Escuchando Emociones | Ortega Ibarra Jaime Jesus

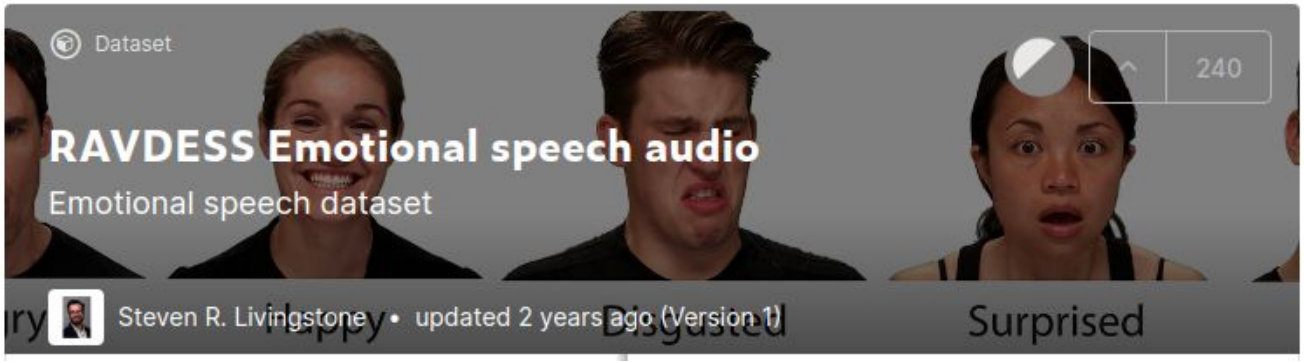
Proyecto Final
Introducción al Aprendizaje Profundo
Licenciatura en Ciencia de Datos

Recordando...

Dado el crecimiento exponencial de los datos en la actualidad y a su vez el crecimiento en los diferentes formatos en que encontramos estos, El proyecto Escuchando Emociones plantea poder realizar un análisis de emociones dado un cierto conjunto de datos.

Para dicho proyecto se utilizará la base de Datos “RAVDESS”, la cual podremos encontrar en el siguiente enlace:

https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio



Herramientas

Datos

Lenguaje de Programación

Entorno

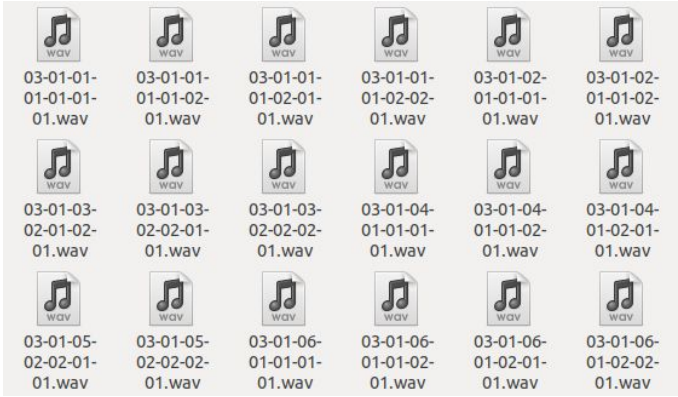
Modelo

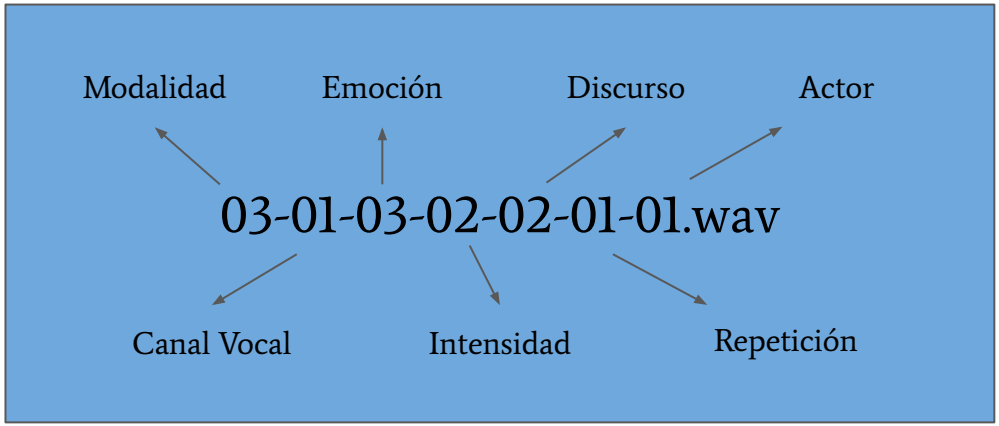
Procesamiento de Datos

Visualización

Formato y Etiquetado

Contamos con 1440 Audios correspondientes a los 24 diferentes actores, estos se encuentran distribuidos en 24 diferentes carpetas (Divididas por actor) y se muestran en formato *.wav.*
El tamaño por auido ronda entre los *350Kb* y *750Kb.*





Como podemos observar, el nombre del archivo contiene el etiquetado correspondiente, este coincide con el tipo de emoción expresada, género del actor, número de actor, intensidad y discurso.

Diccionario de Etiquetado

Modalidad:
- 01: AV Completo
- 02: Solo video
- 03: Solo Audio

Canal Vocal:
- 01: Habla
- 02: Canción

Repetición:
- 01: 1a Repetición
- 02: 2a Repetición

Intensidad Emocional:
- 01: Normal
- 02: Fuerte

Actor:
- Impar: Hombre
- Par: Mujer

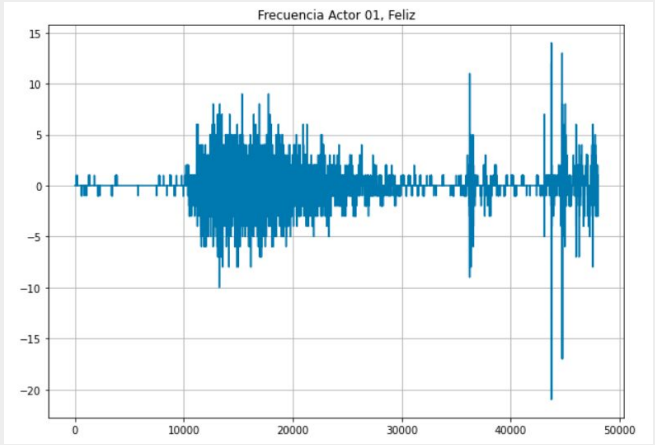
Discurso:
- 01: “Kids are talking by the door”.
- 02: “Dogs are sitting by the door”.

Emoción:
- 01: Neutral
- 02: Calma
- 03: Feliz
- 04: Triste
- 05: Enojado
- 06: Temeroso
- 07: Disgustado
- 08: Sorprendido

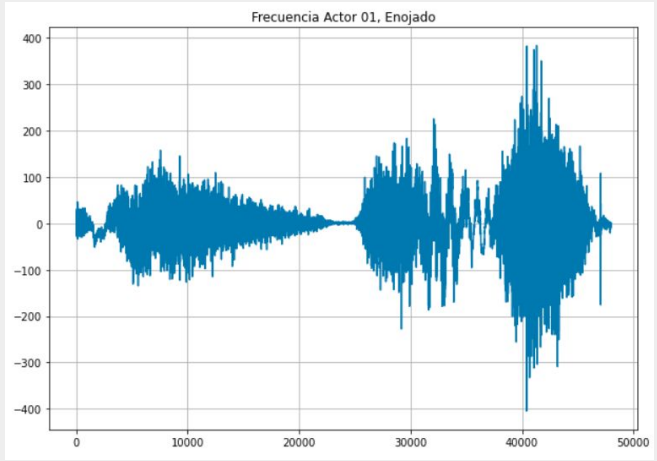
Visualización de nuestros datos.

Con ayuda de librerías tales como *IPhyton, Scipy y Librosa,* hemos podido observar la frecuencia de nuestros audios, en el siguiente ejemplo, se tomaron dos audios de un mismo actor, pero con diferente emoción:

Actor 01
Emoción: Feliz



Actor 01
Emoción: Enojado



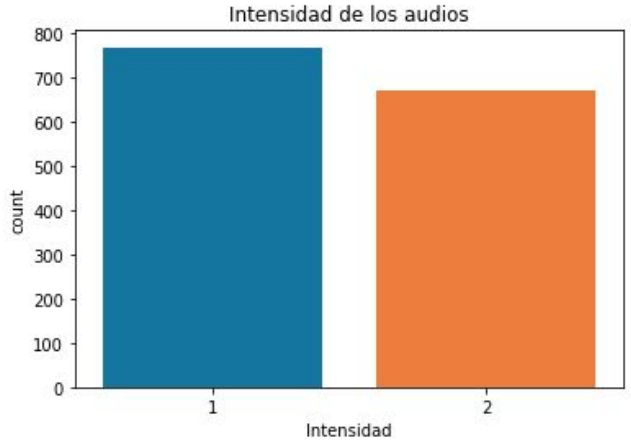
Desarrollo

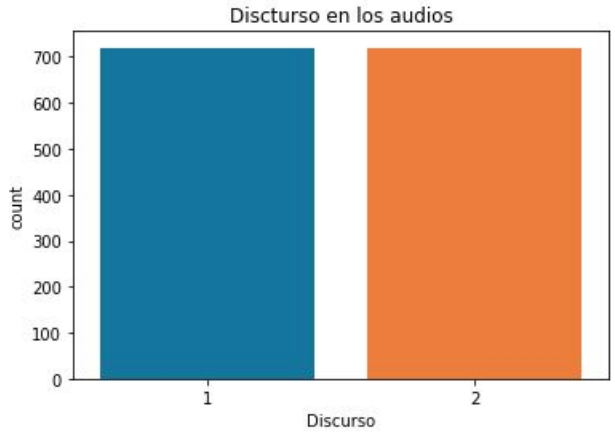
Tipo de Problema:

Dado el tipo de problema, en el cual deseamos detectar una posible emoción a partir de la voz, este se puede denominar como un problema de Clasificación.
Pues mediante los datos entrenados con diversas etiquetas, deseamos clasificar datos posteriores.

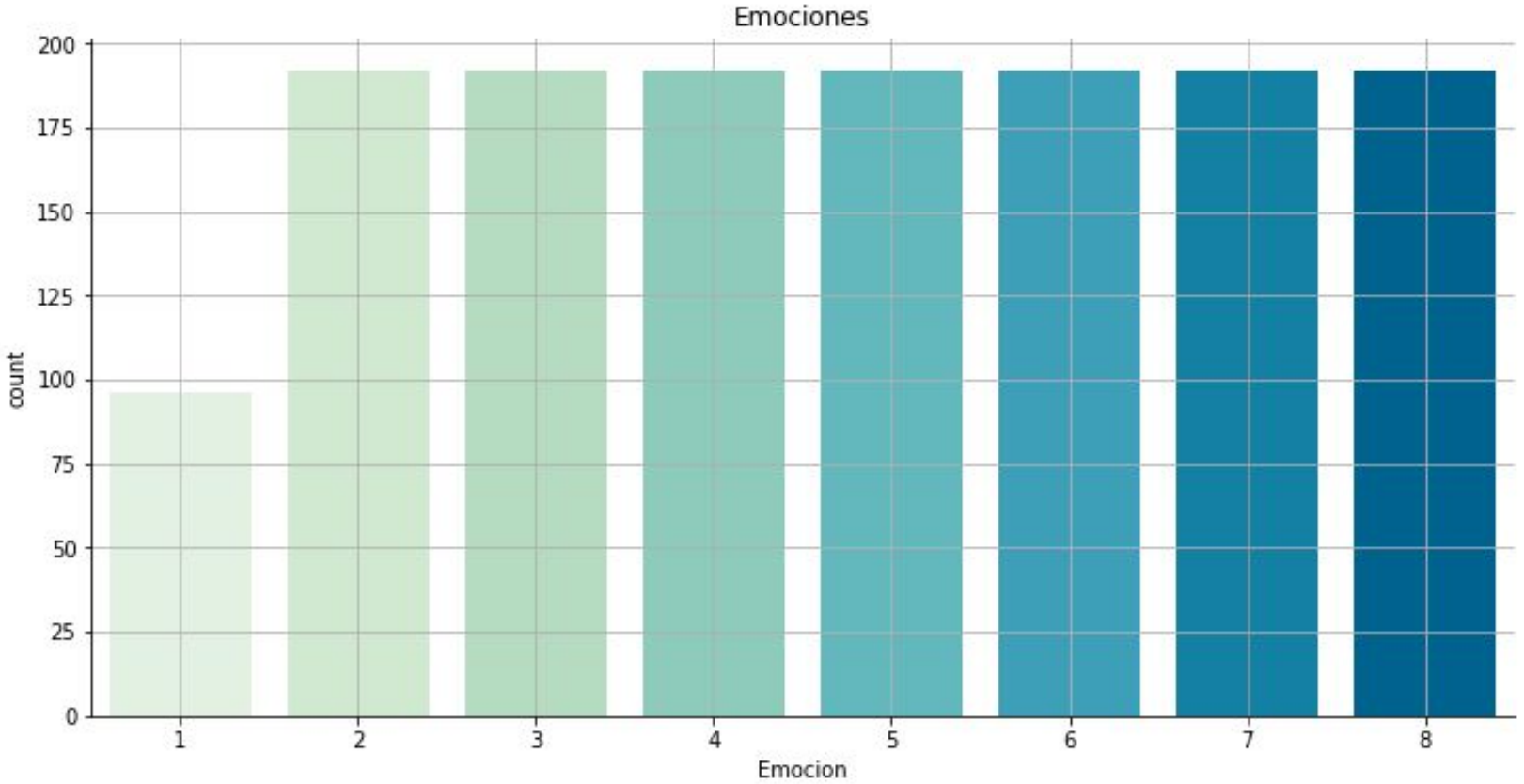
Exploración de los datos:

Para poder abordar el problema de la mejor forma, se ha realizado un análisis exploratorio sobre los diversos datos.





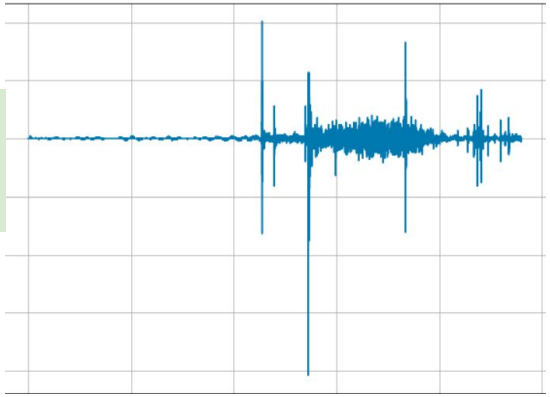
También es importante verificar que nuestras clases se encuentren equilibradas, para ello se obtuvo una gráfica mostrando la distribución de estas, la cual fue muy favorable:



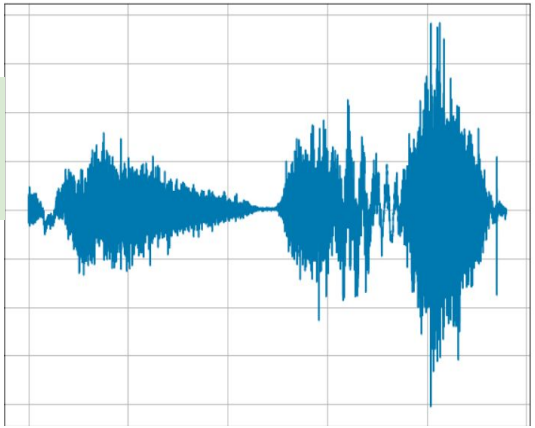
Limpieza y procesamiento de audios

Mediante la visualización de las frecuencias, he notado diferencia en los picos de onda, por lo que se optó graficar la distribución de la duración de nuestros audios para evaluar.

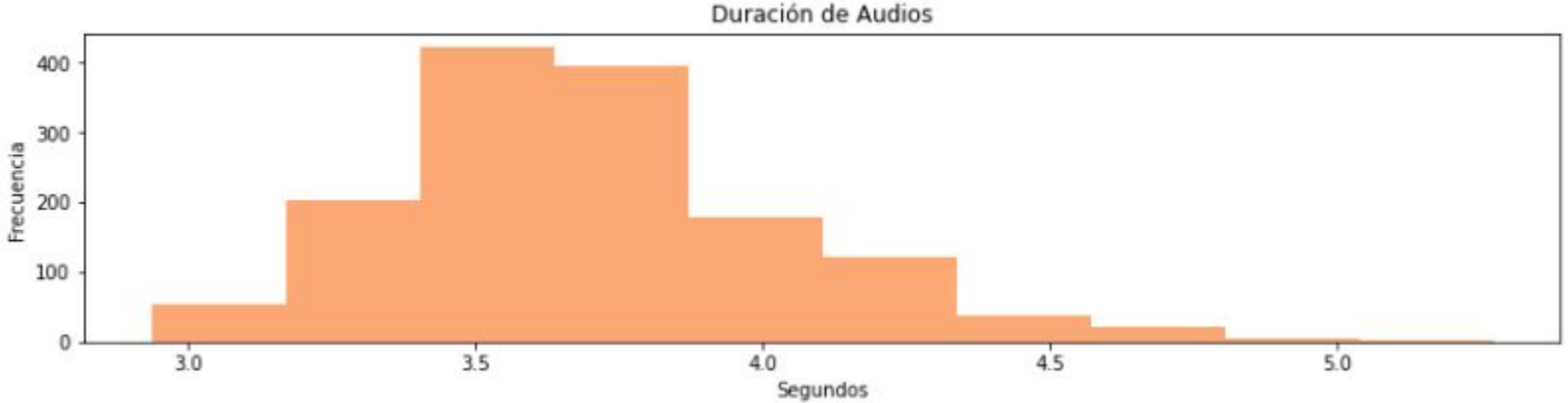
Audio con Silencio Inicial



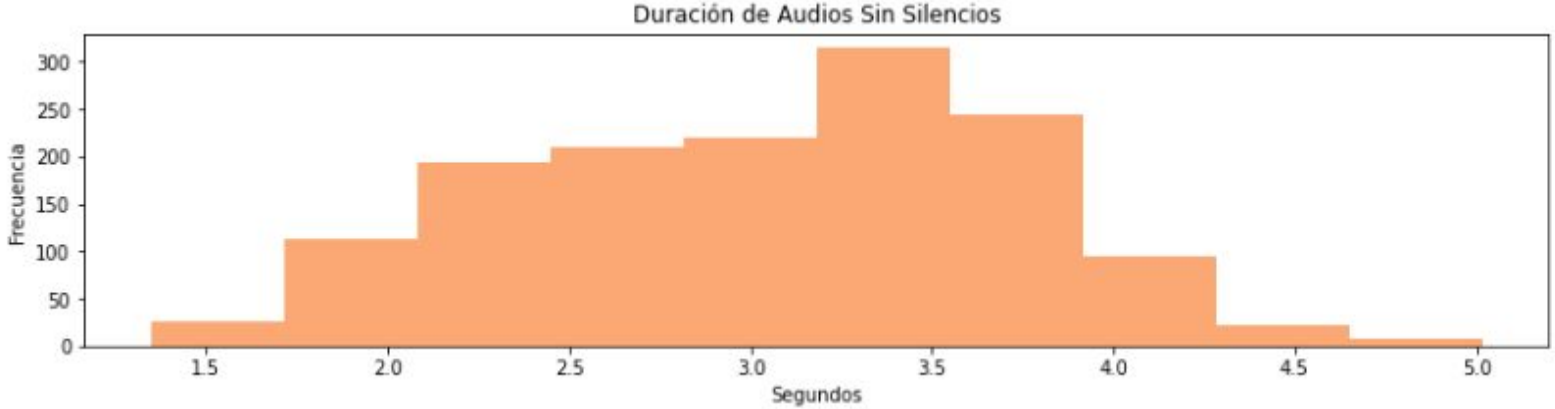
Audio sin Silencio Inicial



Eliminación de Silencios

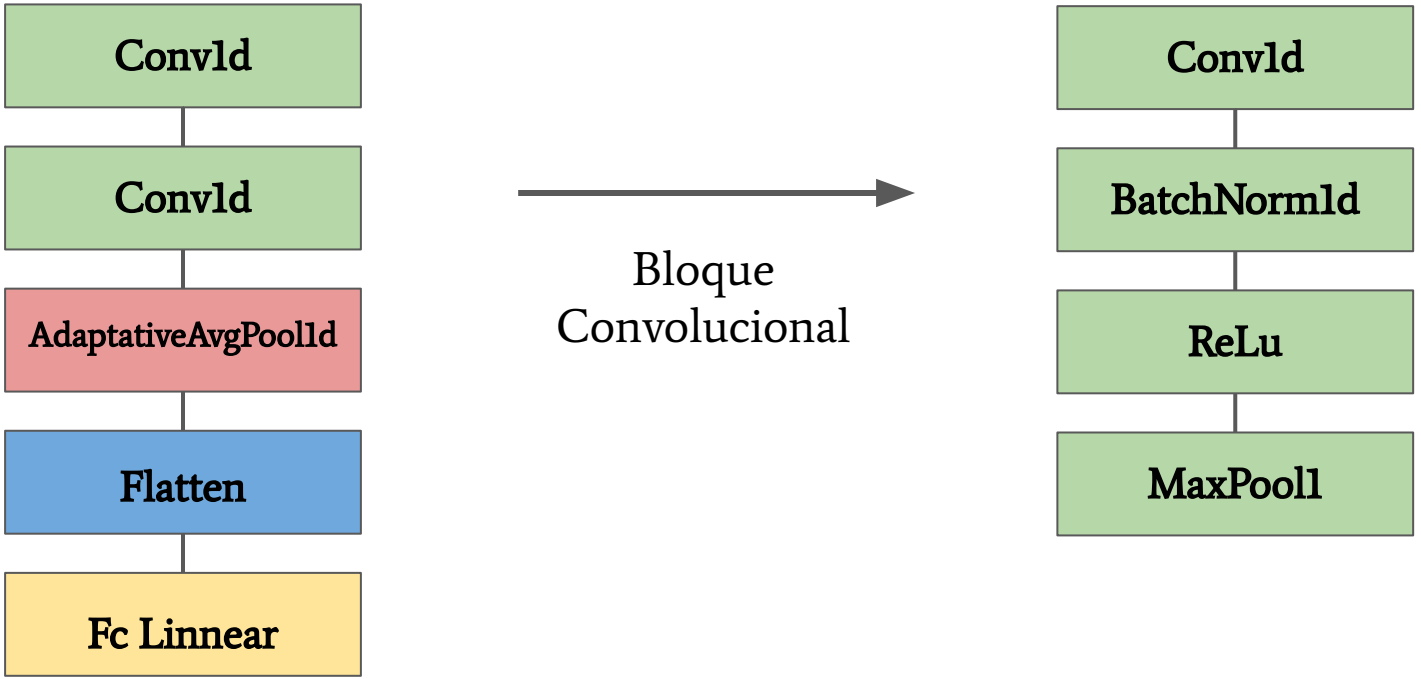


Con ayuda de la función effects.trim(), podemos eliminar los silencios en nuestros audios, dando los siguientes resultados.

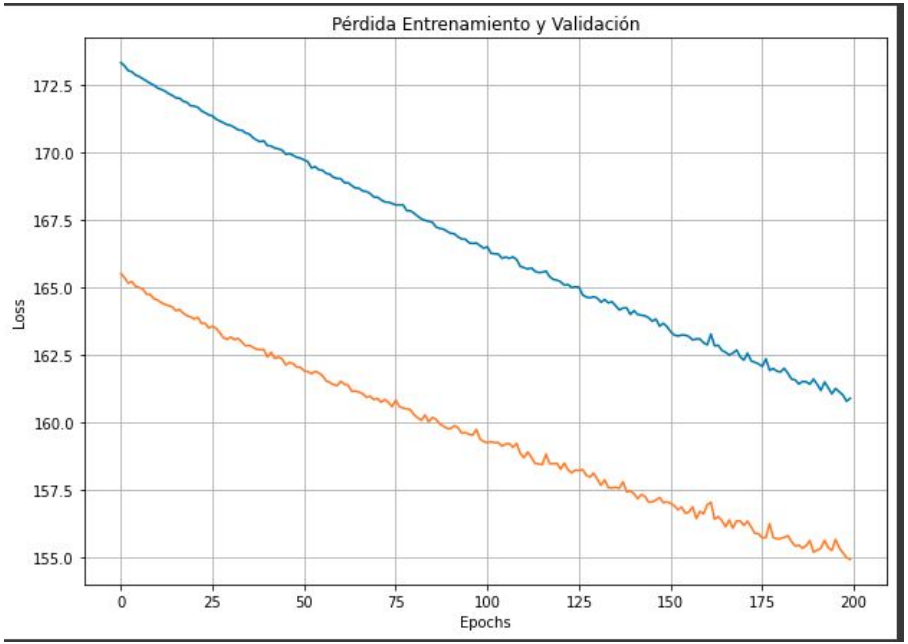


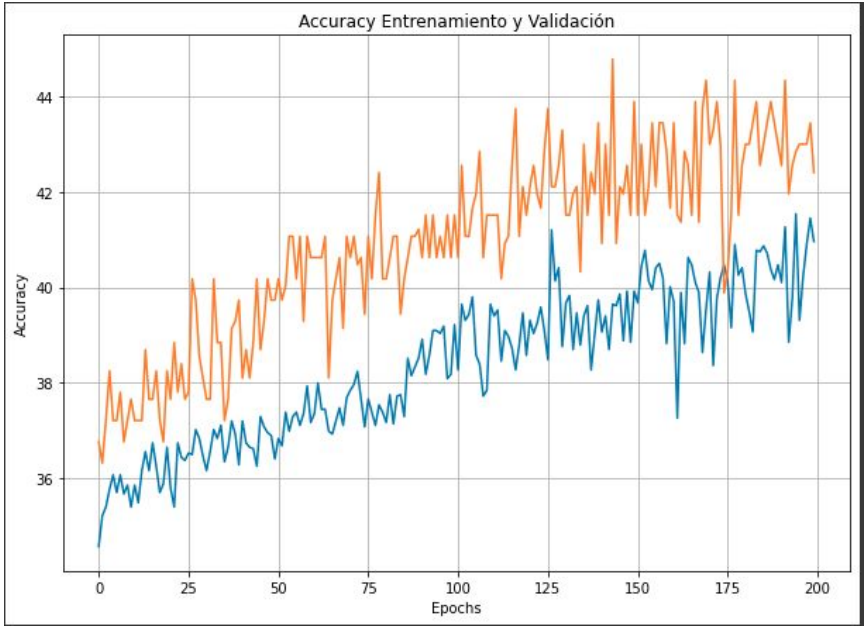
Modelo

Para llevar a cabo este proyecto, decidí enfocarme en Redes Neuronales Convolucionales, específicamente utilizar una arquitectura M5, nos dará buenos resultados al momento de entrenar nuestros audios en crudo, la arquitectura fue la vista en clase, es decir:



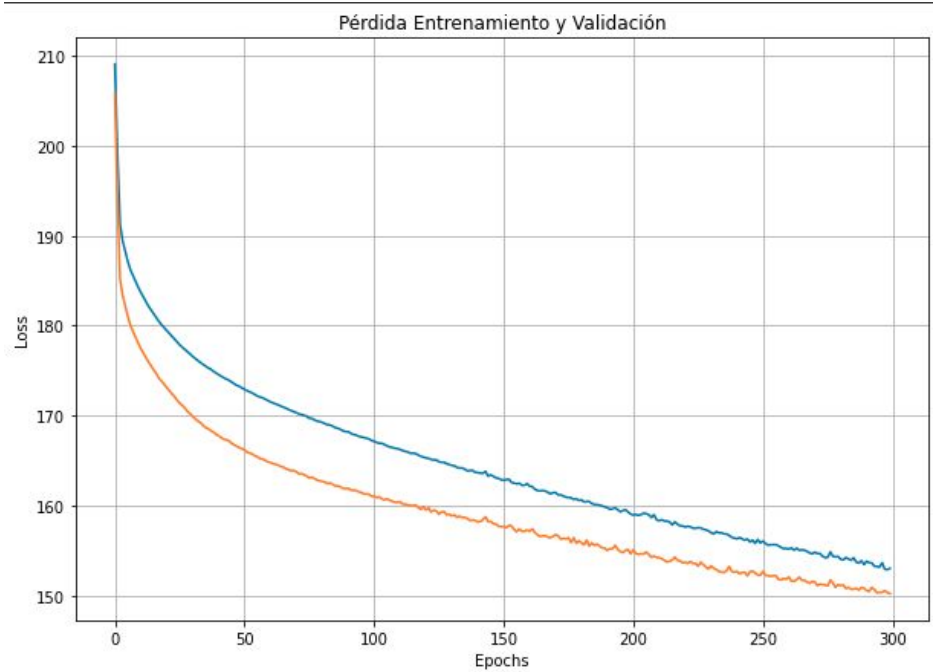
Resultados Audios sin Silencio

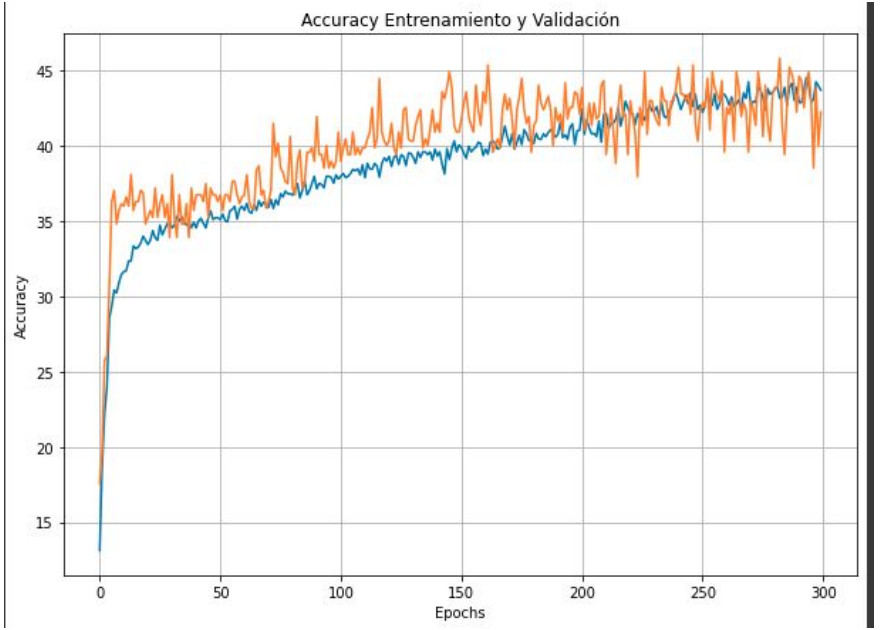




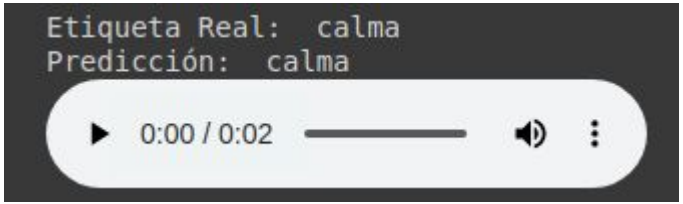
Resultados

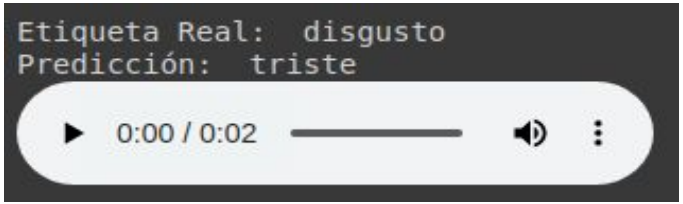
Al entrenar nuestro modelo con 300 épocas, la pérdida y el accuracy fue el siguiente:

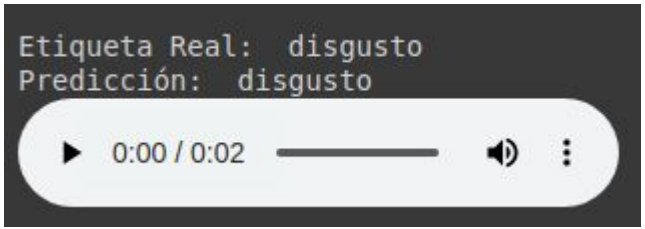


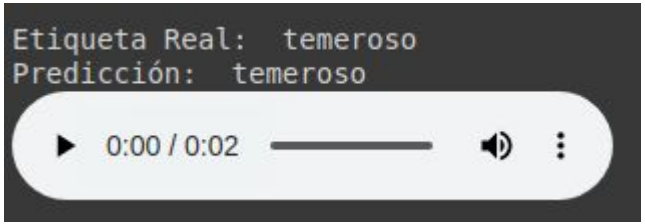


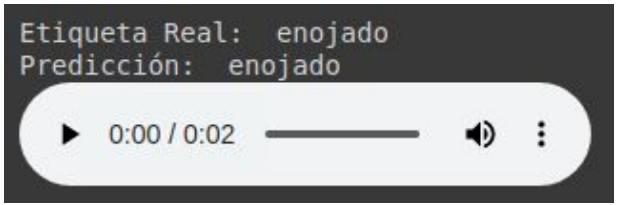
Predicciones

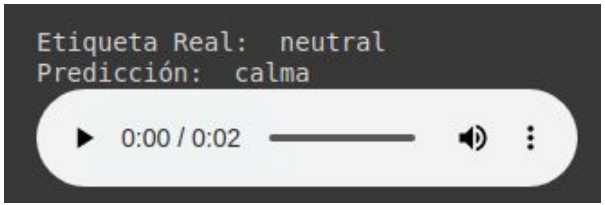












Conclusiones

Se encontraron diversas dificultades al momento de tratar los audios, pues no todas las técnicas para procesar son favorables, tal como lo vimos al quitar los silencios, pues nuestro modelo daba peores resultados, además la falta de datos puede influir de cierta manera, ya que un audio de mayor duración, podría proporcionar mayor cantidad de características para un entrenamiento más preciso, cabe mencionar que la etiqueta en la que mayores fallas tenía, era en la etiqueta neutral, lo cual lo podemos relacionar con la falta de datos en dicho atributo.