Recomendación Musical con características numéricas y procesamiento de lenguaje natural

David Alfonso Barbosa Gómez

December 2024

1 Introducción

El presente proyecto tiene como objetivo desarrollar un sistema de recomendación musical, utilizando diversas técnicas de análisis y ciencia de datos.

En un entorno donde las plataformas de streaming de música han transformado la manera en que los usuarios descubren y consumen contenido, los sistemas de recomendación juegan un papel crucial para ofrecer sugerencias personalizadas y mejorar la experiencia del usuario.

El proyecto se basa en el uso de un conjunto de datos que incluye diversos atributos cuantitativos de las canciones, como el tempo, la energía, la intensidad y la tonalidad, entre otros. Estos atributos permiten caracterizar musicalmente las canciones, lo que facilita la creación de modelos de predicción, para clasificar y recomendar música acorde a los gustos de los usuarios.

Además de las características numéricas, se incorpora el análisis de las letras de las canciones mediante técnicas de procesamiento de lenguaje natural. A través de la extracción de temas o "topics" presentes en las letras, se pretende agregar una dimensión adicional al sistema de recomendación, teniendo en cuenta el contenido emocional y temático de las canciones.

Este enfoque combinado busca proporcionar recomendaciones más personalizadas, alineando tanto los atributos musicales como los emocionales.

2 Trabajos relacionados

Tanto en la industria como en el nicho de investigación existen varios trabajos relacionados con el uso de características musicales, análisis de letras y técnicas de aprendizaje automático para la recomendación de canciones.

Tomando Spotify como ejemplo, usan principalmente dos enfoques:

- Filtrado Basado en Contenido: Este método se enfoca en analizar los metadatos de las canciones, como género, información del artista y características acústicas de las canciones (como el tempo, la energía, la tonalidad, etc.).

- Filtrado Colaborativo: Este enfoque se basa en analizar los hábitos de escucha de los usuarios para encontrar patrones y recomendar canciones que podrían gustar a otros usuarios con gustos similares.

3 Definición de problema y Herramientas

3.1 Definición de problema

En la actualidad, las plataformas de música en línea generan recomendaciones personalizadas basadas en las preferencias de los usuarios. Sin embargo, la mayoría de estos sistemas de recomendación tienden a utilizar únicamente características como las interacciones previas de los usuarios (me gusta, reproducciones, etc.) o las características acústicas de las canciones. Esto puede limitar la capacidad de ofrecer recomendaciones precisas y variadas, especialmente cuando los usuarios buscan nuevas canciones basadas en temáticas específicas o en la letra de las canciones.

El objetivo principal de este proyecto es diseñar un sistema de recomendación musical que combine tanto características numéricas (como que tan bailable es, energía, tempo, etc.) como procesamiento de lenguaje natural (PLN) para analizar las letras de las canciones.

3.2 Herramientas

- Python: El lenguaje de programación principal para la implementación del sistema.
- Pandas y Numpy: Para la manipulación de datos y la gestión de grandes conjuntos de información.
- Scikit-learn: librería de Python que ofrece herramientas para la construcción de modelos de aprendizaje automático.
- Transformers de Hugging Face: Esta librería se utiliza para cargar y usar modelos preentrenados de PLN como BERT.
- Joblib: Para la serialización y carga de objetos.
- Matplotlib/Seaborn: Para visualización de datos, útil en el análisis exploratorio de datos (EDA).
- Google Colab: Entorno de desarrollo que permite ejecutar código Python de manera eficiente en la nube.

4 Evaluación Experimental

4.1 Datos

4.1.1 Conjunto de Datos sobre características acústicas

El conjunto de datos incluye una serie de características numéricas extraídas de las canciones. Estas características son representaciones cuantitativas de diversas propiedades musicales que ayudan a describir el estilo y el género de la canción.

- Danceability: Mide la facilidad con la que una canción puede inducir a la danza, basada en el ritmo y la claridad del ritmo.
- Energy: Refleja la intensidad y el nivel de actividad de la canción.
- Loudness: El volumen promedio de la canción medido en decibelios.
- Tempo: El ritmo de la canción, medido en beats por minuto (BPM).
- Valence: Mide el estado de ánimo general de la canción, indicando si tiene una tonalidad positiva o negativa.
- Speechiness: Indica la cantidad de contenido hablado en una canción.
- Acousticness: Refleja el grado en que una canción es acústica, es decir, si utiliza sonidos naturales o electrónicos.

Extraído de: `https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs`

4.1.2 Conjunto de Datos de Letras

El conjunto de datos incluye las mismas características del conjunto de datos acústicos. Ademas de la letra de la canción, su tema y otras características que no son relevantes para este estudio.

Extraído de: `https://www.kaggle.com/datasets/saurabhshahane/music-dataset-1950-to-2019`

4.1.3 Conjunto de datos relacionado a la salud mental

El conjunto de datos busca explorar posibles correlaciones entre los gustos musicales de los individuos y su salud mental auto informada.

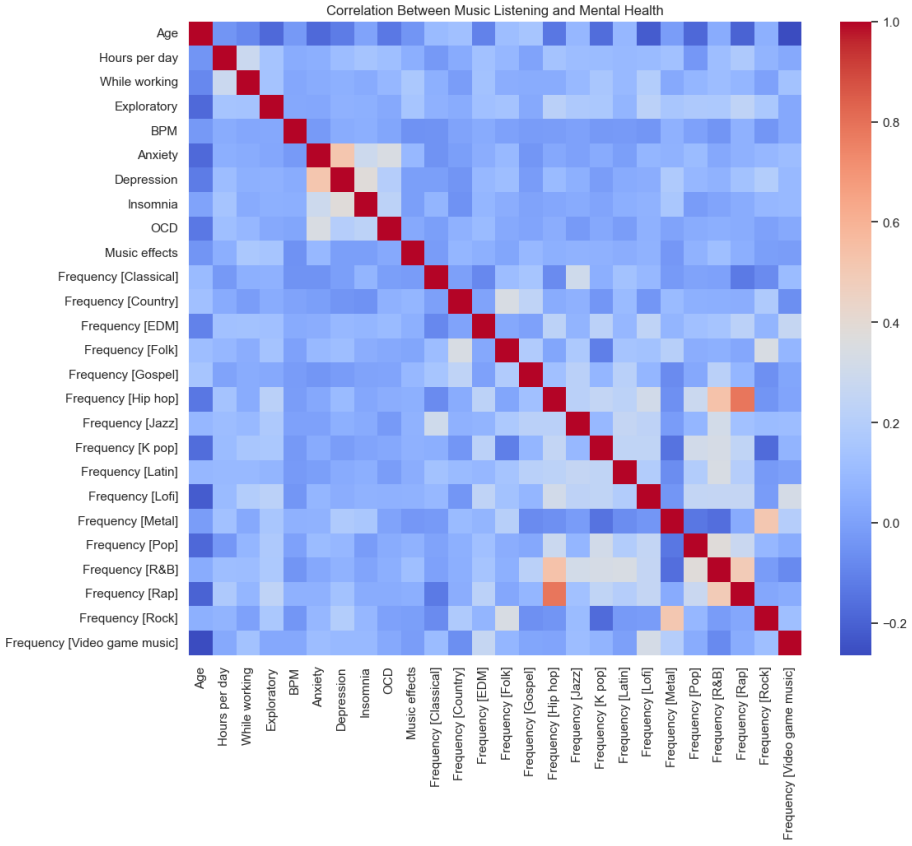
Extraído de: `https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results`

4.2 Metodología

4.3 Preprocesamiento de Datos y Análisis Exploratorio

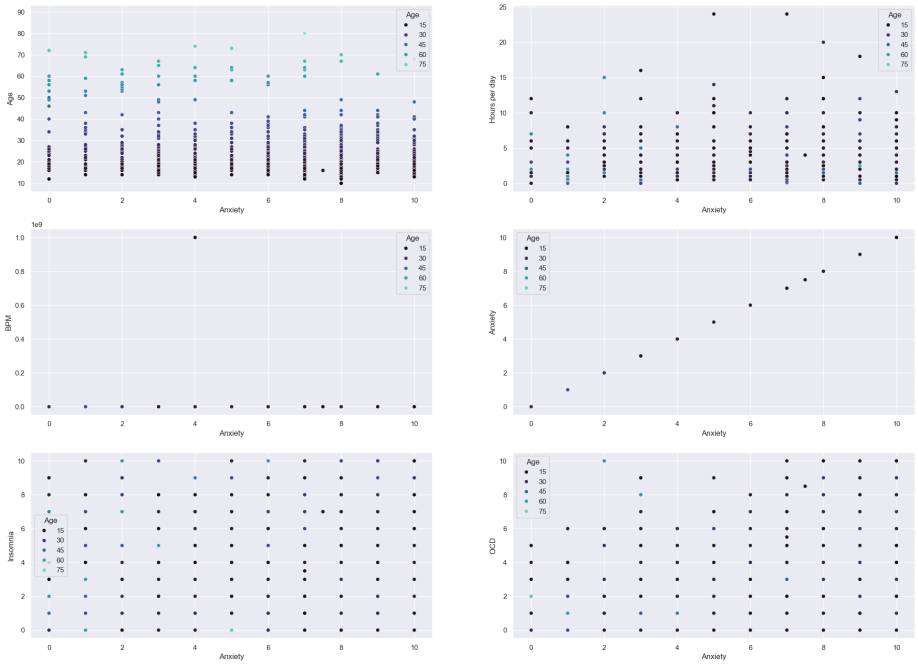
4.3.1 Conjunto de datos relacionado a la salud mental

- Histogramas para observar las distribuciones de los datos, dando resultados curiosos como el siguiente:
- Diagrama de bigotes para observar datos que difieren de la media. Las clases que presentan datos atípicos son las siguientes:
- Análisis de correccional de variables, para observar la correlación entre las variables interesantes.

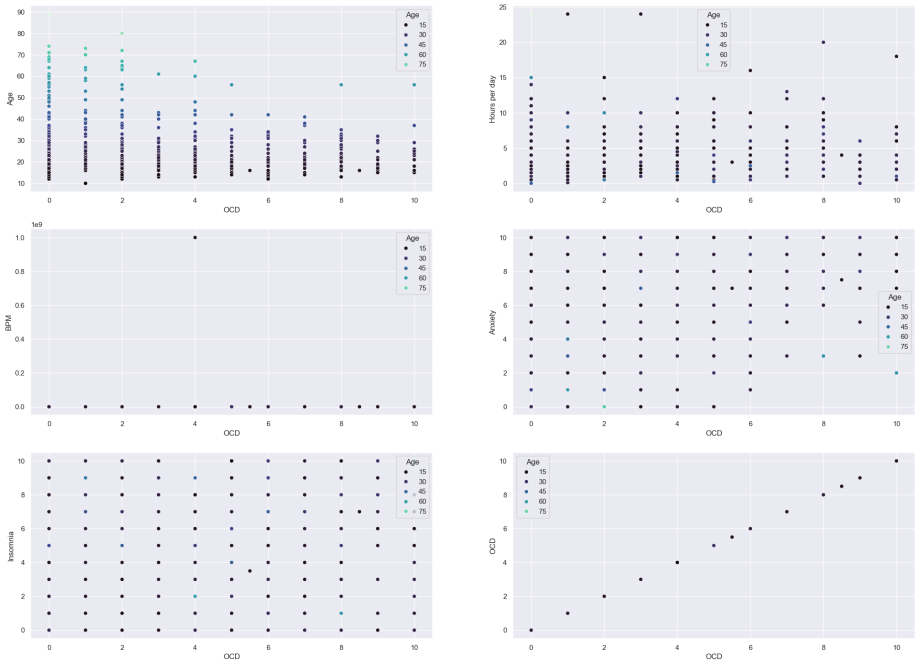


- scatterplot para mostrar las relaciones entre las columnas de interés

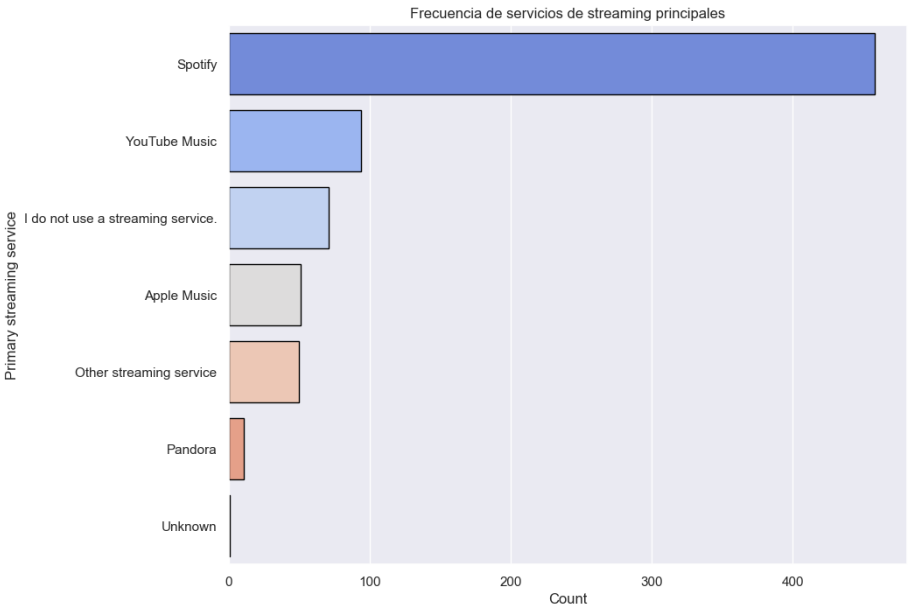
  Ansiedad VS columnas de interes



OCD VS columnas de interes

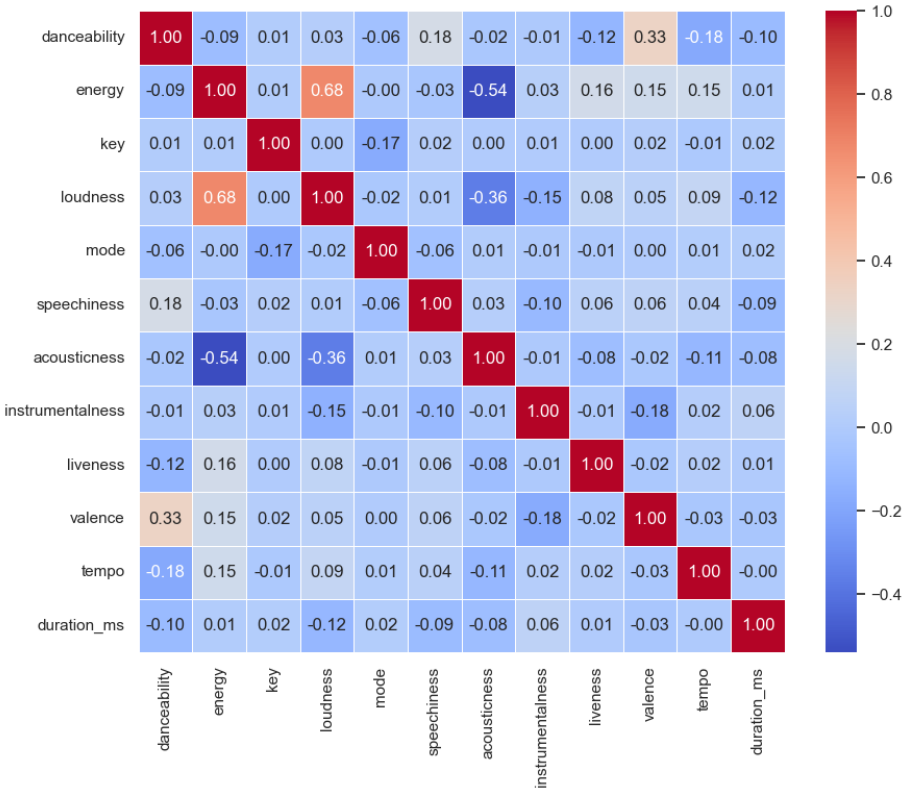


- Gráfico de barras para determinar cual es la aplicación de streaming de música mas usada.



4.3.2 Conjunto de Datos sobre características acústicas

- Histogramas para observar las distribuciones de los datos.
- Análisis de correccional de variables, para observar la correlación entre las variables interesantes.



- Gráfico de barras para determinar si hay desbalance en los datos.
  Diagrama de barras para los géneros:

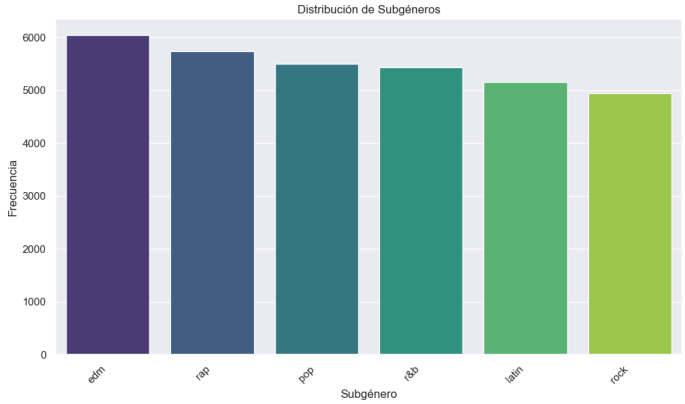
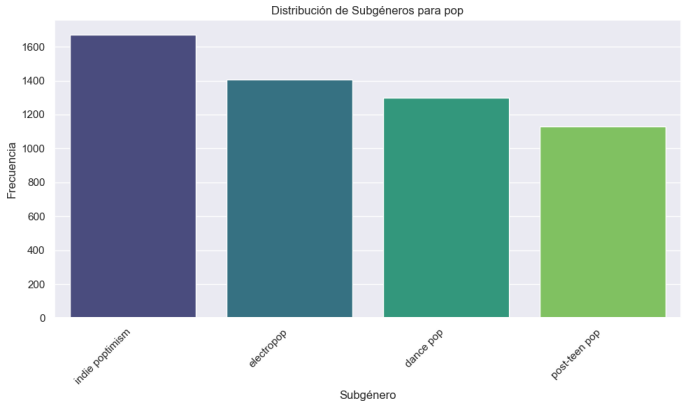
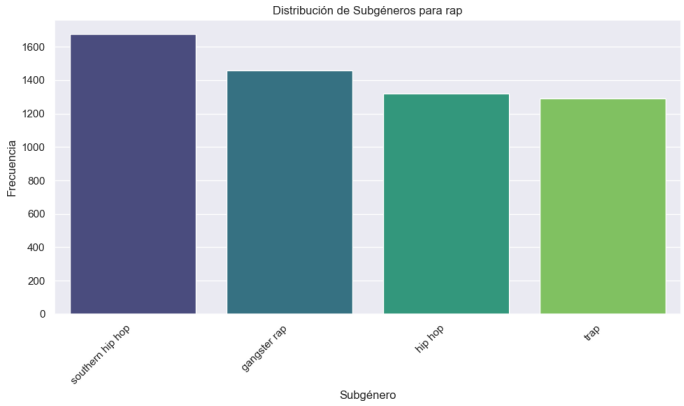
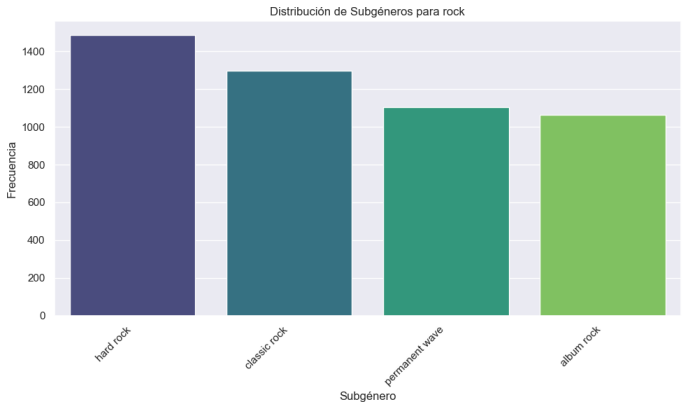
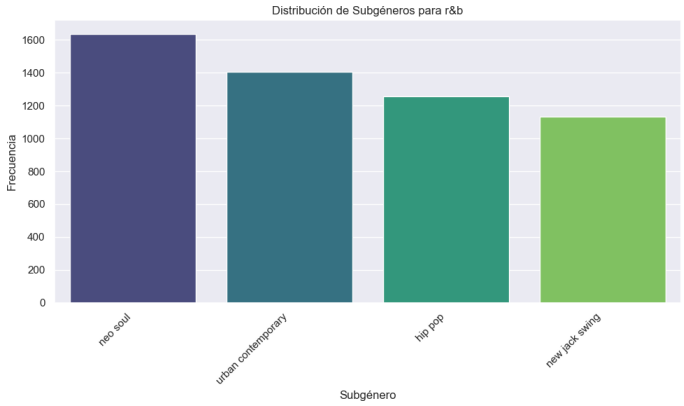


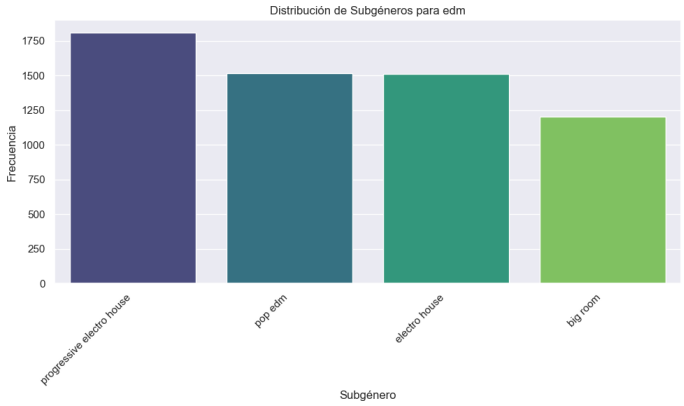
  Diagrama de barras para subgéneros:


Distribución de Subgéneros para pop
Frecuencia
0
200
400
600
800
1000
1200
1400
1600
indie poptimism
electropop
dance pop
post-teen pop
Subgénero


Distribución de Subgéneros para rap
Frecuencia
0
200
400
600
800
1000
1200
1400
1600
southern hip hop
gangster rap
hip hop
trap
Subgénero


Distribución de Subgéneros para rock
Frecuencia
0
200
400
600
800
1000
1200
1400
hard rock
classic rock
permanent wave
album rock
Subgénero





- Gráfico de densidad Kernel.
- Clustering con k-means para agrupar de manera optima los subgéneros

4.3.3 Conjunto de Datos de Letras

Se recortaron los datos de acorde a la frecuencia de estos

4.4 Modelado

4.4.1 Conjunto de Datos sobre características acústicas

Se utiliza un modelo de aprendizaje supervisado (Random Forest Classifier) para predecir el subgénero de las canciones basándose en sus características numéricas.

4.4.2 Conjunto de Datos de Letras

Un modelo preentrenado de lenguaje en este caso *DistilBertTokenizer* identifica el tema principal de las letras.

4.5 Resultados

4.5.1 Conjunto de Datos sobre características acústicas

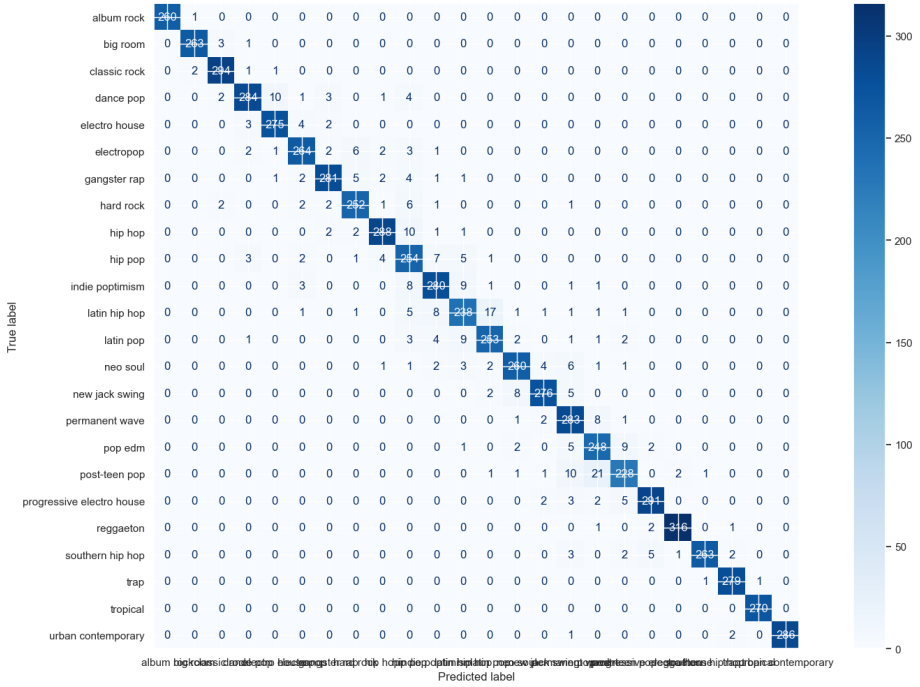
```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       261
           1       0.99      0.99      0.99       267
           2       0.98      0.99      0.98       298
           3       0.96      0.93      0.95       305
           4       0.95      0.97      0.96       284
           5       0.95      0.94      0.94       281
           6       0.96      0.95      0.95       297
           7       0.94      0.94      0.94       267
           8       0.96      0.95      0.96       304
           9       0.85      0.92      0.88       277
          10       0.92      0.92      0.92       303
          11       0.89      0.87      0.88       275
          12       0.91      0.92      0.92       276
          13       0.95      0.93      0.94       281
          14       0.97      0.95      0.96       291
          15       0.88      0.96      0.92       295
          16       0.87      0.93      0.90       267
          17       0.92      0.86      0.89       265
          18       0.97      0.96      0.97       303
          19       0.99      0.99      0.99       320
          20       0.99      0.95      0.97       276
...
weighted avg       0.95      0.95      0.95      6833

MSE: 0.29679496560807844
R²: 0.9937594777488891
```

El modelo tiene un rendimiento sobresaliente en términos de clasificación y regresión.

Ademas de que no presenta confusión relevante a la hora de clasificar los subgéneros.



4.5.2 Conjunto de Datos de Letras

```
Validation Loss: 1.1356577159216006
Validation Accuracy: 0.819327731092437
```

4.6 Combinación de los dos modelos para la recomendación

Algoritmo usado, para la implementación

1. **Seleccionar las características musicales:**
   Se toman características numéricas como "energía", "valencia" o "tempo" de la canción de entrada para procesarlas.

2. **Normalizar las características:**
   Se utiliza un modelo de escalamiento (*scaler*) para ajustar estas características al mismo rango que las del modelo ya entrenado.

3. **Predecir el subgénero:**
   Un modelo de clasificación (*Random Forest Classifier*) predice el subgénero al que pertenece la canción basándose en las características musicales normalizadas.

Se transforma el resultado del modelo (que es un número) al nombre del subgénero usando un *label encoder*.

4. **Filtrar canciones por subgénero:**
   Del conjunto de datos, se seleccionan todas las canciones que pertenecen al subgénero predicho.

5. **Predecir el tema de las letras:**
   Se pasa la letra de la canción a un modelo de procesamiento de lenguaje natural (usando un tokenizador y un modelo entrenado) para identificar el tema.
   Se convierte el número predicho al nombre del tema usando un segundo *label encoder*.

6. **Filtrar por tema (si aplica):**
   Si hay canciones en el subgénero predicho que también coinciden con el tema de las letras, se filtran aún más.
   Si no se encuentran coincidencias, se usa solo el filtro por subgénero.

7. **Seleccionar canciones recomendadas:**
   Se devuelven las primeras canciones filtradas (por defecto, las 5 mejores recomendaciones) con detalles como nombre, subgénero, duración y tema.

5 Discusión

Los resultados obtenidos en este proyecto destacan el potencial de combinar características numéricas y análisis de lenguaje natural para mejorar los sistemas de recomendación musical. La precisión del modelo al identificar subgéneros musicales y temas líricos demuestra la capacidad del enfoque propuesto para capturar patrones complejos tanto en los datos estructurados como en las letras de las canciones.

6 Conclusiones

El presente proyecto demostró que es posible construir un sistema de recomendación musical efectivo al combinar características numéricas de las canciones con análisis de lenguaje natural aplicado a las letras.

Entre los principales hallazgos, se destacó la capacidad del modelo de clasificación basado en *Random Forest* para predecir subgéneros musicales con alta precisión.

Además, el uso de un modelo de lenguaje preentrenado para el análisis de letras permitió identificar temáticas líricas de manera eficiente, pero no tan acertado como el análisis acústico