Predicción de géneros musicales

Alexis Hernández Morales

Facultad de Ciencias Físico Matemáticas, Universidad Autónoma de Nuevo León, San Nicolás de los Garza, México.

1. Introducción

El reconocimiento automático de géneros musicales ha sido un área de gran interés en la investigación, con aplicaciones en la organización de bibliotecas musicales y recomendaciones personalizadas. Tradicionalmente, la categorización de música se ha basado en la percepción humana, considerando aspectos como la instrumentación, el ritmo y la armonía. Sin embargo, el creciente volumen de datos musicales y la necesidad de clasificaciones precisas han impulsado el desarrollo de enfoques automatizados basados en el análisis de audio y modelos de aprendizaje automático.

2. Antecedentes

Para este proyecto se utilizaron como soporte las siguientes investigaciones previas:

El investigador Pavan utilizando características extraídas de archivos de audio comparo modelos de Bosque Aleatorio y Árboles de Decisión en términos de precisión, ayudando a predecir de forma correcta el género musical de los archivos. Se apoyó en Coeficientes Cepstrales de Frecuencia Mel principalmente. Aprovechando estos datos logró una precisión de predicción del 71,78 % con el modelo de Bosque Aleatorio[1].

Tzanetakis and Cook, realiza un modelo para la clasificación musical, caracterizado por el uso de señales de audio, aplicando métodos estadísticos de entrenamiento y reconocimiento de patrones. [2].

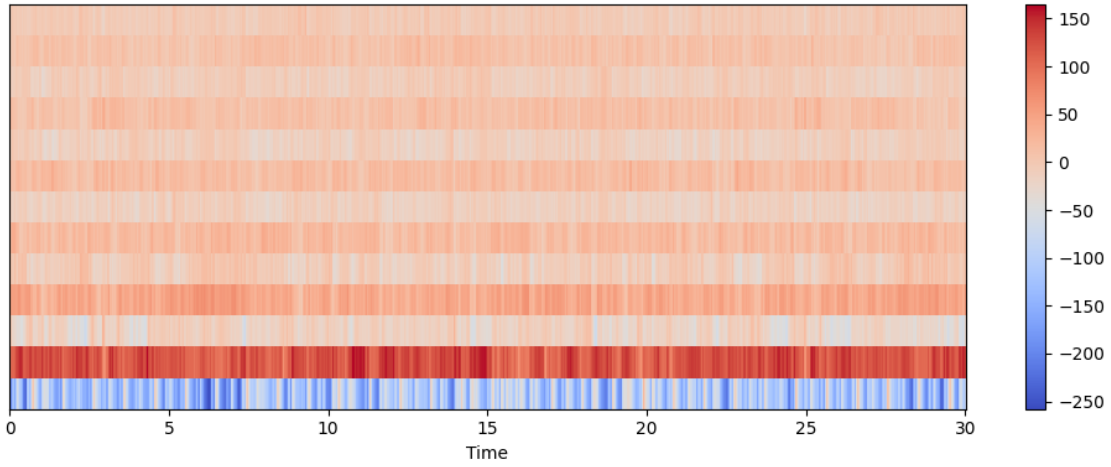
3. Descripción de los datos

Esta base de datos se descargó de *Kaggle* originalmente con 1 000 archivos de audio divididos por género musical, teniendo 100 archivos de audio por cada uno, con una longitud de 30 segundos por audio.

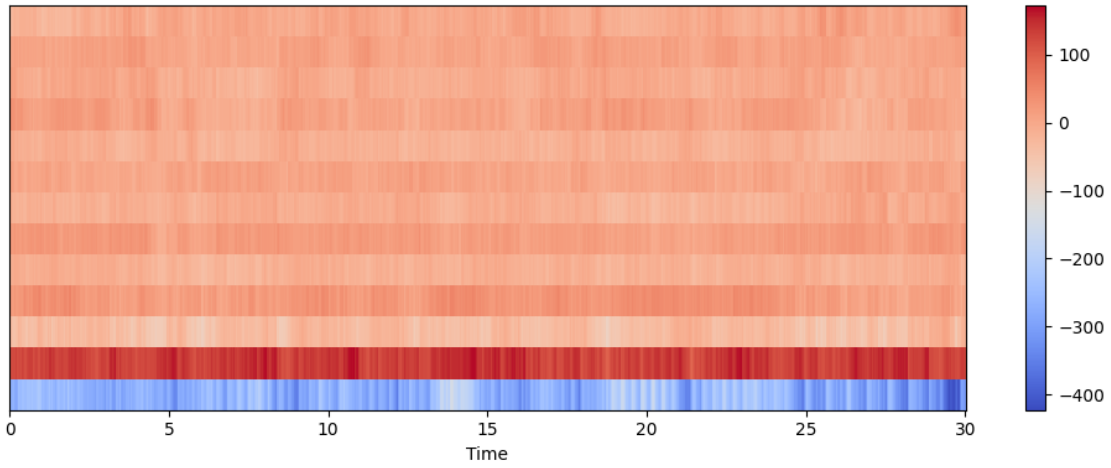
Teniendo en mente el procesamiento exhaustivo de los archivos, se decidió utilizar solamente 150 audios clasificados en 5 géneros diferentes, con 30 audios cada uno.

4. Estudio estadístico

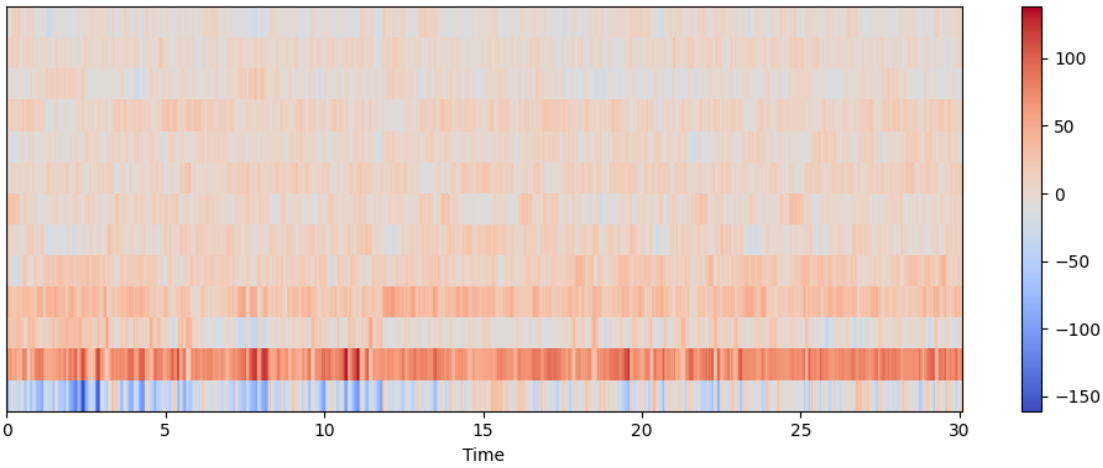
A continuación se muestran gráficas sobre Coeficientes Cepstrales de Frecuencia Mel y espectrogramas.



(a) `Blues`

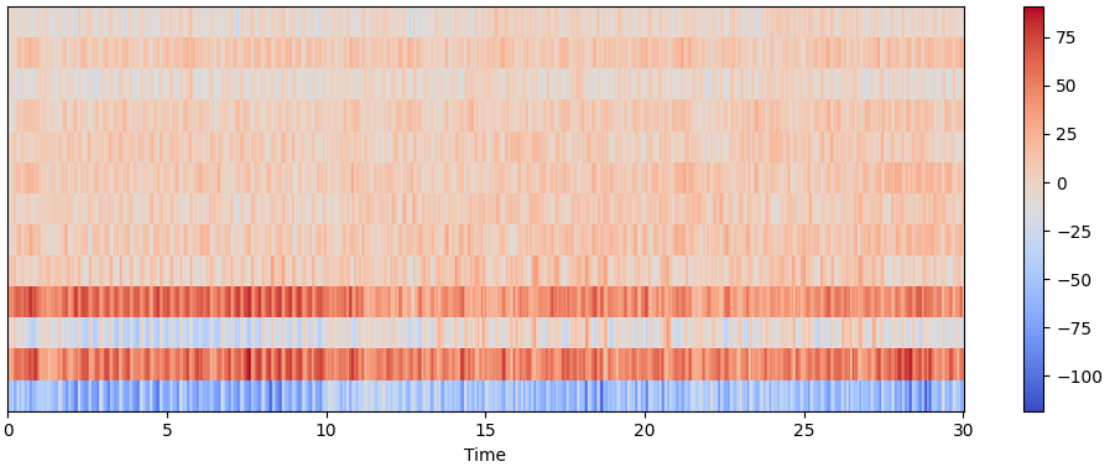


(b) `Classical`

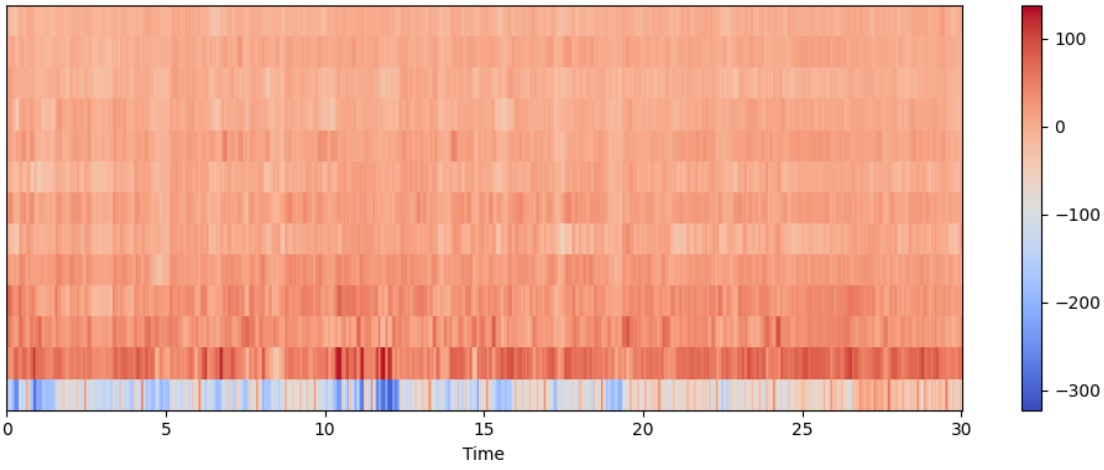


(c) `Country`

Figura 1: Coeficientes Cepstrales de Frecuencia Mel (Parte 1)

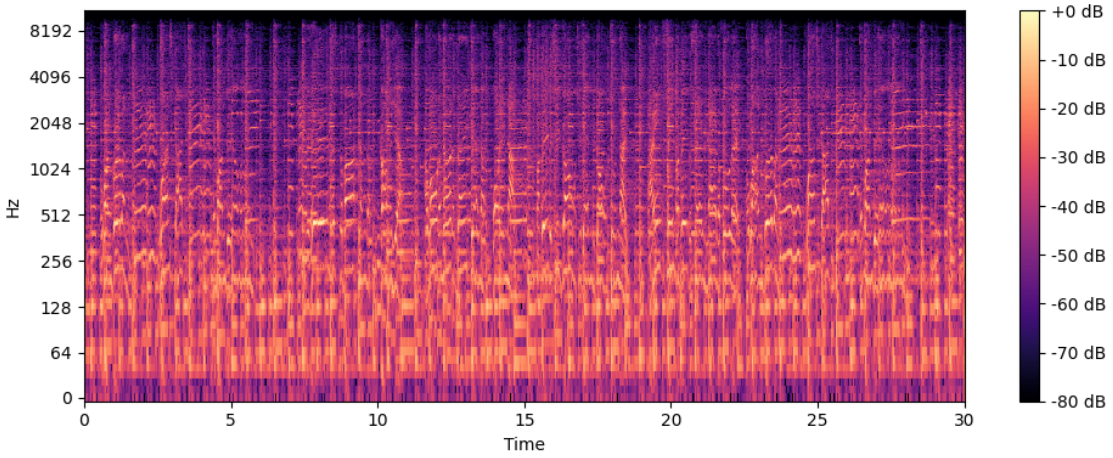


(a) `Metal`

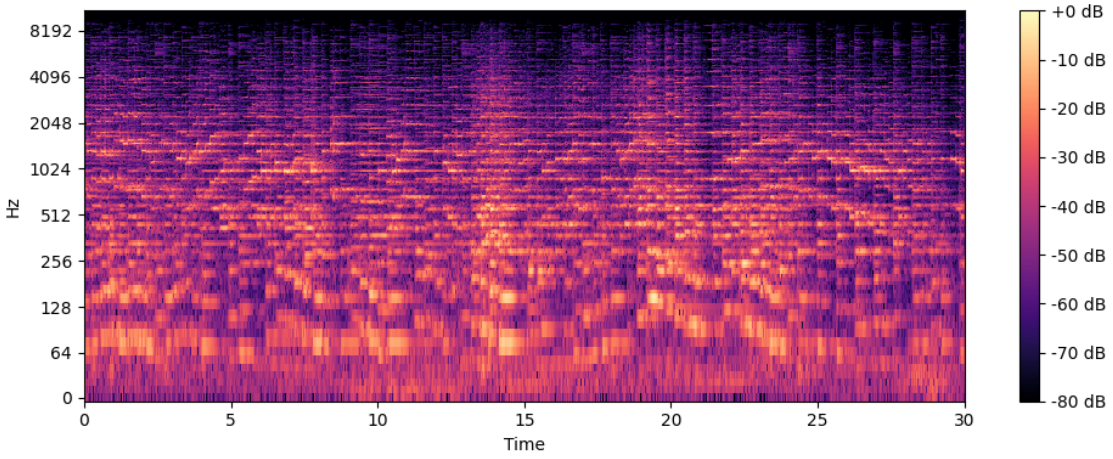


(b) `Pop`

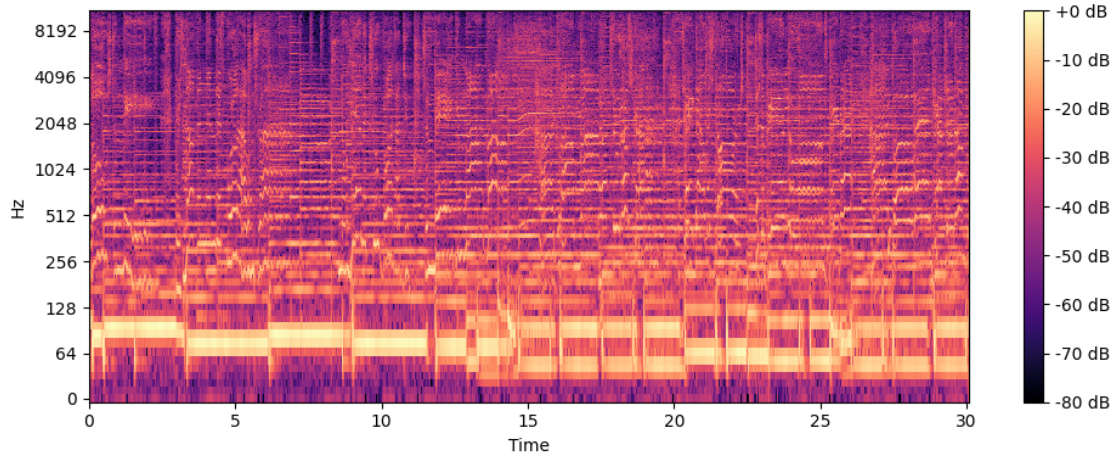
Figura 2: Coeficientes Cepstrales de Frecuencia Mel (Parte 2)



(a) `Blues`

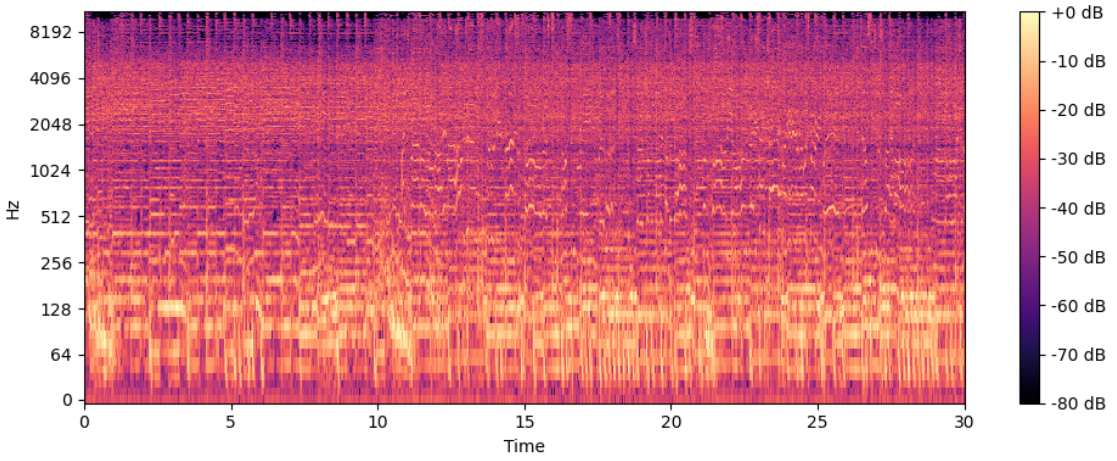


(b) `Classical`

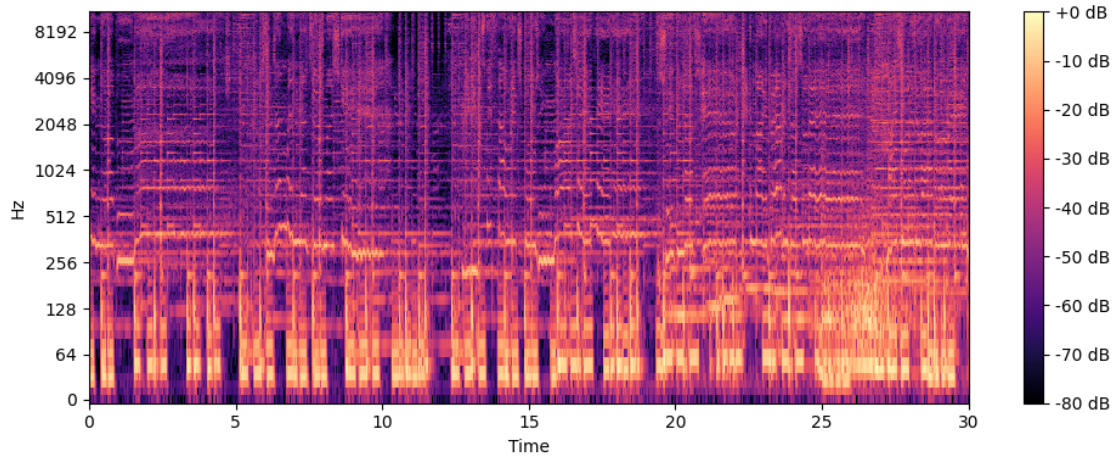


(c) `Country`

Figura 3: Espectrogramas (Parte 1)



(a) `Metal`



(b) `Pop`

Figura 4: Espectrogramas (Parte 2)

5. Metodología

Se aplicaron una variedad de métricas a los archivos de audio, permitiendo un análisis de mayor amplitud con una perspectiva mas completa hacia los audios, de esta forma se obtuvieron parámetros de gran valor definiendo con claridad características de los audios y de los géneros que representaban.

Con esto en cuenta se hizo uso de el modelo de aprendizaje automático de Bosque Aleatorio con la intención de aprovechar las variables previamente mencionadas y así cargar esta información al modelo.

A continuación se presenta el modelo, métricas y procedimientos aplicados.

5.1. Tasa de Cruces por Cero

La Tasa de Cruces por Cero es una métrica utilizada en el procesamiento de señales de audio para describir la estructura temporal y espectral de una señal. Define la frecuencia con la que la señal cambia de signo en un intervalo de tiempo determinado, es decir, cuántas veces pasa de valores positivos a negativos o viceversa.

Señales con altos valores corresponden a sonidos con cambios rápidos en la forma de onda, como ruido blanco, percusión o consonantes fricativas en el habla.

Señales con bajos valores suelen estar asociadas a sonidos más continuos y armónicos, como vocales, cuerdas o señales de baja frecuencia.

5.2. Coeficientes Cepstrales de Frecuencia Mel

Los Coeficientes Cepstrales de Frecuencia Mel son una representación matemática de las señales de audio diseñada para capturar características espectrales de manera similar a cómo el sistema auditivo humano percibe el sonido. Utilizan una escala de frecuencia perceptual conocida como escala Mel. Esta escala refleja la manera en que los humanos perciben las diferencias entre frecuencias, donde las variaciones en las bajas frecuencias son más notorias que en las altas.

5.3. Centroide Espectral

El Centroide Espectral es una métrica utilizada en el análisis de señales de audio para describir la distribución de energía en el espectro de frecuencia de una señal. Indica si la mayor parte de la energía espectral se encuentra en frecuencias bajas o altas y se define como el promedio ponderado de las frecuencias presentes en la señal.

5.4. Caída Espectral

La Caída Espectral es una técnica utilizada en el análisis de audio para examinar cómo las características espectrales de una señal acústica cambian a lo largo del tiempo, especialmente en respuesta a variaciones en el entorno o el material que genera o transmite el sonido. Se enfoca en cómo disminuye la intensidad de las frecuencias de una señal de audio a medida que pasa a través de diferentes superficies o es afectada por factores como la absorción acústica.

5.5. Características de *Chroma*

El análisis de *chroma* se centra en extraer las características tonales de una señal de audio relacionadas con las notas musicales y sus alturas, sin considerar las variaciones de timbre o el ritmo. Se utiliza para representar la tonalidad y la estructura armónica de la música a través de las intensidades de las notas musicales presentes en una pieza. Se enfoca exclusivamente en las frecuencias asociadas con las doce notas de la escala cromática.

5.6. Bosque Aleatorio

El Bosque Aleatorio es un método de aprendizaje automático que utiliza Árboles de Decisión para crear predicciones y clasificaciones. Cuando el Bosque Aleatorio está prediciendo un nuevo ítem basado en ciertos atributos, cada Árbol de Decisión da su propia clasificación como resultado, entonces el resultado general del Bosque Aleatorio será el mayor número de taxonomía. En el caso de una regresión, el resultado será el valor promedio de todos los Árboles de Decisión.

6. Métricas de desempeño

A partir de los valores obtenidos del modelo se utilizan métricas para medir el desempeño con el que el modelo está realizando estas predicciones y así saber lo bien o mal que se está evaluando.

- Exactitud: Es la proporción de verdaderos positivos y verdaderos negativos con respecto a todas las observaciones positivas y negativas.
- Puntuación $F1$: La puntuación $F1$ es la media armónica de la precisión y la puntuación de recuperación. Da el mismo peso tanto a la precisión como a la recuperación para medir su rendimiento en términos de precisión.
- Matriz de Confusión: Representación tabular que muestra cómo las predicciones del modelo se comparan con los valores reales, es decir, las etiquetas verdaderas.

7. Resultados

Las métricas calculadas al igual que sus medias y varianzas fueron de gran utilidad dando perspectiva de lo que define un audio, siendo en este caso especifico muestras de canciones, esta utilidad se mostró al momento de la aplicación del modelo de Bosque Aleatorio con el que se pretendió clasificar de forma eficiente el género musical de los archivos de audio.

Al procesar los datos mediante el método de Bosque Aleatorio enfocado a la predicción del género musical se obtuvo una exactitud del 84,1 % y una puntuación $F1$ del 82 %.

Observando la Matriz de Confusión se tiene que en `Blues`, `Classical`, `Metal` y `Pop` la mayoría de las predicciones están en la diagonal principal, mientras que en `Country` tres instancias fueron clasificadas erróneamente.

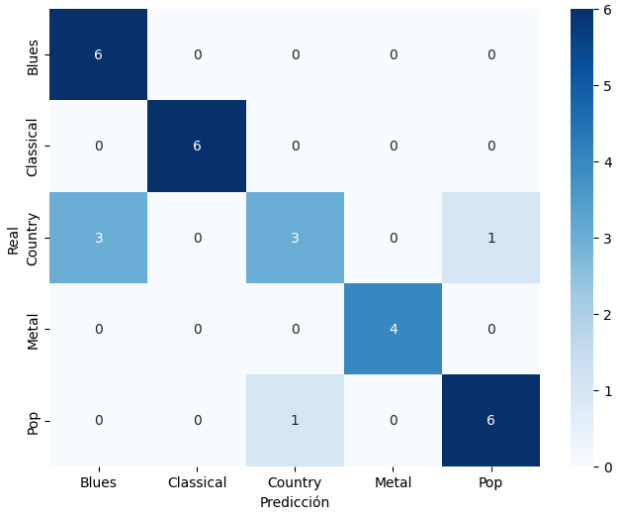


Figura 5: Matriz de Confusión

8. Conclusiones

El modelo mostró un desempeño aceptable en la tarea de predicción con el porcentaje de precisión. La matriz de confusión muestra que el modelo clasifica correctamente la mayoría de los casos, especialmente en géneros como `Blues`, `Classical`, `Metal` y `Pop`, donde las predicciones correctas son altas. Sin embargo, existen errores en la clasificación del género `Country`, que presenta confusión con `Blues` y `Pop`.

A pesar del buen desempeño general del modelo, la presencia de falsas clasificaciones en algunos géneros sugiere oportunidades de mejora.

Referencias

[1] R. Pavan, V. y Dhanalakshmi. Analysis of audio data and prediction of the genre using novel random forest and decision tree. pages 1773–1777, 2022. doi: 10.1109/ICIRCA54612.2022.9985019.

[2] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002. doi: 10.1109/TSA.2002.800560.