma personalizada para el usuario.

Objetivo Final

Realizar un sistema de recomendación que sea personalizado para los usuarios de tal forma que mejore la experiencia de usuario.

Objetivo Secundario

Predecir qué artistas y qué géneros de música lanzarán la música más popular en el año siguiente, en base a los datos con lo que ya contamos.

Planeación de la Herramienta a Utilizar

Se utilizará **árboles de decisión y clustering** para obtener los artistas, géneros y canciones que puedan ser de interés para el usuario, así como la realización de una **predicción** para la recomendación de una lista de canciones que estén en el mismo rango de interés del usuario.

Sistema de recomendación: Herramienta que establece un conjunto de criterios y valoraciones sobre los datos de lo usuarios para realizar predicciones sobre recomendaciones de elementos que puedan ser de interés para el usuario. Desarrollaremos un sistema de recomendación basado en contenido, ya que con datos del historial del usuario vamos a predecir qué busca el usuario y qué sugerencias similares se le pueden mostrar.

Clustering: Utilizaremos esta técnica para dividir los datos en clústeres. Los clústeres estarán divididos según sus características (tales como género, artistas, popularidad, año). Se generarán recomendaciones basadas en el clúster especifico que es similar al usuario que necesita una recomendación. Para saber en qué clúster se encuentra un usuario se tomarán ciertos datos de entrada.

Árbol de decisión: Modelo predictivo que basado en los valores de entrada predice un valor de salida. Cada nodo del árbol corresponde a un atributo (un género, una canción, un artista, un año) y cada arco que va del nodo padre al nodo hijo representa un posible valor para ese atributo. Se reciben los datos de entrada (una canción, un género, un artista, un año) y se van dividiendo para que cada nodo hijo sean solo los atributos que coincidan con esos valores. El proceso se repite de manera recursiva hasta que ya no se pueda dividir.