**Sistema inteligente de clasificación por voz**

**Autor**: Antoni Cano Aladid

Curso de especialización en big data e inteligencia artificial

IES Pere Maria Orts

# Índice

Contenido

[Índice 2](#_Toc188667534)

# Introducción

Tenemos el dataset **Audio MNIST**, un conjunto de datos que contiene 30.000 muestras de audio de dígitos hablados (0-9) por 60 hablantes diferentes.

## Objetivo

Construir un sistema que pueda extraer información del audio y realizar varios sistemas de clasificación.

1. **Clasificación por género**: Determinar si la voz pertenece a un hombre o a una mujer.
2. **Clasificación del acento**: Identificar el acento del hablante.
3. **Clasificación del origen**: Determinar el país o región de origen del hablante.
4. **Clasificación de hablante nativo o no nativo**: Estimar si el hablante es nativo del idioma.
5. **Clasificación de la edad**: Predecir el rango de edad del hablante.

# Metodología

## Preprocesamiento

Tenemos un archivo .txt que contiene la información de los 60 participantes del dataset. Contiene información relevante tal como:

* Acento
* Edad
* Género
* Hablante nativo
* Origen

Con este archivo y los audios procedemos a realizar un script de Python para automatizar la tarea de procesar estos audios y obtener de estos las características que nos interesen.

Para obtener estos datos vamos a hacer uso de la librería de Python [Librosa](https://librosa.org/doc/latest/index.html)

Con esta librería obtenemos características básicas de los audios, como centroides espectrales, MFCC (coeficientes cepstrales de frecuencia mel), tasa de cruces por cero, etc.

Estos datos los vamos a guardar en un .csv para tenerlos guardados y no tener que procesar los audios cada vez que queramos entrenar un modelo de inteligencia artificial.

### Características

#### Chroma Features

Representan las energías en las 12 notas musicales de una octava. Útil para analizar tonalidad y armonía.

#### RMS

Energía de la señal, útil para medir intensidad.

#### Spectral Centroid

Representa el "centro de gravedad" del espectro. Indica qué tan brillante es un sonido.

#### Spectral Bandwidth

Mide la dispersión del espectro alrededor del centroid. Ayuda a distinguir entre sonidos suaves y agresivos.

#### Spectral Rolloff

Frecuencia por debajo de la cual se encuentra un porcentaje acumulado (generalmente 85%) de la energía espectral. Relacionado con la cantidad de energía en las frecuencias altas.

#### Zero-Crossing Rate

Número de veces que la señal cruza por el eje cero.

Útil para clasificar sonidos con patrones de vibración (como consonantes frente a vocales).

#### MFCC(Mean y Var)

Captura la envolvente del espectro en la escala mel. Es crucial para tareas como reconocimiento de voz o análisis del timbre.

Se han tomado estas características porque la clasificación por género, acento, origen y edad requieren características que capturen propiedades del timbre, intensidad, y estructura espectral de la voz. MFCC y Spectral Centroid son ideales para capturar el timbre y las diferencias en el habla. RMS, Spectral Bandwidth, y Spectral Rolloff ayudan a diferenciar entre voces graves y agudas (útil para distinguir géneros y edades).

### Modelos

Se han elegido los siguientes modelos:

#### Decision Tree

Para los clasificadores de género y nativo/no nativo

Los árboles de decisión son intuitivos, fáciles de interpretar y efectivos para manejar datos categóricos y numéricos, como los que tienes (género, acento, origen, etc.). Además:

* Pueden capturar relaciones no lineales en los datos.
* No requieren mucho preprocesamiento.
* Funcionan bien para problemas con características claramente diferenciadas.

#### Máquina de Soporte Vectorial

Para los clasificadores de acento y origen

Las SVM son muy potentes para encontrar límites claros entre clases, especialmente cuando los datos no son linealmente separables. Además:

* Funcionan bien con datos de alta dimensionalidad (como características extraídas de audios).
* El kernel trick permite manejar relaciones no lineales entre las características.

#### K-Nearest Neighbors (KNN)

Para la clasificación por edad

KNN clasifica basándose en la similitud entre las observaciones. Es una técnica sencilla y efectiva, especialmente si tienes un conjunto de datos relativamente equilibrado.

* No asume ninguna distribución de los datos, lo que es útil en problemas donde las características tienen patrones complejos.
* Los datos de audio, al tener múltiples dimensiones, pueden beneficiarse de este enfoque basado en distancias.

#### DNN

TBD