Análisis Avanzado de Datos - Práctica Final

Dario, Muñoz Muñoz (NIA: 100405982)

16 de marzo de 2023

I. Introducción

Spotify es una plataforma de streaming de música que ofrece a sus usuarios acceso a millones de canciones de diversos géneros y épocas. Para mejorar la experiencia de los usuarios, Spotify utiliza la inteligencia artificial (IA) para analizar las características acústicas y la popularidad de las canciones, y así poder ofrecer recomendaciones personalizadas, listas de reproducción y radios. El conjunto de datos Spotify Dataset 1921-2020, 600k+ Tracks recoge información sobre más de 600 mil canciones publicadas entre 1921 y 2020, incluyendo datos como el nombre, los artistas, la fecha de lanzamiento, la duración, la presencia de contenido explícito y varios atributos acústicos como la bailabilidad, la energía o el tempo. En este trabajo se pretende explorar este conjunto de datos utilizando dos tipos de aprendizaje automático: el supervisado y el no supervisado. El primero se basa en utilizar datos etiquetados para entrenar un modelo que pueda predecir la salida o la clase de nuevos datos. El segundo se basa en utilizar datos sin etiquetar para descubrir patrones ocultos o agrupaciones en los datos. El objetivo es aplicar ambos tipos de aprendizaje para obtener información relevante sobre las canciones y sus características.

Para el aprendizaje supervisado, nos centraremos en predecir la popularidad de las canciones a partir de sus atributos acústicos. La popularidad es un valor numérico entre 0 y 100 que indica el grado de éxito o reconocimiento de una canción en Spotify.

Para el aprendizaje no supervisado, nos centraremos en agrupar las canciones por su danzabilidad a partir de sus atributos acústicos. La danzabilidad es una categoría que describe el grado de facilidad para bailar la canción.

II. Dataset

El dataset escogido es de Yamac Eern AY y se llama "Spotify Dataset 1921-2020, 600k+ Tracks". Contiene 600 mil canciones lanzadas entre 1921 y 2020. Este dataset puede ser usado para tareas de aprendizaje supervisado y no supervisado, como clasificación, regresión o agrupamiento. El dataset contiene dos archivos, uno llamado "tracks.csv"tiene 20 columnas con los siguientes datos:

- **id**: el identificador único de cada canción en Spotify (cadena alfanumérica).
- **name**: el nombre de la canción (cadena de texto).
- **popularity**: el nivel de popularidad de la canción en una escala de 0 a 100 (número entero).
- **duration_ms**: la duración de la canción en milisegundos (número entero).
- **explicit**: si la canción tiene contenido explícito o no (0 = no, 1 = sí) (número binario).
- **artists**: los artistas que participan en la canción (lista de cadenas de texto).
- **id_artists**: los identificadores únicos de los artistas en Spotify (lista de cadenas alfanuméricas).
- **release_date**: la fecha de lanzamiento de la canción en formato YYYY-MM-DD (cadena de texto).
- **danceability**: el grado de facilidad para bailar la canción en una escala de 0 a 1 (número decimal).
- **energy**: el nivel de energía e intensidad sonora de la canción en una escala de 0 a 1 (número decimal).
- **key**: la tonalidad musical principal de la canción en una escala cromática estándar que va del 0 al 11, donde el 0 es Do y el 11 es Si (número entero).
- **loudness**: el nivel promedio del volumen sonoro en decibeles (dB) durante toda la canción (número decimal negativo o cero).
- **mode**: el modo musical principal de la canción, donde el 0 es menor y el 1 es mayor (número binario).
- **speechiness**: el grado en que se habla o se canta durante la canción en una escala del 0 al 1, donde valores cercanos al cero indican más música instrumental y valores cercanos al uno indican más voz humana (número decimal).
- **acousticness**: el grado en que se usan instrumentos acústicos o eléctricos durante la canción en una escala del 0 al uno, donde valores cercanos al cero indican más instrumentos eléctricos y valores cercanos al uno indican más instrumentos acústicos (número decimal).
- **instrumentalness**: el grado en que se usan instrumentos musicales o voces humanas durante la canción en una escala del cero al uno, donde valores cercanos al cero indican más voces humanas y valores cercanos al uno indican más instrumentos musicales(número decimal).
- **liveness**: el grado en que se percibe que hay público presente durante la grabación o reproducción de la canción en una escala del cero al uno, donde valores cercanos al cero indican menos presencia pública y valores cercanos al uno indican más presencia pública(número decimal).
- **valence**: el grado en que se transmite un sentimiento positivo o negativo durante la canción en una escala del cero al uno, donde valores cercanos al cero indican sentimientos más negativos como tristeza o ira y valores cercanos al uno indican sentimientos más positivos como alegría o euforia(número decimal).
- **tempo**: el ritmo promedio expresado como las pulsaciones por minuto(BPM) durante toda la canción(número decimal).
- **time_signature**:el número promedio expresado como fracciones simples(4/4 ,3/4 , etc.)de las pulsaciones por compás musical durante toda la canción(número entero).

Este dataset es muy útil para analizar las tendencias musicales a lo largo del tiempo y las preferencias del público. También puede servir para generar recomendaciones personalizadas o crear nuevas composiciones musicales basadas en los atributos acústicos.

III. Análisis del Dataset

Para realizar un buen análisis de datos del Dataset, primero mostraremos la matriz de correlación y la explicaremos. La matriz de correlación nos permite ver la relación lineal entre las variables numéricas del Dataset. Algunas de las correlaciones más destacadas son:

- La variable loudness tiene una alta correlación positiva con la variable energy (0.76), lo que indica que a mayor intensidad sonora, mayor nivel de energía percibida en las canciones.
- La variable valence tiene una moderada correlación positiva con la variable danceability (0.53), lo que sugiere que a mayor grado de positividad o alegría expresada en las canciones, mayor facilidad para bailarlas.
- La variable acousticness tiene una alta correlación negativa con la variable energy (-0.72) y una moderada correlación negativa con la variable loudness (-0.52), lo que implica que a mayor presencia de elementos acústicos en las canciones, menor nivel de energía y menor intensidad sonora.

Estos resultados nos permiten conocer mejor las características y relaciones entre las variables del Dataset y nos ayudan a entender mejor las características de las canciones del Dataset y a identificar posibles patrones o tendencias.

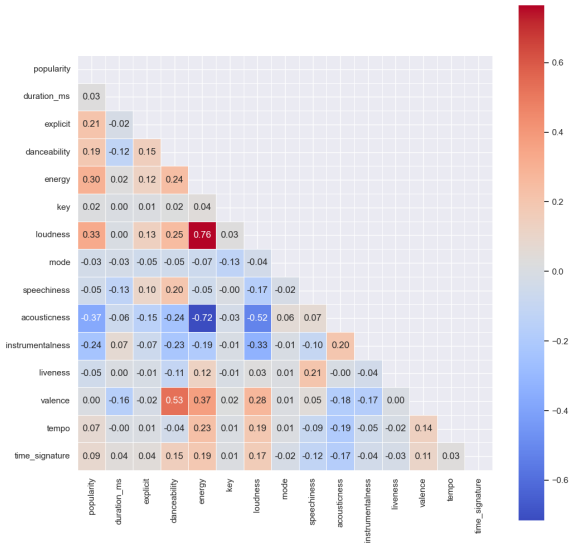

Figura 1. Matriz de correlación

También para analizar el dataset realizaremos un histograma en la Figura2, de cada variable que vamos a predecir, es decir popularidad y danzabilidad. Esto para ver cómo están distribuidas estas variables en el dataset.

También se ha realizado una comparación entre variables representativas como son la popularidad y la intensidad sonora, en la Figura3. en este se puede ver que tienen más popularidad las canciones con mayot intensidad sonora.

Por último también se ha realizado una comparación entre las variables bailabilidad y cuanto se habla en una canción en la Figura4.

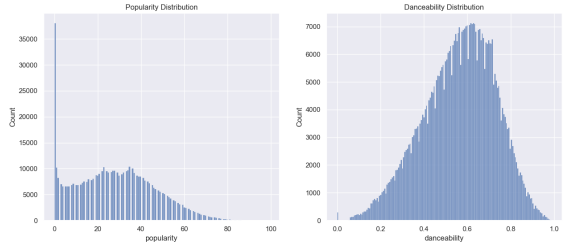

Figura 2. Histograma de popularidad y danzabilidad

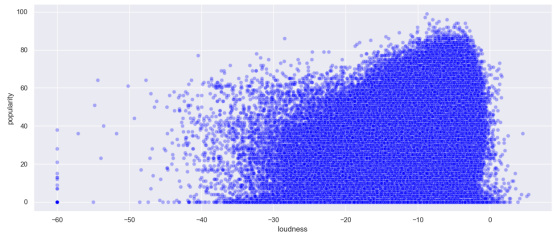

Figura 3. Relación de popularidad e intensidad sonora

Viéndose que la mayoría de las canciones con mayor nivel de danzabilidad tienen muy poca locuencia.

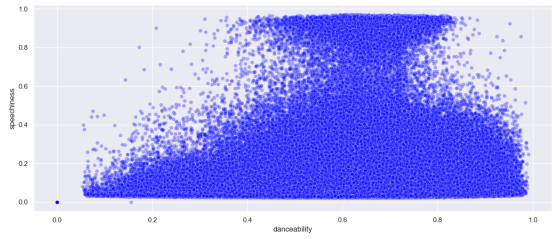

Figura 4. Relación de danzabilidad y locuencia

IV. Método

Primero se realizará un preprocesado de los datos eliminando algunos inservibles como los id de la canción. Y se dividirá el dataset en dos, uno para entrenamiento y otro para test.

1. Aprendizaje supervisado

Para el aprendizaje supervisado se realizará mediante técnicas de regresión. Una vez entrenado el modelo, este se evaluará con el conjunto de test para obtener diferentes métricas que sirvan como comprobación de que el modelo está funcionando correctamente.

2. Aprendizaje no supervisado

Respecto al aprendizaje no supervisado se aplicaran técnicas de clustering y de reducción de dimensionalidad. El clustering nos ayudará a ver cómo está segmentado el mercado y en que se fijan los consumidores y con la reducción de dimensionalidad conseguiremos hacer un modelo más pequeño sin tantas variables pero con una eficacia muy parecida.