

“BAG OF SONGS”: EXTRACCIÓN DE CARACTERÍSTICAS EN SEÑALES DE AUDIO

Pablo Reina Jiménez
María Lourdes Linares Barrera



ÍNDICE DE CONTENIDOS

- 1** _____
Presentación
- 2** _____
Conjunto de datos
- 3** _____
**Extracción de características
y análisis cualitativo**
- 4** _____
**Entrenamiento de
modelos**
- 5** _____
Conclusiones y cierre

01

PRESENTACIÓN

- 1) Objetivos del proyecto
- 2) Estructura del proyecto



Introducción

Conjunto de
datos

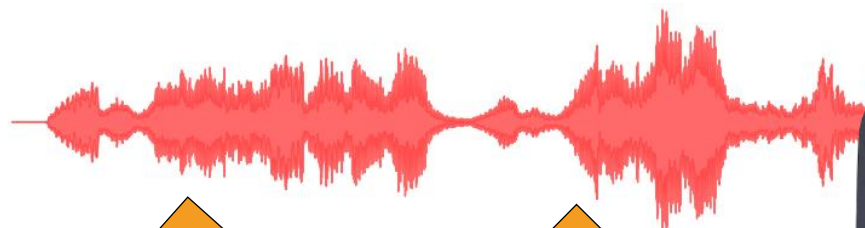
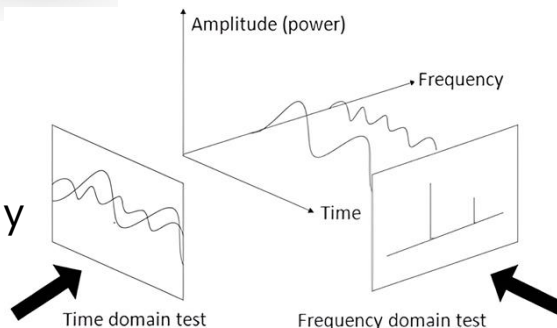
Extracción de
características

Entrenamiento de
modelos

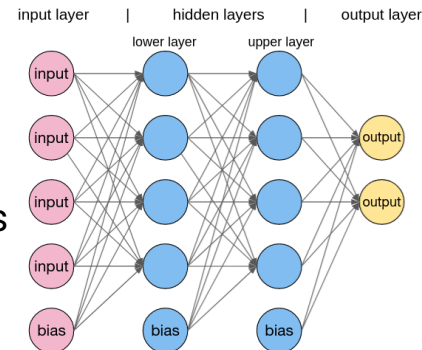
TEMÁTICA



Estudiar las
características del
dominio frecuencial y



Clasificación y
distinción de
géneros musicales





Introducción

Conjunto de datos

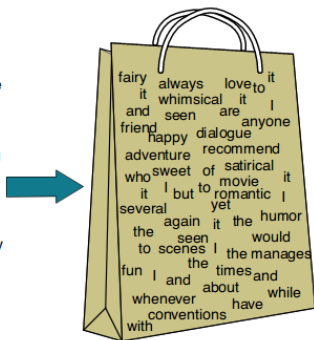
Extracción de características

Entrenamiento de modelos

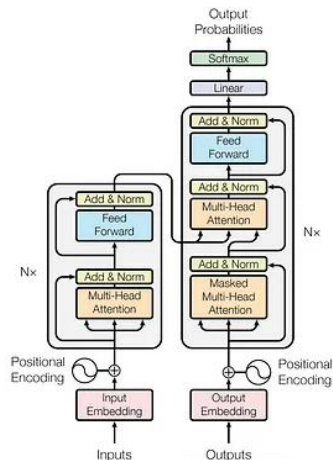
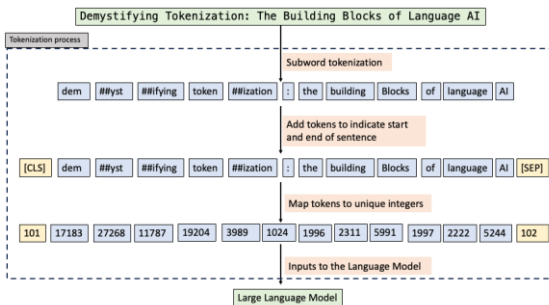


TEXTOS

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...



Bag of words: tfidf (datos tabulares), modelos ML (SVC, RF...) o MLP.

Tokenizado, embeddings, modelos LSTM, transformers...

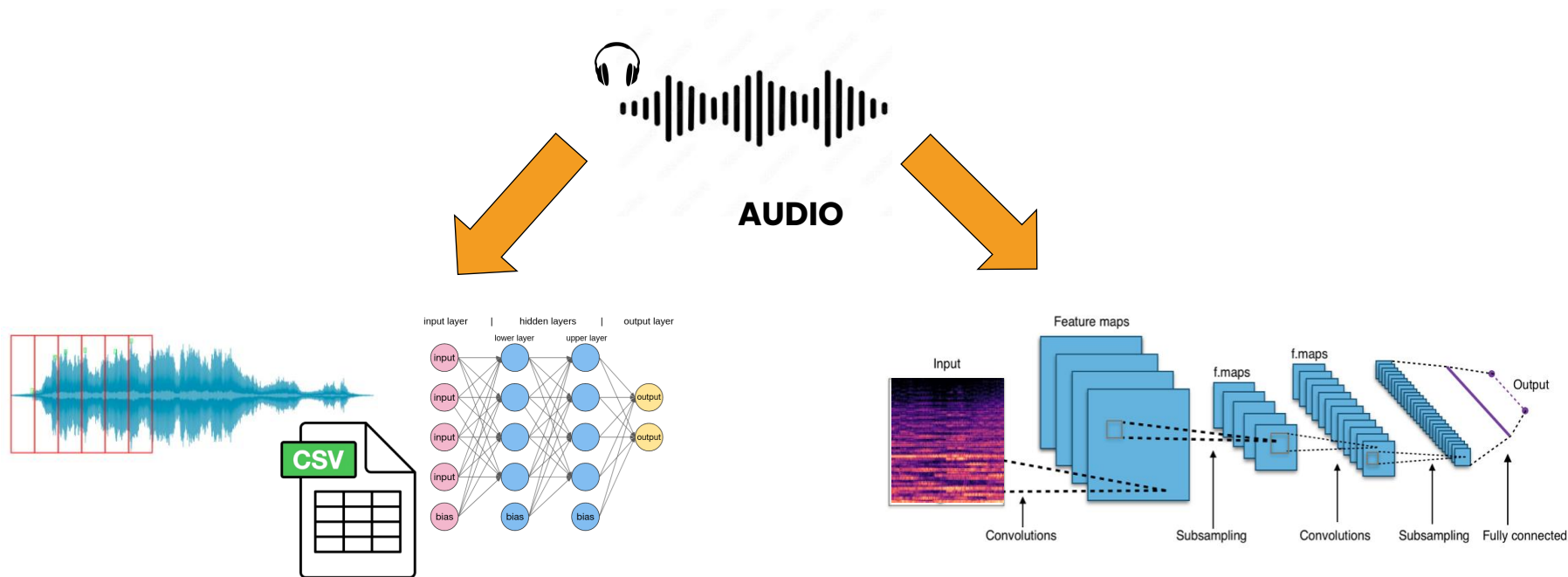


Introducción

Conjunto de datos

Extracción de características

Entrenamiento de modelos



- 1) "Bag of songs": extracción de características (datos tabulares), modelos ML (SVC, RF...) o MLP.
- 2) Serie temporal + modelos autoregresivos.

- 1) Espectrogramas + CNNs
- 2) RNNs
- 3) Transformers...



ESTRUCTURA DEL PROYECTO

▼ BAG-OF-SONGS

> ccmusic

> ccmusic2

> img

📁 .gitignore

📄 1-data-preparation.ipynb

📄 2-features-explanation.ipynb

📄 3-models-prediction.ipynb

📖 README.md



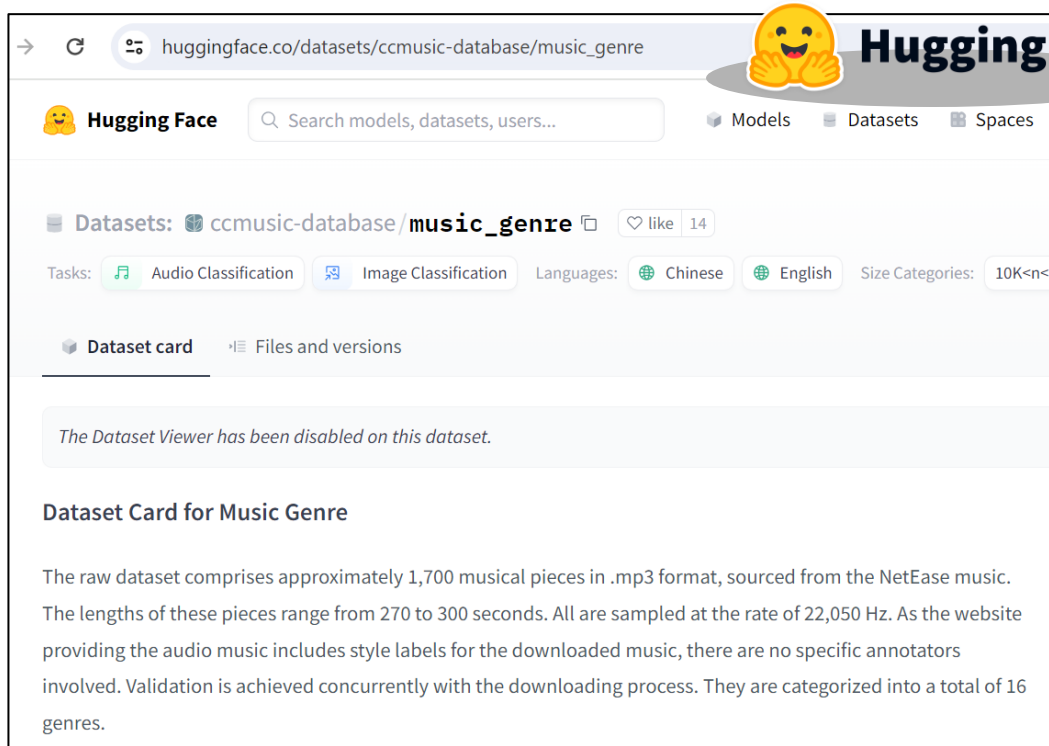
02

CONJUNTO DE DATOS

- 1) Descripción de los datos
- 2) Carga de los datos



CONJUNTO DE DATOS: CCMUSIC



The screenshot shows the Hugging Face interface for the dataset 'ccmusic-database/music_genre'. The page includes a search bar, navigation tabs for Models, Datasets, and Spaces, and a section for the dataset card. The dataset card is titled 'Dataset Card for Music Genre' and contains the following text:

The raw dataset comprises approximately 1,700 musical pieces in .mp3 format, sourced from the NetEase music. The lengths of these pieces range from 270 to 300 seconds. All are sampled at the rate of 22,050 Hz. As the website providing the audio music includes style labels for the downloaded music, there are no specific annotators involved. Validation is achieved concurrently with the downloading process. They are categorized into a total of 16 genres.

**Hugging Face**

> 15GB en piezas musicales



1 700 piezas musicales

- 1370 entrenamiento
- 171 validación
- 172 test



CONJUNTO DE DATOS: CCMUSIC

```
# Visualizamos la estructura del corpus
ccmusic_corpus = datasets.load_dataset(["ccmusic-database/music_genre", name="default", trust_remote_code=True])
print(ccmusic_corpus)
```

```
DatasetDict({
  train: Dataset({
    features: ['audio', 'mel', 'fst_level_label', 'sec_level_label', 'thr_level_label'],
    num_rows: 1370
  })
  validation: Dataset({
    features: ['audio', 'mel', 'fst_level_label', 'sec_level_label', 'thr_level_label'],
    num_rows: 171
  })
  test: Dataset({
    features: ['audio', 'mel', 'fst_level_label', 'sec_level_label', 'thr_level_label'],
    num_rows: 172
  })
})
```

> 15GB en piezas musicales



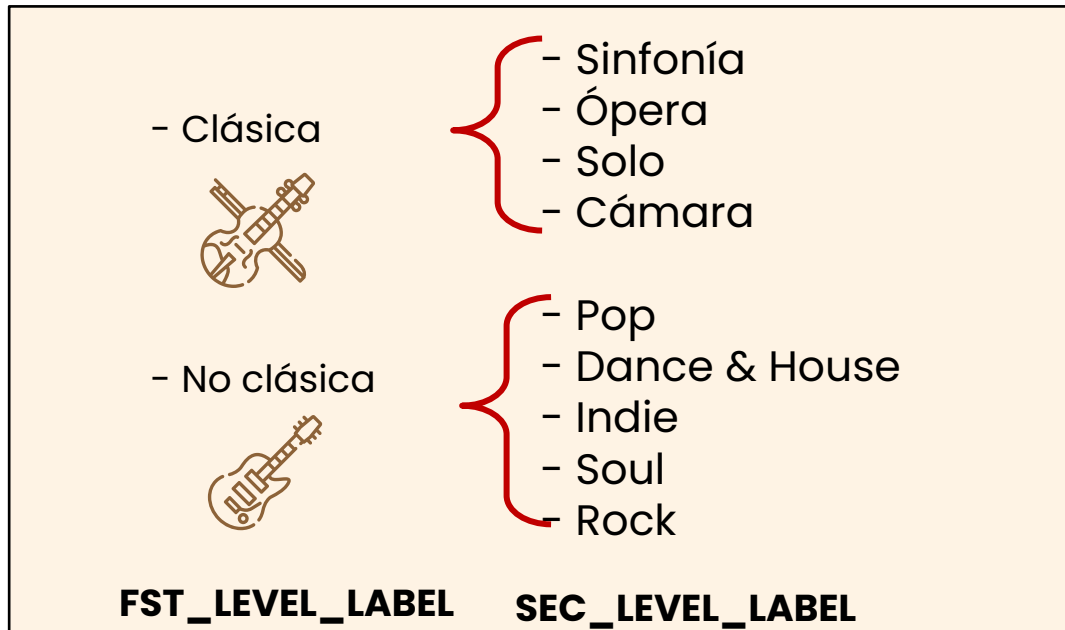
1 700 piezas musicales

- 1370 entrenamiento
- 171 validación
- 172 test



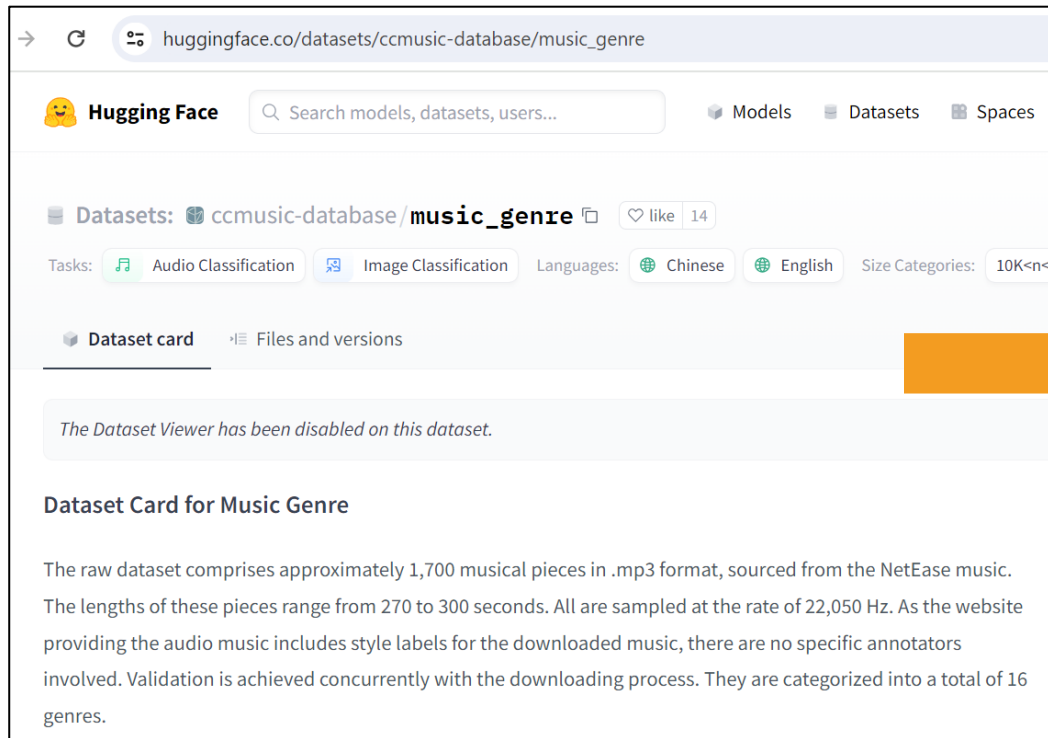
CONJUNTO DE DATOS: CCMUSIC

- Dataset orientado a clasificación.
- Maneja 3 jerarquías (**fst_level_label**, **sec_level_label**, **thr_level_label**).





NOTEBOOK 1 – CARGA DE LOS DATOS



→ huggingface.co/datasets/ccmusic-database/music_genre

Hugging Face Search models, datasets, users... Models Datasets Spaces

Datasets: ccmusic-database/music_genre like 14

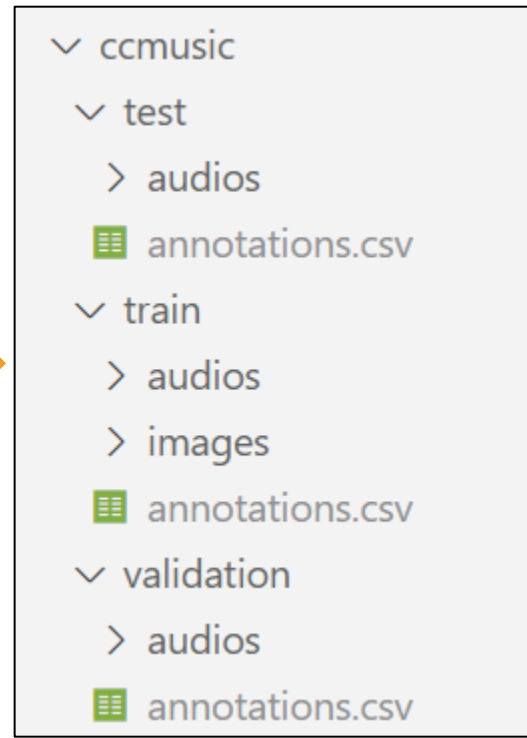
Tasks: Audio Classification Image Classification Languages: Chinese English Size Categories: 10K<nn<

Dataset card Files and versions

The Dataset Viewer has been disabled on this dataset.

Dataset Card for Music Genre

The raw dataset comprises approximately 1,700 musical pieces in .mp3 format, sourced from the NetEase music. The lengths of these pieces range from 270 to 300 seconds. All are sampled at the rate of 22,050 Hz. As the website providing the audio music includes style labels for the downloaded music, there are no specific annotators involved. Validation is achieved concurrently with the downloading process. They are categorized into a total of 16 genres.



- ccmusic
 - test
 - audios
 - annotations.csv
 - train
 - audios
 - images
 - annotations.csv
 - validation
 - audios
 - annotations.csv



NOTEBOOK 1 – CARGA DE LOS DATOS

1. Carga del dataset de Hugging Face.

2. Creación de funciones.

- Almacenamiento de ficheros WAV.
- Registrar anotaciones (categoría).
- Homogeneización (= sample rate, = mono, ≠ número de muestras): recorte/padding →
$$\text{Número de Muestras} = \text{Tasa de muestreo (Hz)} \times \text{Duración (s)} = 22050 \times 30 = 661500$$

3. Creación de estructura de directorios. Para cada partición:

- Creamos el fichero de anotaciones y la subcarpeta de anotaciones.
- Almacenamos el audio recortado.
- Creamos una entrada fichero-clase en anotaciones.

03

EXTRACCIÓN DE CARACTERÍSTICAS

- 1) Conceptos previos: Dominios en audio y segmentación
- 2) Características de señales de audio
- 3) Generación de dataset y caso práctico



EXTRACCIÓN DE CARACTERÍSTICAS

Dominios en señales de audio

Características de una señal de audio

Generación de dataset de características





EXTRACCIÓN DE CARACTERÍSTICAS

Dominios en señales de audio

Características de una señal de audio

Generación de dataset de características
y caso práctico (análisis descriptivo)





DOMINIO TEMPORAL

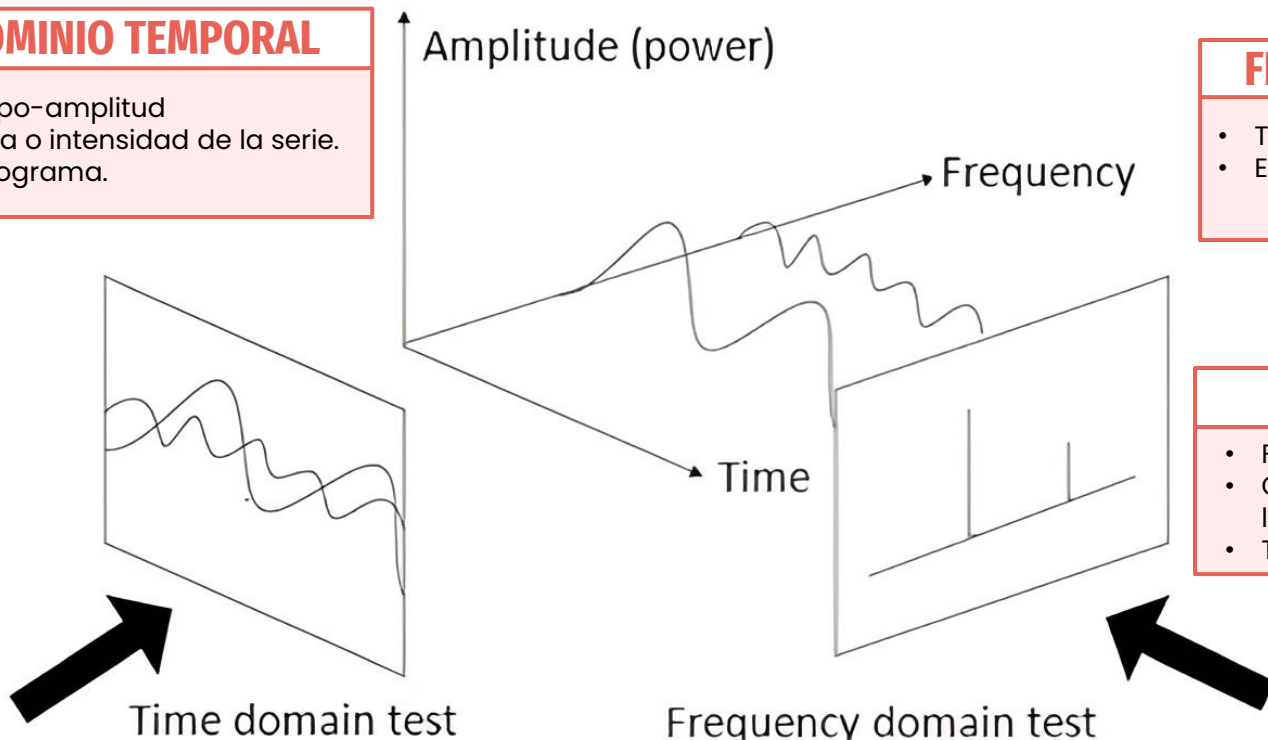
- Tiempo-amplitud
- Fuerza o intensidad de la serie.
- Oscilograma.

FRECUENCIAL-TEMPORAL

- Tiempo-frecuencia-magnitud.
- Espectrograma.

DOMINIO FRECUENCIAL

- Frecuencia-magnitud.
- Contribución de cada frecuencia a la señal global.
- Transformada de Fourier.





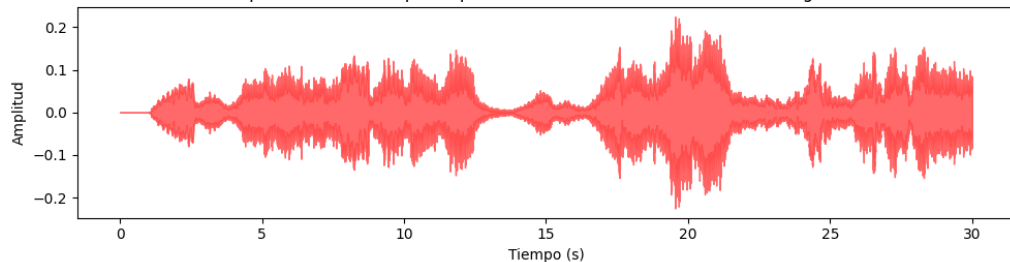
Introducción

Conjunto de datos

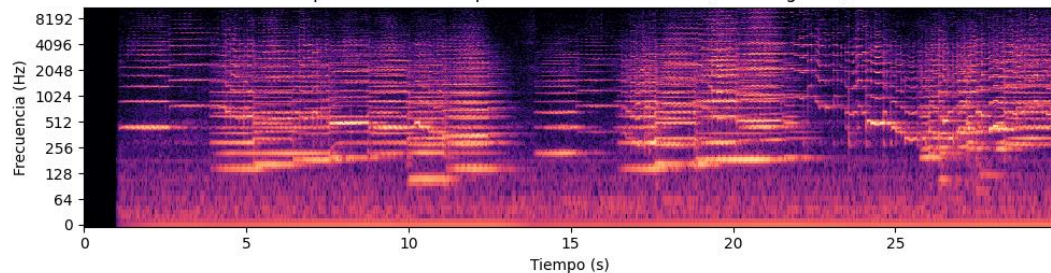
Extracción de características

Entrenamiento de modelos

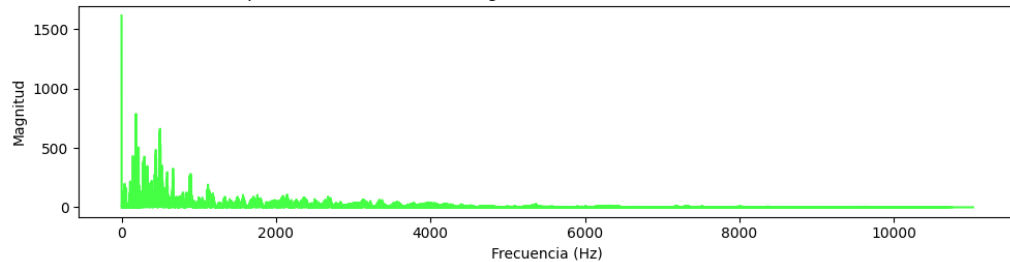
Representación tiempo-amplitud de la señal (Forma de onda/oscilograma)



Representación tiempo-frecuencia de la señal (estetrograma)



Representación frecuencia-magnitud de la señal (Transformada de Fourier)





EXTRACCIÓN DE CARACTERÍSTICAS

Dominios en señales de audio

Características de una señal de audio

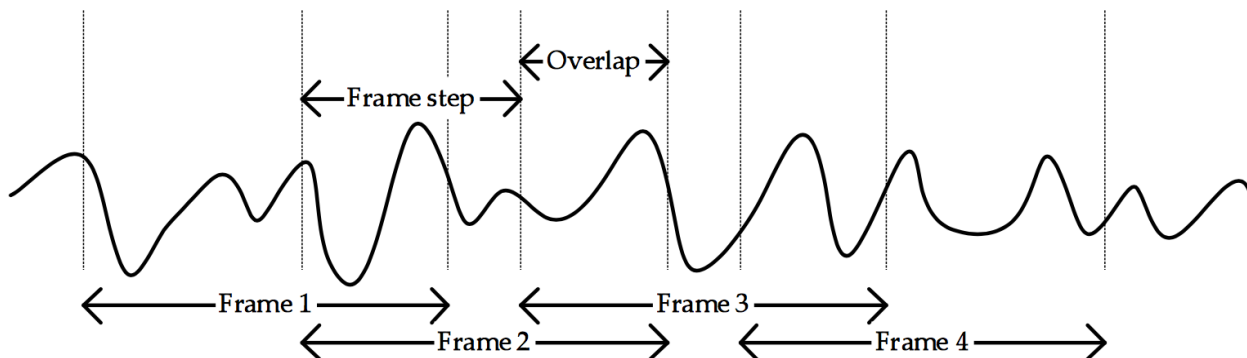
Generación de dataset de características
y caso práctico (análisis descriptivo)





Notación: sea S una señal segmentada en bloques

- s_i ($i = 0, \dots, N-1$) \equiv señal con sample rate sr
- $F \equiv$ tamaño de bloque (frame size).
- $H \equiv$ salto de bloque (hop).
- Bloques $k = 1, \dots, T - 1$. El k -ésimo bloque: $\left[\frac{H*k}{sr}, \frac{H*k+F-1}{sr}\right]$





ENVOLVENTE

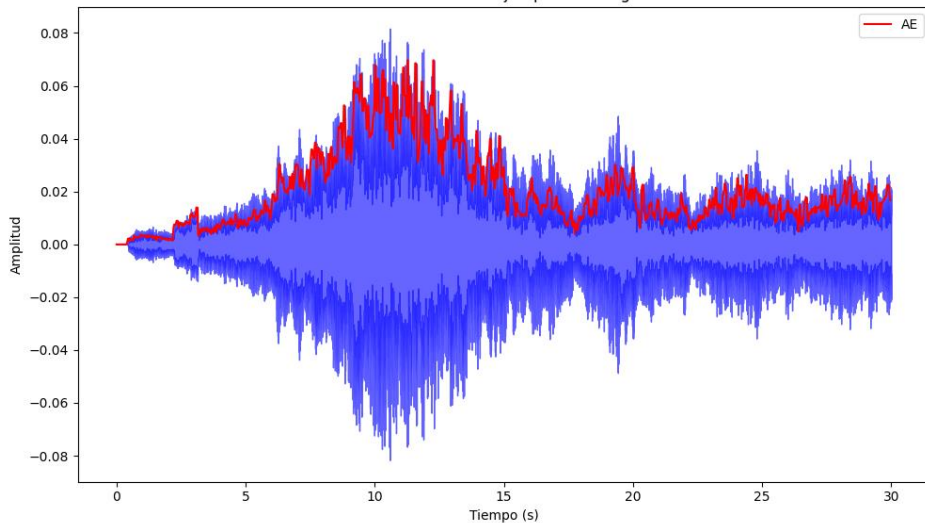
Dominio temporal.

Borde o silueta de la señal. Da una intuición más interpretable de cómo varía la señal a lo largo del tiempo.

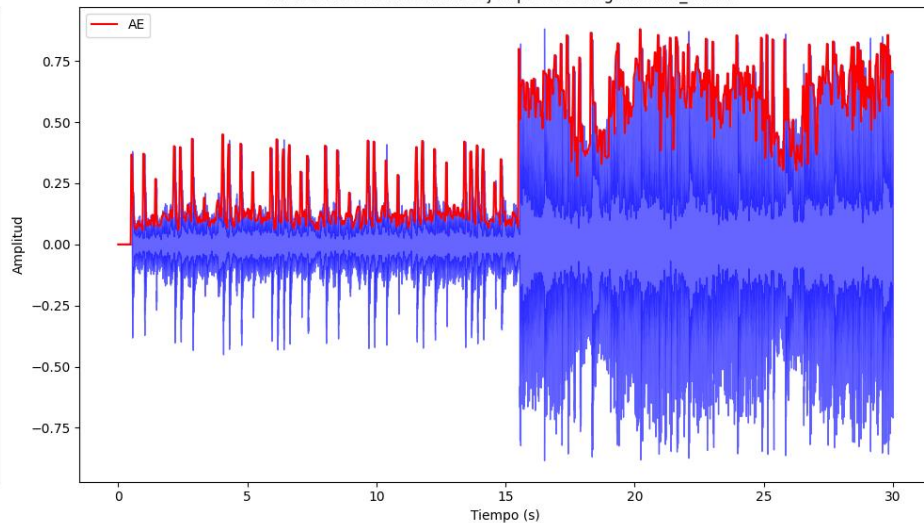
$$AE_k = \max_{\{i=kH\}}^{\{kH+F-1\}} s(i)$$



Envolvente de la señal de ejemplo de categoría Classic



Envolvente de la señal de ejemplo de categoría Non_classic



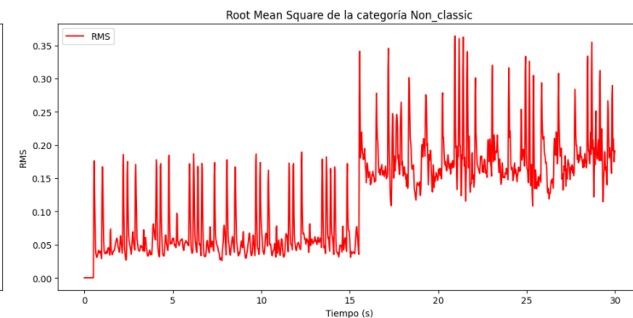
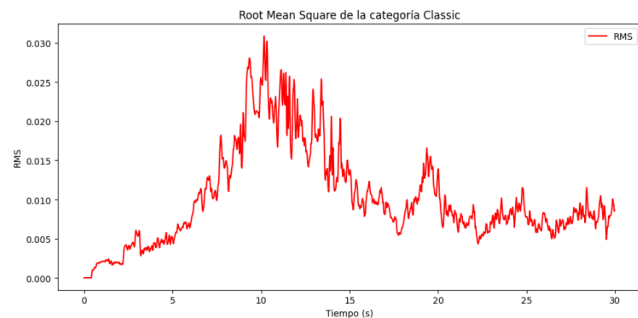
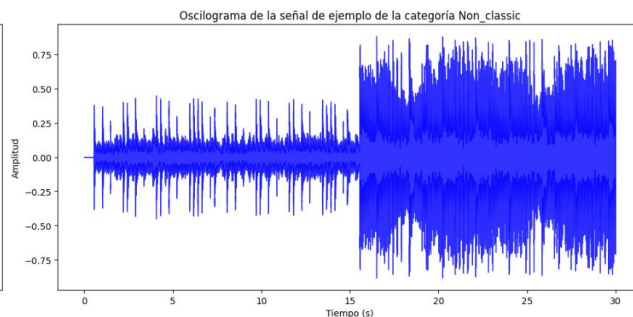
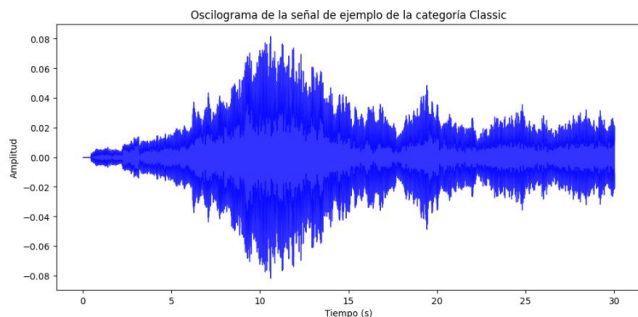


ROOT MEAN SQUARE

Dominio temporal.

Estima la energía de la señal a lo largo del tiempo. Sirve para detectar silencios y la dinámica de la señal.

$$RMS_k = \sqrt{\frac{1}{F} \cdot \sum_{i=k \cdot F}^{(k+1) \cdot F - 1} s(i)^2}$$



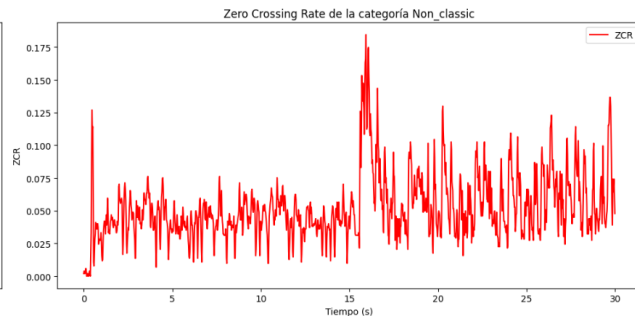
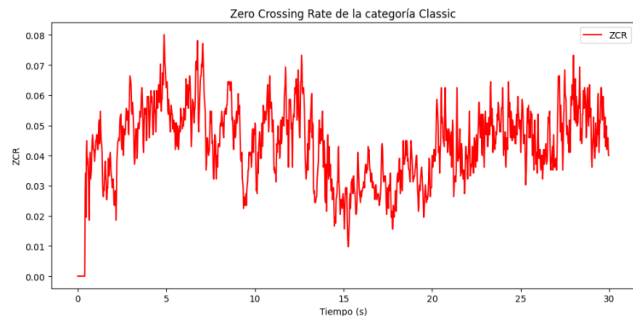
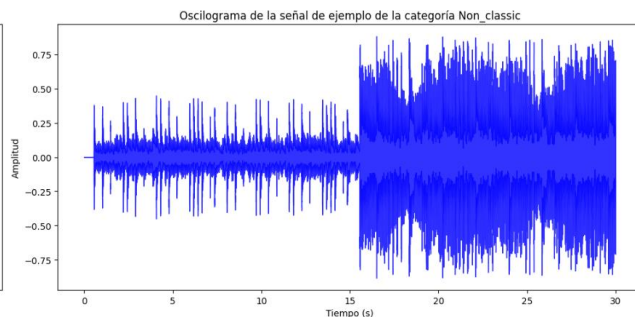
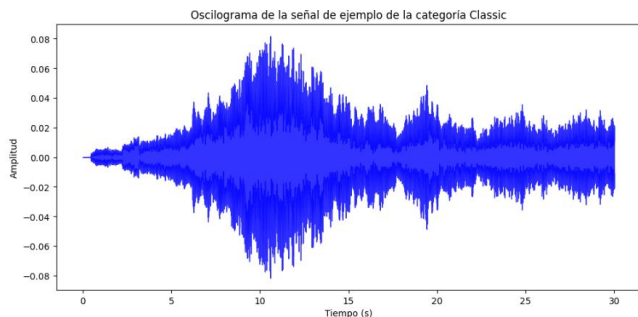


ZERO CROSSING RATE

Dominio temporal.

Calcula el promedio de cuántas veces la amplitud de la señal cruza el eje horizontal.

$$ZCR_k = \sum_{i=k \cdot F}^{(k+1) \cdot F - 1} \frac{1}{2} |\text{sgn}(s(i)) - \text{sgn}(s(i+1))|$$



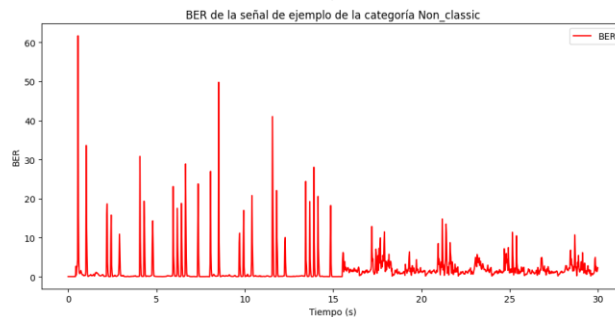
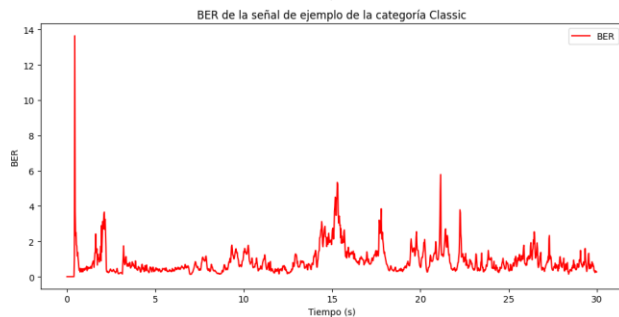
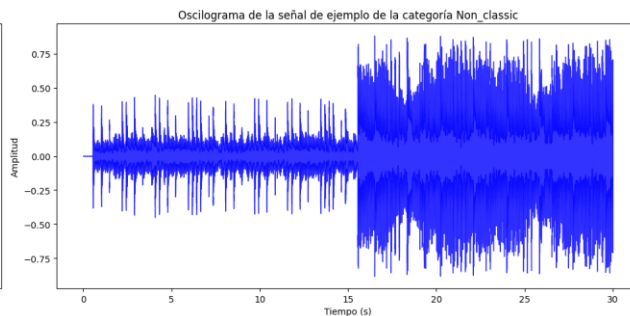
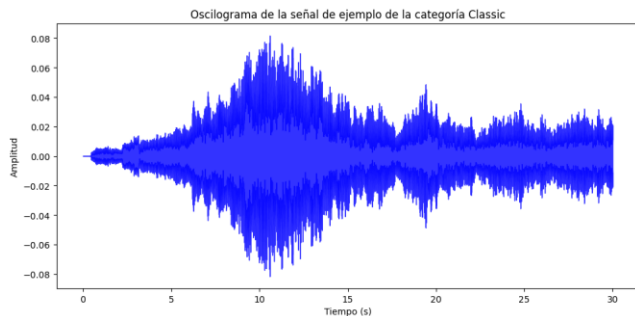


BER

Dominio frecuencial.

Cuánta energía acumulada hay en las frecuencias bajas frente a las frecuencias altas.

$$\text{BER}_k = \frac{\text{Energía banda baja}_k}{\text{Energía banda alta}_k} = \frac{\sum_{n=0}^{FR-1} m_k(n)^2}{\sum_{n=FR}^N m_k(n)^2}$$



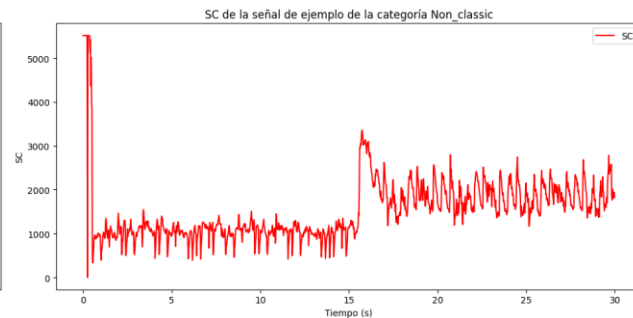
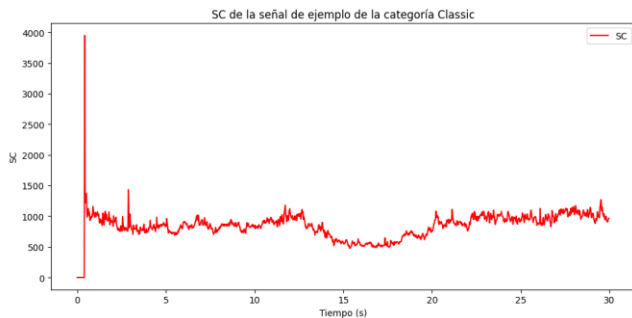
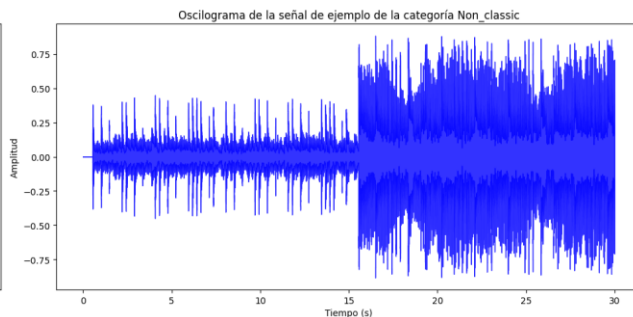
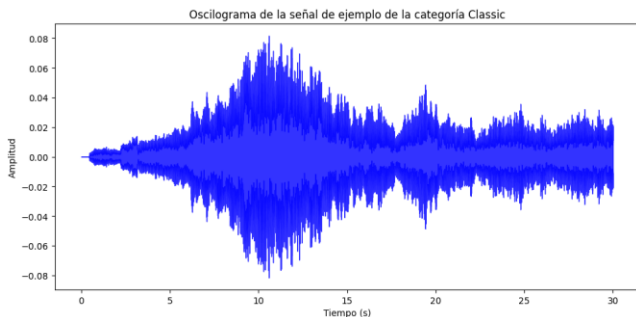


SPECTRAL CENTROID

Dominio frecuencial.

Centro de gravedad, banda de frecuencias en torno a la cual se concentra la mayor parte de la energía.

$$SC_k = \frac{\sum_{n=0}^N n \cdot m_k(n)}{\sum_{n=0}^N m_k(n)}$$



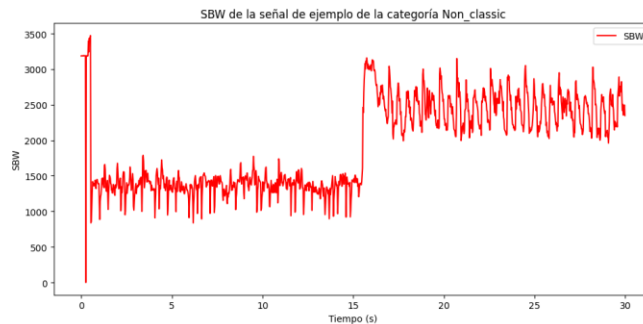
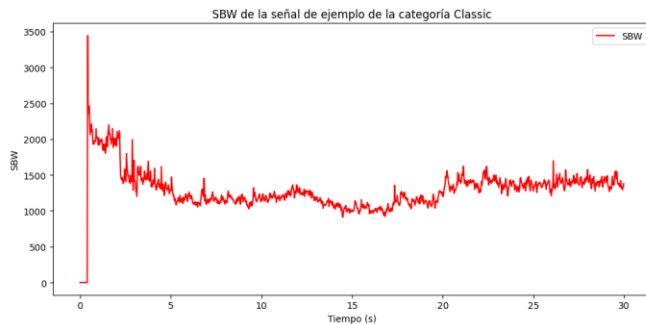
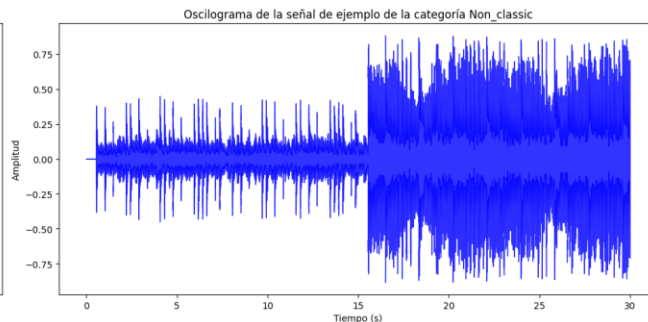
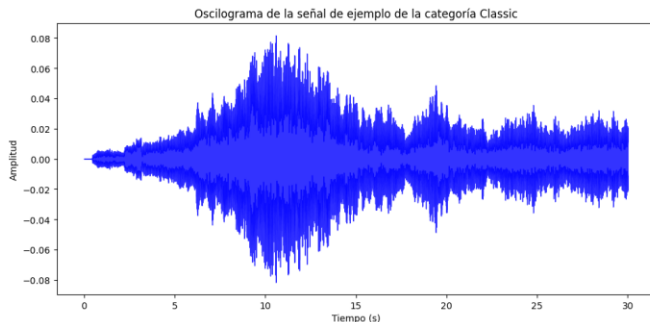


SPECTRAL BANDWIDTH

Dominio frecuencial.

Cómo de dispersas están las frecuencias en la señal de audio.

$$SBW_k = \sqrt{\frac{\sum_{n=0}^N (n - SC_k)^2 \cdot m_k(n)}{\sum_{n=0}^N m_k(n)}}$$



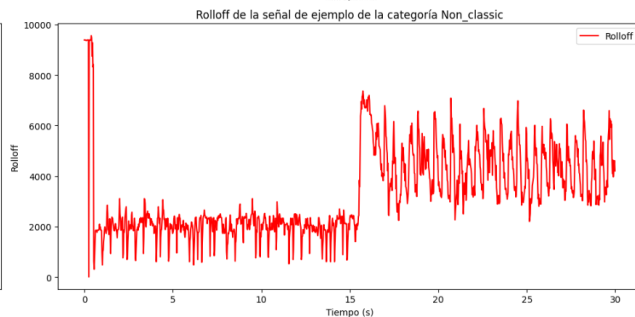
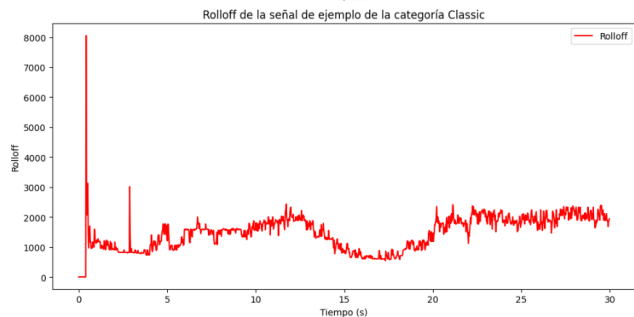
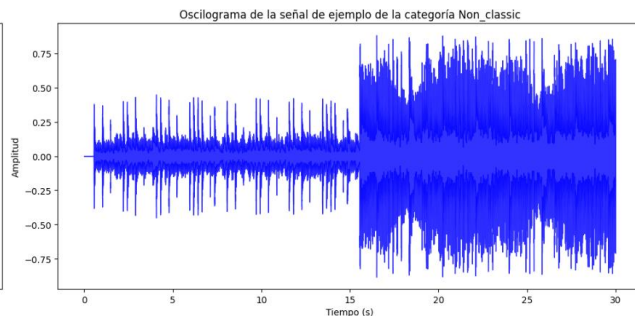
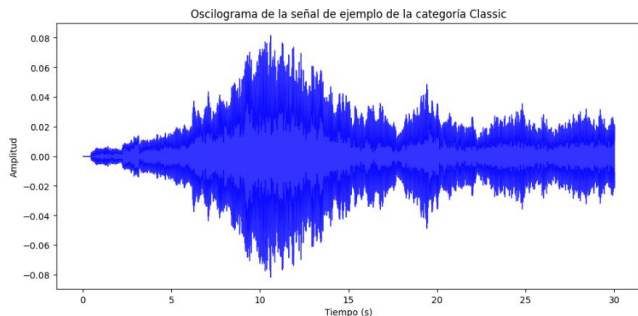


SPECTRAL ROLLOFF

Dominio frecuencial.

Determina el límite superior de las frecuencias del audio, reflejando el punto por debajo del cual se encuentra un porcentaje determinado de la energía espectral total.

$$R(k) = \min \left(\omega : \sum_{i=0}^{\omega} |X(k, i)| \geq 0.85 \times \sum_{i=0}^{N-1} |X(k, i)| \right)$$



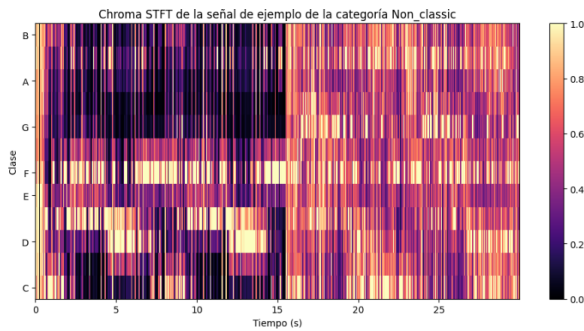
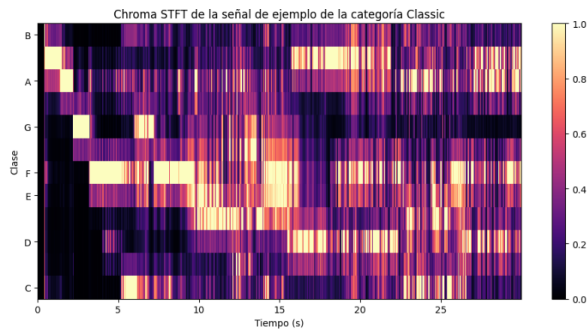
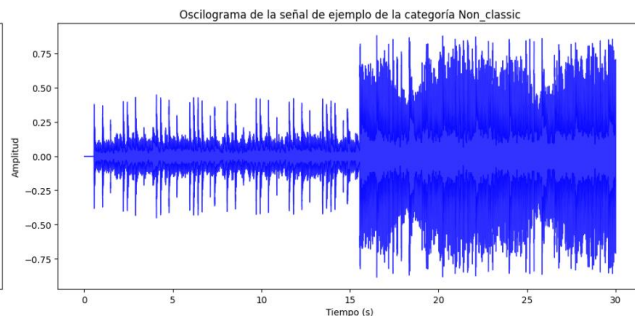
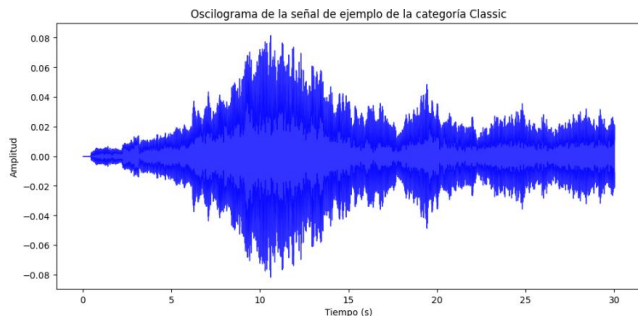


CHROMA STFT

Dominio frecuencial.

Representa el audio en términos de las 12 notas de la escala cromática.

$$C(k, m) = \sum_{\omega \in \text{bin}(m)} |X(k, \omega)|$$





Introducción

Conjunto de
datos

Extracción de
características

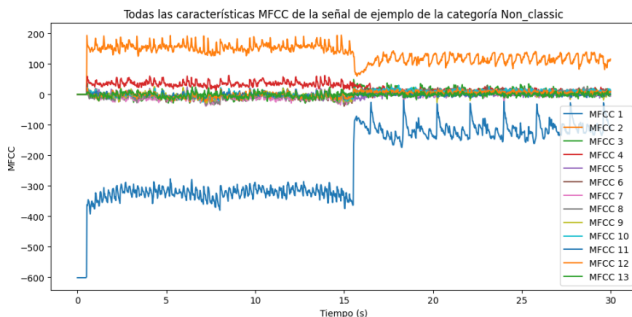
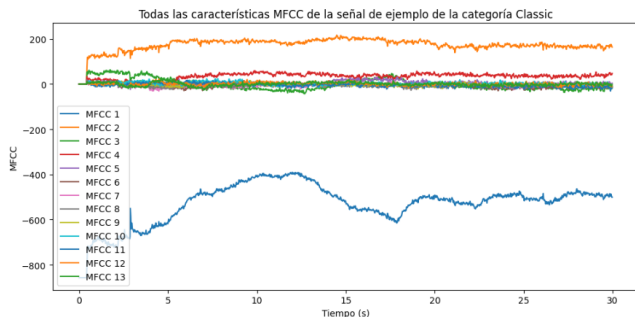
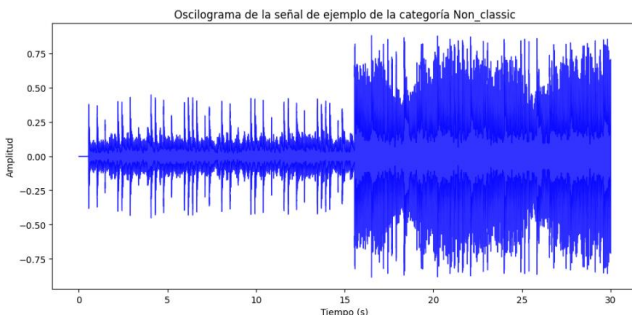
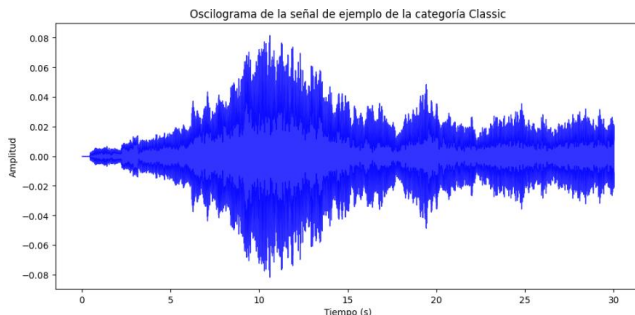
Entrenamiento de
modelos

MEL FREQUENCY CEPSTRAL COEFFICIENTS

Dominio frecuencial.

*Capturan características de la señal basándose en la percepción
auditiva humana.*

$$MFCC(k, c) = \text{DCT}(\log(M(k, m))) = \sum_{m=0}^{M-1} \log(M(k, m)) \cdot \cos\left(\frac{\pi c(2m+1)}{2M}\right)$$





EXTRACCIÓN DE CARACTERÍSTICAS

Dominios en señales de audio

Características de una señal de audio

Generación de dataset de características
y caso práctico (análisis descriptivo)

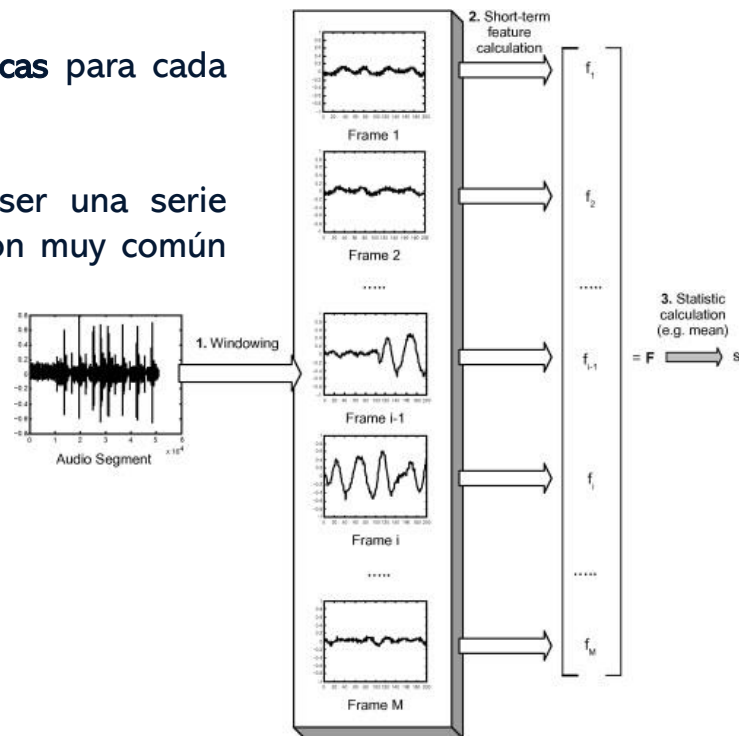




GENERACIÓN DEL DATASET DE CARACTERÍSTICAS

1. Definir funciones para el cálculo y generar las características para cada señal de cada partición.

2. Aproximación “*naive*”: para cada característica pese a ser una serie tenemos que almacenar un único valor tabular. Una aproximación muy común en la literatura es asociar la media.





ANÁLISIS DESCRIPTIVO

32 columnas de características

	audio_file	label	mean_envelope	mean_rms	mean_zcr	mean_ber	mean_spec_cent	mean_spec_bw	mean_rolloff	mean_chroma_stftC
0	ccmusic2/train/audios/audio_train_0.wav	Rock	0.118342	0.043627	0.060294	4.037650	1350.649029	1758.828856	2675.950486	0.303799
1	ccmusic2/train/audios/audio_train_1.wav	Soul_or_r_and_b	0.314424	0.103811	0.138103	4.370842	2636.363229	2593.828616	5751.798553	0.449251
2	ccmusic2/train/audios/audio_train_2.wav	Symphony	0.181952	0.065241	0.090361	1.856970	1452.552736	1590.335734	2817.732962	0.370835
3	ccmusic2/train/audios/audio_train_3.wav	Dance_and_house	0.163407	0.063564	0.090101	6.817037	1675.637336	1715.368994	3338.713145	0.240659
4	ccmusic2/train/audios/audio_train_4.wav	Soul_or_r_and_b	0.348335	0.128825	0.100390	17.659496	2222.363681	2569.843206	4858.120624	0.570061

DATASET TABULAR

→ ETC



¿SERÁ SUFICIENTE PARA
DISTINGUIR GÉNEROS?



¿DISTINGUEN LAS CARACTERÍSTICAS
REALMENTE DISTINTOS GÉNEROS
MUSICALES?

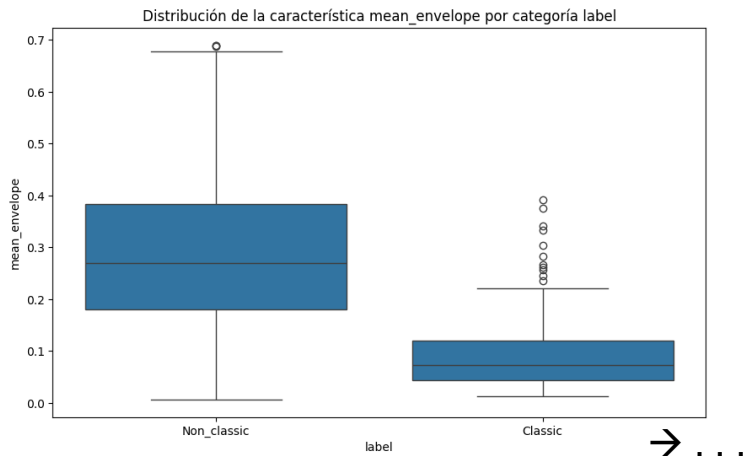


ANÁLISIS DESCRIPTIVO

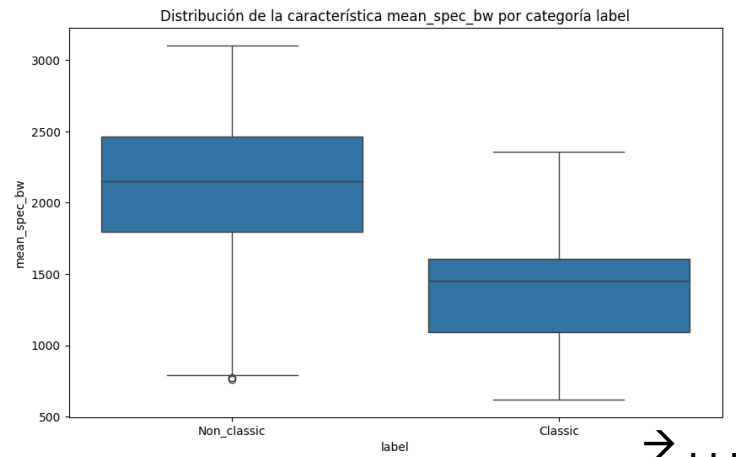
- **Análisis descriptivo** para comprobar si las características distinguen entre ambas clases.
- **Dos enfoques:** boxplots y correlaciones.

BOXPLOTS

Dominio temporal: Envelope



Dominio frecuencial: Spectral bandwidth





Introducción

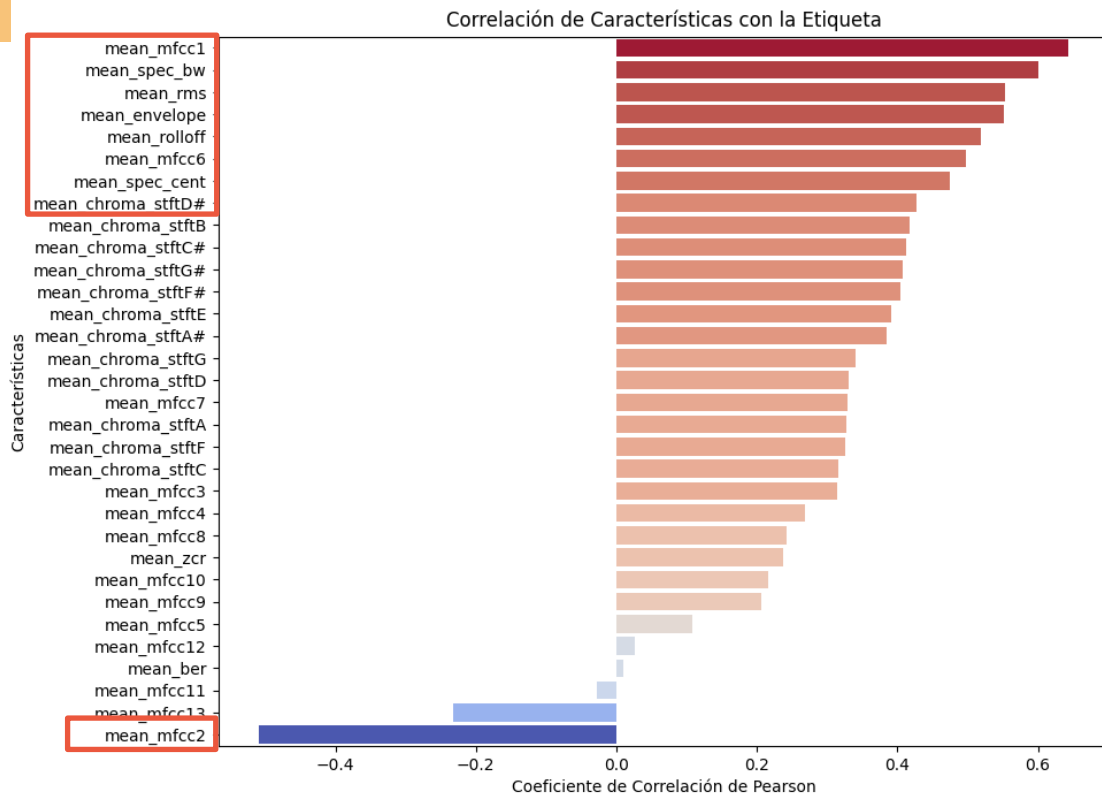
Conjunto de
datos

Extracción de
características

Entrenamiento de
modelos

ANÁLISIS DESCRIPTIVO

CORRELACIÓN LINEAL



04

ENTRENAMIENTO DE MODELOS

- 1) Corpus tabular
- 2) Entrenamiento e inferencia
modelo características



Introducción

Conjunto de
datos

Extracción de
características

Entrenamiento de
modelos

CREACIÓN CORPUS



```
class TabularDataset(torch.utils.data.Dataset):  
  
    def __init__(self, features_file, scaler=None):  
  
        df = pd.read_csv(features_file)  
  
        self.X = df.drop(['audio_file', 'label'], axis=1)  
        self.y = df['label'].values.astype(np.int64)  
  
        if scaler:  
            self.X = scaler.transform(self.X)  
        else:  
            self.scaler = StandardScaler()  
            self.X = self.scaler.fit_transform(self.X)  
  
    def get_scaler(self):  
        return self.scaler  
  
    def __len__(self):  
        return len(self.X)  
  
    def __getitem__(self, idx):  
        if isinstance(idx, torch.Tensor):  
            idx = idx.tolist()  
  
        return torch.tensor(self.X[idx], dtype=torch.float32), torch.tensor(self.y[idx], dtype=torch.long)
```



CREACIÓN MODELO



GREEN AI

- MLP con 3 capas (128, 64 y 1 neuronas).
- 50 épocas. Parada temprano sobre el conjunto de validación con (paciencia=5).
- Tamaño de batch de 32 y learning rate de 0.001.
- Función de activación BCEWithLogitsLoss, optimizador Adam.
- Mismo entrenamiento podría desempeñarse utilizando SVC o RF.



Introducción

Conjunto de
datos

Extracción de
características

Entrenamiento de
modelos

RESULTADOS

Época	Loss (Validation)
Epoch 1	0.0883
Epoch 10	0.0002
Epoch 25	1.7858 e-05

Métrica	Valor (Test)
Accuracy	0.96
F1 Score	0.97

05

CONCLUSIONES

Y CIERRE

- 1) Conclusiones
- 2) Trabajos futuros

CONCLUSIONES



Extracción previa de características que ayudan a distinguir géneros musicales.



Buenos resultados con modelos menos costosos computacionalmente

TRABAJO FUTURO



Clasificación para las CCMUSIC con jerarquías más detalladas



Entrenamiento con espectrogramas y modelos TSF. Comparativa.



Modelo híbrido/multimodal



**MUCHAS GRACIAS
POR SU ATENCIÓN**

