Universidad del Valle de Guatemala
Facultad de ingeniería

Data Science
Resultados Parciales - Proyecto
Sistema de recomendación de Spotify

Pablo Gonzalez
Javier Mombiela
Jose Hernandez
Jun Woo Lee
Andrés de la Roca

Guatemala, 2023

Índice

Introducción

En un mundo donde la música se ha convertido en un componente sumamente importante dentro de nuestras vidas, con la capacidad de acceder a una infinidad de música a través de las diferentes plataformas de streaming, esto se ha convertido en más que una simple comodidad. Las diferentes plataformas de streaming de música han desempeñado un papel destacable en ofrecer diferentes métodos para que los usuarios puedan acceder a su música preferida en donde quiera que estén.

La proliferación de estos servicios, junto con el crecimiento de sus librerías de canciones y artistas, se ha presentado un desafío intrigante: ¿Como podemos hacer para que los usuarios encuentren la música que verdaderamente les gustara? La respuesta a esta pregunta se encuentra en los sistemas de recomendación, uno de los acercamientos más poderosos que han surgido como parte del desarrollo de estos sistemas es la implementación de modelos de Machine Learning.

Por medio de este proyecto se busca explorar el uso de los modelos de Machine Learning para el desarrollo de un sistema de recomendación que en base a datos de listas de reproducción recopiladas por usuarios de la plataforma de Spotify pueda ayudarnos a resolver el desafío planteado, ofreciendo predicciones/recomendaciones de lo que le posiblemente le pueda gustar escuchar a partir del nombre de un artista.

El objetivo principal de este informe es mostrar el planteamiento inicial del modelo y aplicaciones planteada a partir de la problemática a resolver, se mostraran los objetivos en los que se basa este proyecto, se mencionara la teoría estudiada para elaborar este proyecto así como información importante que se tomó en cuenta al momento de analizar los datos utilizados. Adicionalmente se mencionara la metodología usada para desarrollar esta primera iteración de la solución y se presentarán resultados que puedan demostrar la eficiencia y precisión del modelo utilizado para realizar la solución

Objetivos

Objetivo General

- Proponer un sistema de recomendación efectivo y preciso que le ofrezca recomendaciones valiosas a los usuarios de Spotify por medio de Machine Learning.

Objetivos Específicos

- Reconocer patrones en los gustos musicales que ayuden a predecir la relevancia de las recomendaciones que se le hagan a los usuarios
- Identificar qué variables representan mejor los gustos musicales de los usuarios.
- Relacionar de manera precisa las canciones con las características que buscan los usuarios en su música preferida.
- Desarrollar un modelo de manera iterativa para encontrar la mejor combinación de parámetros para el sistema de recomendación

Marco Teórico

Para empezar, el Machine Learning es un campo de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las máquinas aprender patrones a partir de datos y tomar decisiones o hacer recomendaciones, como es en este caso que se quieren realizar recomendaciones de artistas a partir la información encontrada acerca de las listas de reproducción de Spotify creadas por usuarios.

Los sistemas de recomendación son una parte fundamental en la experiencia de usuario en las aplicaciones de streaming de música, el tipo de análisis que realizan estos sistemas suelen variar de aplicación a aplicación, incluso hasta dentro de una misma plataforma suelen haber diferentes sistemas que realizan análisis diferentes. Para este proyecto el análisis del sistema de recomendación desarrollado tiene un enfoque para que el modelo tome en cuenta la información de las playlists creadas, a partir de esto observar la frecuencia con la que se repiten ciertos artistas y/o canciones y a partir de esto el modelo poder calcular y medir el parentesco de unas canciones con otras o unos artistas con otros.

Respecto al conjunto de datos utilizado para entrenar al modelo se puede mencionar que este consiste en aproximadamente 10,000 playlists públicas de Spotify que han sido creadas entre enero 2010 y noviembre 2017. Cada playlist contiene datos como; el nombre de la playlist y la lista de canciones, la lista de canciones contiene información específica referente a cada una de las canciones, a continuación se profundizará más acerca de qué significa cada uno de los campos del conjunto de datos.

El proceso de análisis exploratorio de datos es fundamental en la investigación, ya que permite comprender la naturaleza y las relaciones en un conjunto de datos. En este proyecto, se llevaron a cabo tareas críticas de limpieza y preprocesamiento de datos en un archivo JSON llamado 'challenge_set.json'. Se extrajeron detalles relevantes de cada playlist, como la posición, nombre del artista, URI del track, URI del artista, nombre del track, URI del álbum, duración en milisegundos y nombre del álbum. Además, se identificaron y gestionaron las listas de tracks vacías, asegurando la integridad de los datos.

Como se mencionó anteriormente el conjunto de datos consta de 10,000 listas de reproducción aproximadamente y muestra una gran variación en términos de la cantidad de "holdouts", canciones, "samples" y la duración de las canciones. Esto refleja la diversidad de preferencias de los usuarios, con algunas listas siendo más extensas y musicalmente ricas que otras.

El análisis de la matriz de correlación reveló relaciones significativas entre variables. Se encontró una fuerte correlación positiva entre la cantidad de "holdouts" y la cantidad de pistas en una lista, lo que sugiere que las listas más diversas tienden a tener más pistas. Además, se observó una correlación positiva entre la cantidad de pistas y la cantidad de "samples", lo que podría estar relacionado con la riqueza musical de las listas. La correlación casi perfecta entre

la cantidad de "samples" y la duración de las canciones indica una relación importante entre la cantidad de datos de audio y la duración de las canciones.

El análisis de varianza (ANOVA) mostró que la variable categórica 'name' tiene un efecto significativo en las variables numéricas, incluyendo la cantidad de "holdouts", pistas, "samples" y la duración de las canciones. Esto sugiere que las categorías de 'name' influyen en las características numéricas de los datos y puede ser crucial para comprender cómo se relacionan con las preferencias de los usuarios.

El análisis exploratorio de datos realizado en este proyecto proporciona información valiosa para comprender las preferencias de los usuarios en las listas de reproducción, las relaciones entre las variables y el impacto de las categorías categóricas. Estos hallazgos pueden guiar investigaciones futuras y la toma de decisiones relacionadas con la música y el análisis de datos de audio.

Metodología

Antes realziar el análisis y procesamiento de datos, se establecieron las herramientas y bibliotecas esenciales. Se eligió pandas por su eficacia en la manipulación y análisis de datos. La biblioteca json se incorporó para manejar archivos en este formato, dado que el conjunto de datos principal estaba estructurado así. Para el análisis visual, se integraron las bibliotecas matplotlib.pyplot y seaborn, ambas esenciales para obtener insights visuales y representaciones gráficas.

En el lado del frontend, se utilizó un template "mantine-vite-template", que proporciona una estructura base para el desarrollo. Las principales dependencias incluyen "@mantine/core" para componentes de interfaz, "axios" para solicitudes HTTP y "react-router-dom" para la navegación. El proyecto también incluye Storybook para documentar y probar componentes de React de manera aislada.

En cuanto al backend, se empleó Flask para crear el servidor web y se incorporó CORS para manejar las solicitudes de origen cruzado. Se utilizó Keras para cargar modelos y TensorFlow para operaciones relacionadas con redes neuronales. Además, se integraron bibliotecas como joblib y scikit-learn para la manipulación y el procesamiento de modelos.

En la segunda etapa, se cargaron y exploraron inicialmente los datos utilizando el archivo challenge_set.json. Con la ayuda de pandas, se transformó este archivo JSON en un DataFrame, facilitando su manipulación y exploración estructurada.

La tercera etapa se centró en el procesamiento y transformación de datos. Aunque el dataset inicial fue informativo, se requirió una organización más detallada. Se puso un énfasis particular en las pistas individuales de cada playlist. Se extrajeron detalles como el nombre del artista y la URI del tema. Tras esta extracción, los datos se estructuraron y se

reincorporaron al DataFrame principal. Durante esta etapa, se crearon diagramas específicos para visualizar la distribución de características como duration_ms y num_samples. Además, se elaboró una matriz de correlación para identificar posibles relaciones entre las diferentes características del dataset.

En cuanto a la división de datos, el grupo optó por una estrategia común de división de conjuntos de entrenamiento y prueba. Esta elección permitió entrenar los modelos utilizando una porción significativa de los datos, mientras que se reservó un subconjunto para evaluar su desempeño en datos no vistos.

Finalmente, la elección de los algoritmos fue un aspecto crucial del proceso. Se seleccionaron basándose en la naturaleza y estructura de los datos, así como en los objetivos específicos del proyecto. Las redes neuronales, por ejemplo, se eligieron por su capacidad para manejar grandes cantidades de datos y aprender patrones complejos, lo cual es esencial para el análisis de listas de reproducción de música.

En resumen, la metodología adoptada aseguró una comprensión profunda y estructurada de los datos, preparando el terreno para futuros análisis y recomendaciones. La incorporación de visualizaciones específicas, como las relacionadas con duration_ms, num_samples y la matriz de correlación, junto con decisiones informadas sobre la división de datos y la elección de algoritmos, proporcionó una base sólida para el éxito del proyecto.

Resultados y análisis de resultados

El proyecto inició con una exploración del conjunto de datos original para comprender su estructura y características. Este conjunto, contenido en el archivo challenge_set.json, ofrecía una variedad de información relacionada con las listas de reproducción de Spotify. Las variables presentes en el conjunto de datos se describen a continuación:

- Pid: Es el identificador único de la playlist, de naturaleza categórica nominal.
- Name: Representa el nombre de la playlist. En algunas listas, este campo puede estar omitido.
- num_holdouts: Indica el número de canciones que han sido omitidas para esta playlist.
- tracks: Contiene una colección de las canciones presentes en cada playlist.
- pos: Refleja la posición de una canción específica dentro de la playlist.
- track_name: Es el nombre de la canción.
- track_uri: Proporciona el URI específico de la canción en Spotify.
- artist_name: Indica el nombre del artista de la canción.
- artist_uri: Ofrece el URI del artista en Spotify.
- album_name: Representa el nombre del álbum donde se encuentra la canción.
- album_uri: Es el URI específico del álbum en Spotify.
- duration_ms: Muestra la duración de la canción en milisegundos.

- num_samples: Indica el número total de canciones incluidas en la playlist.
- num_tracks: Representa el número total de canciones en la playlist, incluidos los "holdouts".

Esta estructura proporciona una rica fuente de información para el análisis. Ciertas características, como duration_ms y num_samples, se visualizaron para comprender su distribución y tendencias. El análisis reveló que las listas de reproducción en Spotify son altamente personalizadas y variadas. La duración promedio de las canciones es de aproximadamente 3.88 minutos, aunque algunas pistas pueden durar hasta 152.64 minutos.

Además, se realizó un análisis de frecuencia para las variables categóricas. Se descubrió que géneros como "country", "rap" y "oldies" son populares entre los usuarios. Además, ciertos artistas, como Drake, destacaron en términos de menciones, reflejando su popularidad durante ese período.

La fase de carga y exploración inicial transformó este archivo JSON en un DataFrame, utilizando la biblioteca pandas. Esta transformación facilitó la manipulación y exploración de datos, permitiendo una visión más estructurada del conjunto de datos. En la etapa de procesamiento y transformación, se organizó y desglosó la información. Se extrajeron características específicas, como el nombre del artista y la URI del tema, y se crearon visualizaciones para representar la distribución de ciertas características, como duration_ms y num_samples. Además, se elaboró una matriz de correlación para discernir posibles relaciones entre las diferentes variables.

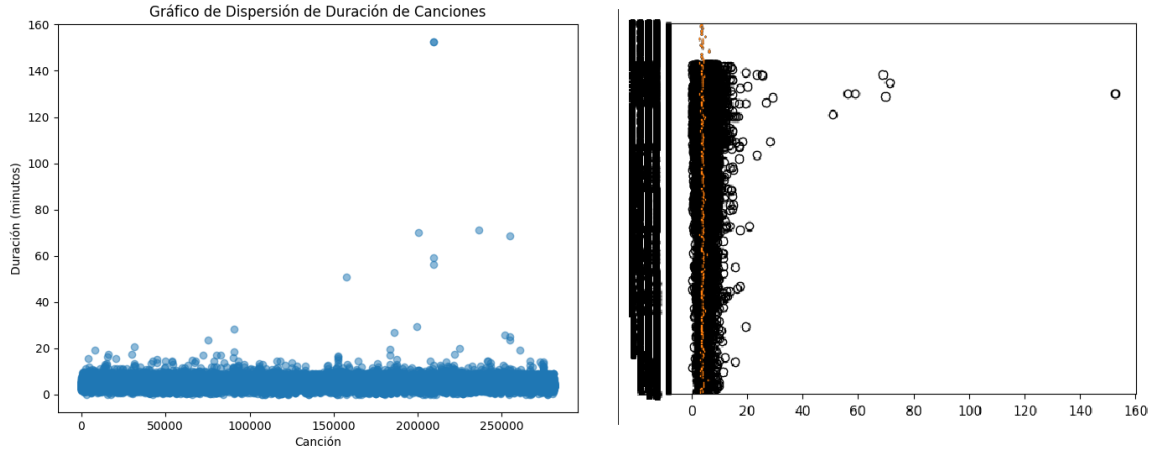

Fig 1. Diagramas sobre Duración de Canciones

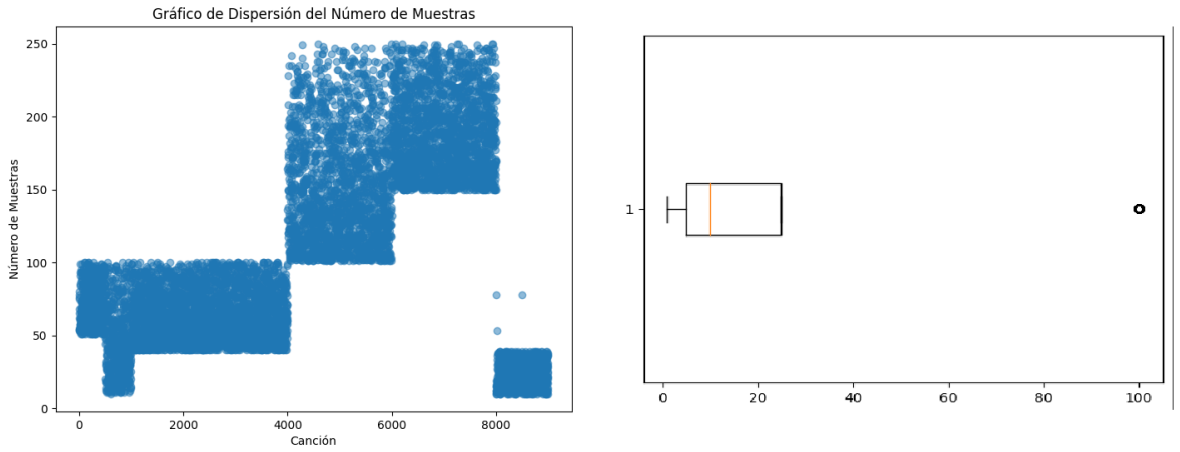

Fig 2. Diagramas sobre Numero de Muestras

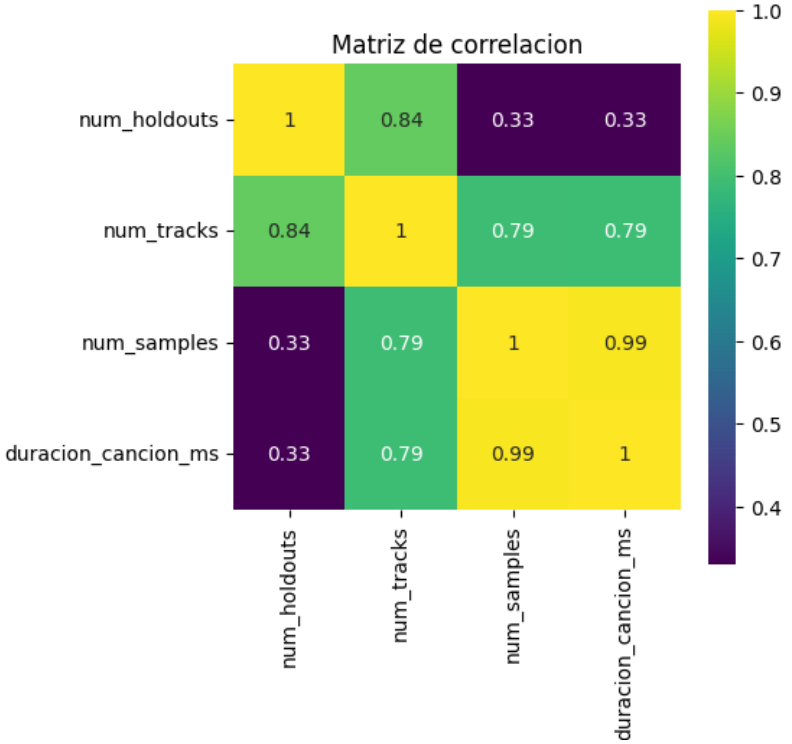

Fig 3. Matriz de Correlación

En cuanto a la selección de conjuntos de datos para entrenamiento y validación, el grupo optó por una división común del 80% para entrenamiento y el 20% para validación. Esta estrategia permitió entrenar los modelos con una porción significativa del conjunto de datos mientras se reservaba un subconjunto para evaluar su desempeño en datos no vistos.

Ambos modelos fueron entrenados utilizando este conjunto de datos. Comparando los dos, el primer modelo demostró un rendimiento superior, alcanzando una precisión del 89.21% en el conjunto de entrenamiento y 89.30% en el conjunto de validación. Por otro lado, el segundo modelo alcanzó una precisión del 82.74% en el entrenamiento y 82.30% en la validación. Estas diferencias en la precisión sugieren que el primer modelo es más adecuado para este

conjunto de datos específico. Además, el primer modelo también presentó una pérdida menor en comparación con el segundo modelo, lo que indica un mejor ajuste a los datos.

La elección de los algoritmos se basó en la naturaleza y estructura del conjunto de datos y en los objetivos específicos del proyecto. Se optó por redes neuronales debido a su capacidad para manejar grandes volúmenes de datos y aprender patrones complejos. Esta elección fue crucial para el análisis de listas de reproducción de música, donde las interacciones y las relaciones entre las variables pueden ser complejas.

En resumen, el proyecto adoptó una metodología estructurada y detallada para garantizar una comprensión profunda del conjunto de datos. La incorporación de visualizaciones y la comparación detallada de los algoritmos, junto con decisiones informadas sobre la división de datos, proporcionaron una base sólida para el éxito del proyecto. La superioridad del primer modelo en términos de precisión y pérdida lo posiciona como la opción más viable para futuras aplicaciones y análisis.

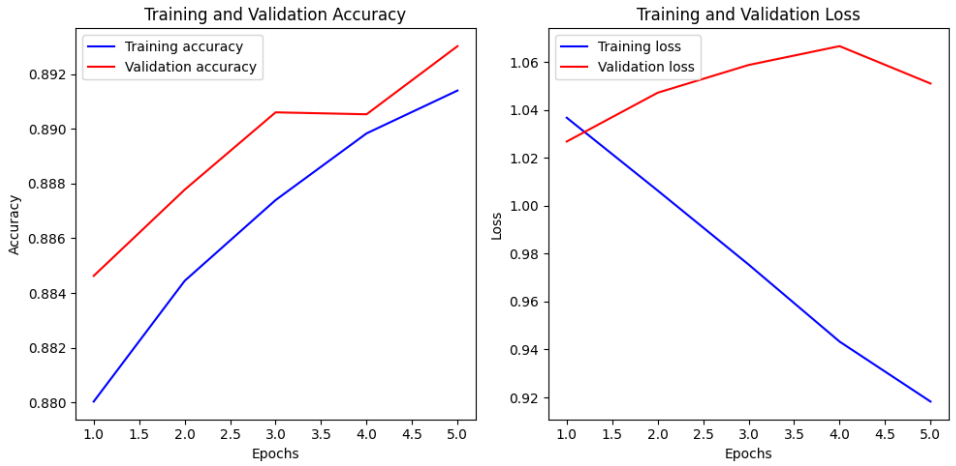

Fig 4. Primer Modelo

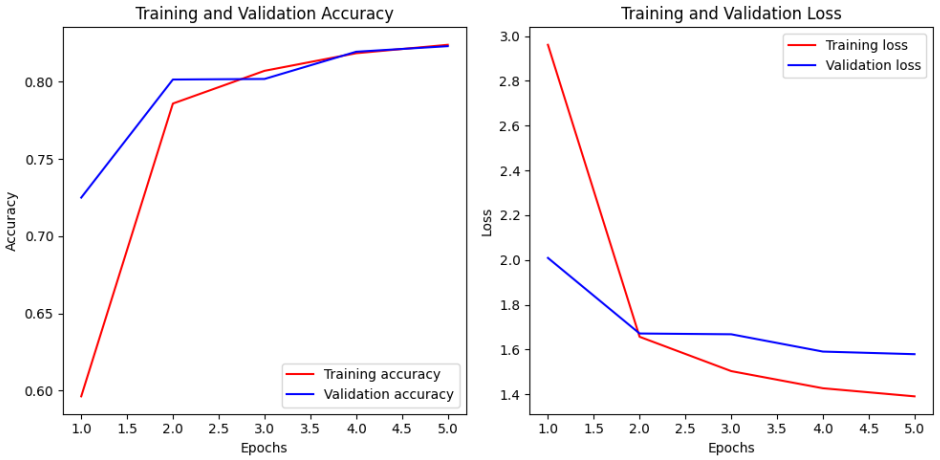

Fig 5. Segundo Modelo

Aplicación

Se creó una aplicación web que tiene tres páginas principales, 1 para simular cada modelo, y la tercera para visualizar las gráficas de los modelos.

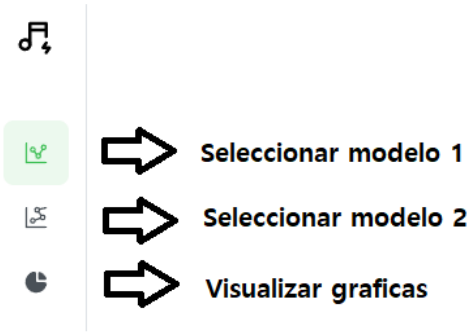

Fig 6. Menú de la aplicación

Al seleccionar un modelo, le manda a una página donde le pide que ingrese el nombre de un artista y luego de buscarlo, regresa con las recomendaciones aportadas por el modelo.

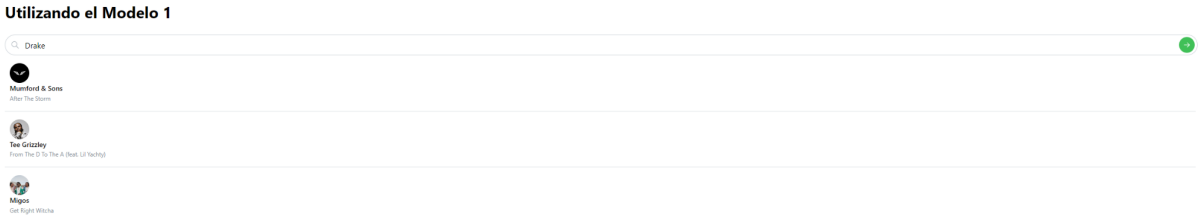

Fig 7. Sistema de recomendación en la aplicación

Referencias bibliográficas

Spotify, AICrowd (2020) Spotify Million Playlist Dataset Challenge. Extraído de AICrowd.

C.W, Chen, P. Lamere, M. Schedl, H. Zamani. (2018) Recsys Challenge 2018: Automatic Music Playlist Continuation, in Proceedings of the 12th ACM Conference on Recommender Systems.

C.W, Chen, P. Lamere, M. Schedl, H. Zamani. (2018) An Analysis of Approaches Taken in the ACM RecSys Challenge 2018 for Automatic Music Playlist Continuation.

Oramas, S., Espinosa-Anke, L., Lawlor, A., Sordo, M., & Saggion, H. (2017). Exploring customer reviews for music genre classification and evolutionary studies. En Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China.

Chollet, F. (2018). Deep learning with Python. Manning Publications Co.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.