Clasificación de Géneros

0:46 — 3:11

UANL UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Introducción

Spotify es un servicio de música y podcasts en streaming que da acceso a millones de canciones y otros contenidos de artistas de todo el mundo.

Spotify®

Vemos la música como una manera de expresión, es más que solo escuchar y cantar, poder transmitir tu estado de ánimo con tus gustos musicales, sentir el poder de la energía que transmite, es una experiencia única.
Sabemos que constantemente los géneros musicales se están reinventando, a tal punto de obtener fusiones inesperadas, es por esto que necesitamos entender y conocer que es lo que en verdad influye al momento de registrar un género musical, así mismo se realizará un análisis puntual para revisar cuales son las métricas necesarias para obtener y categorizar un género musical.

Objetivos

Objetivo principal:
Clasificar las canciones por los géneros que se encuentran en esta base de datos, haciendo uso de la relación que existe con la energía que proyecta, lo fácil que es bailarla, lo ruidosa que es y qué estado de ánimo te transmite.

Objetivo secundario:
Implementar una extensión de búsqueda inteligente tomando como punto focal el BPM de las canciones que se encuentran en la base de datos.

Recursos

Python
Lenguaje de programación en el que fue desarrollado el proyecto.

Jupyter
Consola para ejecutar el lenguaje de Python, la interfaz que brinda es muy amigable.

Base de datos
La información fue tomada de kaggle “Spotify - All Time Top 2000s Mega Dataset” Este conjunto de datos contiene estadísticas de audio de las 2000 pistas principales en Spotify. Los datos contienen aproximadamente 15 columnas, cada una de las cuales describe la pista y sus cualidades.

Librerías

Datos
- Numpy
- Pandas
- Sklearn

scikit learn machine learning in Python

Visualización
- Graphviz
- MatIplotlib
- Seaborn

Metodología

Se hizo una modificación en la base de datos, reclasificando los subgéneros en géneros generales ej.(dutch rock, canadian rock, australian rock = rock), esto para tener una mayor eficiencia al momento de clasificar las canciones.
Con la limpieza nos dimos cuenta de ciertos factores que debemos a tomar en cuenta al momento de construir el árbol, como lo es, que a mayor sea la energía más fuerte es la canción, a cuanto más positivo sea el estado de ánimo más fácil será bailarla y que a cuanto mayor valor enérgico tenga la canción, más positivo será el estado de ánimo.

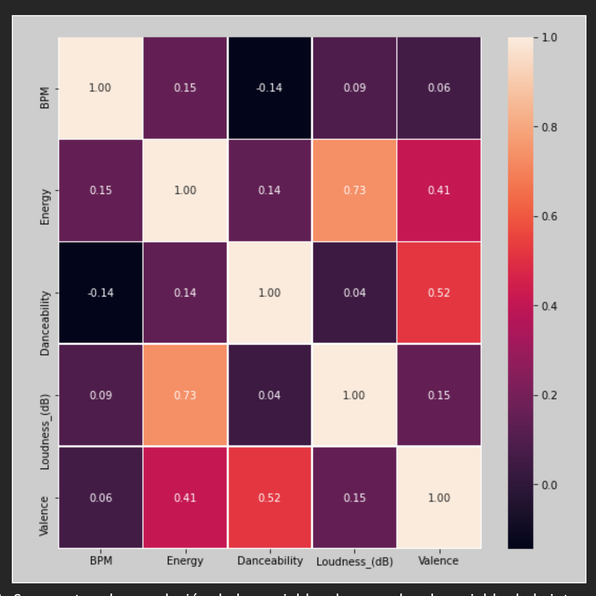

Figura 1: Se muestran la correlación de las variables de acuerdo a la variable de la intersección.

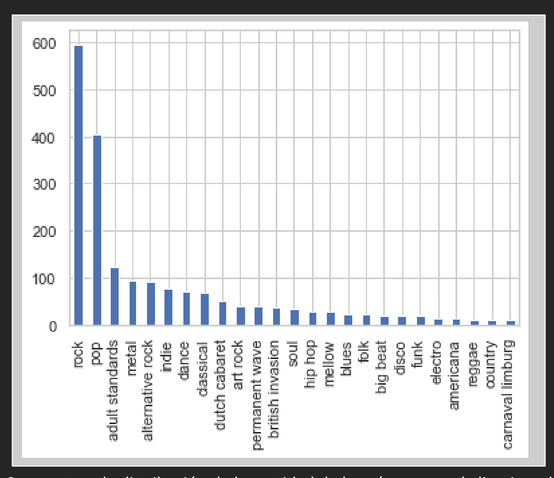

Figura 2: Se muestran la distribución de la cantidad de los géneros con la limpieza de datos.

Resultados

Para obtener los resultados de clasificación, se obtuvo el peso de estas variables predictoras: "BPM", "Energy", "Danceability", "Loudness", "Valance", "Acousticness" y "Speechiness“. Con la finalidad de ir buscando patrones para realizar futuras evaluaciones en las predicciones utilizando la técnica de árboles de decisión.

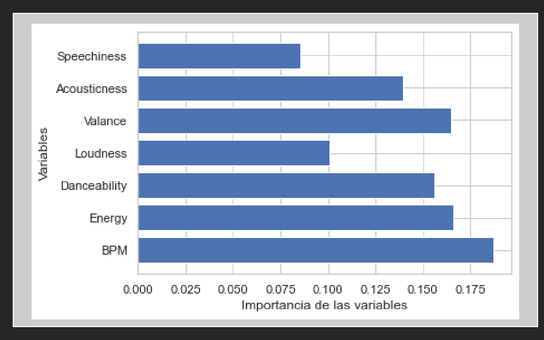

Figura 3: Se muestra el peso de variables en el modelo de árbol de decisión

Con una selección aleatoria utilizamos el 90% de los datos como entrenamiento para la creación del modelo, usando un 10% para probar la efectividad.
El árbol parte del “sample” inicial establecido, dividiendo sus ramificaciones a partir de los 25 géneros en “value” que cumplen con la característica de las variables predictoras, descartando los que no cumplen, dividiéndose hasta clasificarlo en el género correspondiente “class”.
Dando como resultado un aprendizaje del modelo del 40% de efectividad y una predicción del 99.89%, esto quiere decir que, de 80 canciones, clasifica correctamente 32.

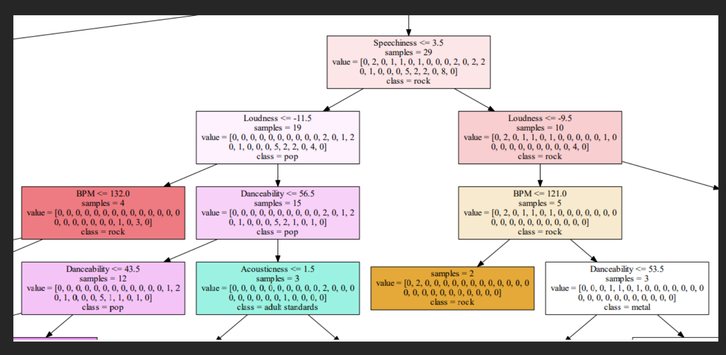

Figura 4: Se muestran un fragmento del diagrama de árbol de decisión, se puede observar completo en el código QR del diagrama. (primer objetivo)

	Title	Top_Genre	BPM
82	Modern World	indie	120
120	Papa Was A Rolling Stone	adult standards	120
131	Talk	permanent wave	120
183	Heavy Cross	dance	120
196	Supermassive Black Hole	rock	120
322	Behind Blue Eyes	metal	120
340	It's My Life	metal	120
402	Alors on danse	pop	120

Figura 5: Se muestra un fragmento del buscador por BPM. (segundo objetivo)

Para la implementación de la búsqueda por medio del BPM se utilizó la base de datos tal cual, filtrando por título y género, obteniendo una estimación del 100%

Conclusión

Se cumplieron los objetivos con una exactitud en un 40% y 100% respectivamente, obteniendo que el “BPM”, “Energy” y “Valance” son las características de más peso al momento de clasificar una canción.

Trabajo a futuro
Concluyendo el proyecto nos percatamos de que los árboles de decisión no son el modelo adecuado para el primer objetivo del proyecto, ya que tienen la tendencia de “sobre ajustar”, esto quiere decir, que tienden a aprender muy bien los datos de entrenamiento con un 99.89% de predicción, pero su generalización no es tan buena con un 40% de efectividad, por lo que concluimos que un modelo más apto serían los Bosques aleatorios, estos hacen que distintos árboles vean distintas porciones de los datos, esto hace que cada árbol se entrene con distintas muestras de datos para un mismo problema, de esta forma, al combinar sus resultados, unos errores se compensan con otros y se puede tener una predicción que generalice mejor.

Diagrama

Autores:
Romero Pascacio M. J.
Vázquez Bocanegra M. J.
Velázquez Rivera S.

Base de datos