**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](#_Toc188797776)

[Introducción 3](#_Toc188797777)

[Objetivo 3](#_Toc188797778)

[Metodología 4](#_Toc188797779)

[Preprocesamiento 4](#_Toc188797780)

[Análisis previo características generales 5](#_Toc188797781)

[Características voz seleccionadas 7](#_Toc188797782)

[Análisis características específicas 9](#_Toc188797783)

[Modelos 12](#_Toc188797784)

[Resultados 14](#_Toc188797785)

[Decision Tree 14](#_Toc188797786)

[Género 14](#_Toc188797787)

[Nativo/No nativo 15](#_Toc188797788)

[Máquina de Soporte Vectorial 16](#_Toc188797789)

[Acento 16](#_Toc188797790)

[Origen 17](#_Toc188797791)

[KNN 19](#_Toc188797792)

[Edad 19](#_Toc188797793)

[Redes neuronales 20](#_Toc188797794)

[Acento 20](#_Toc188797795)

[Origen 21](#_Toc188797796)

# 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.

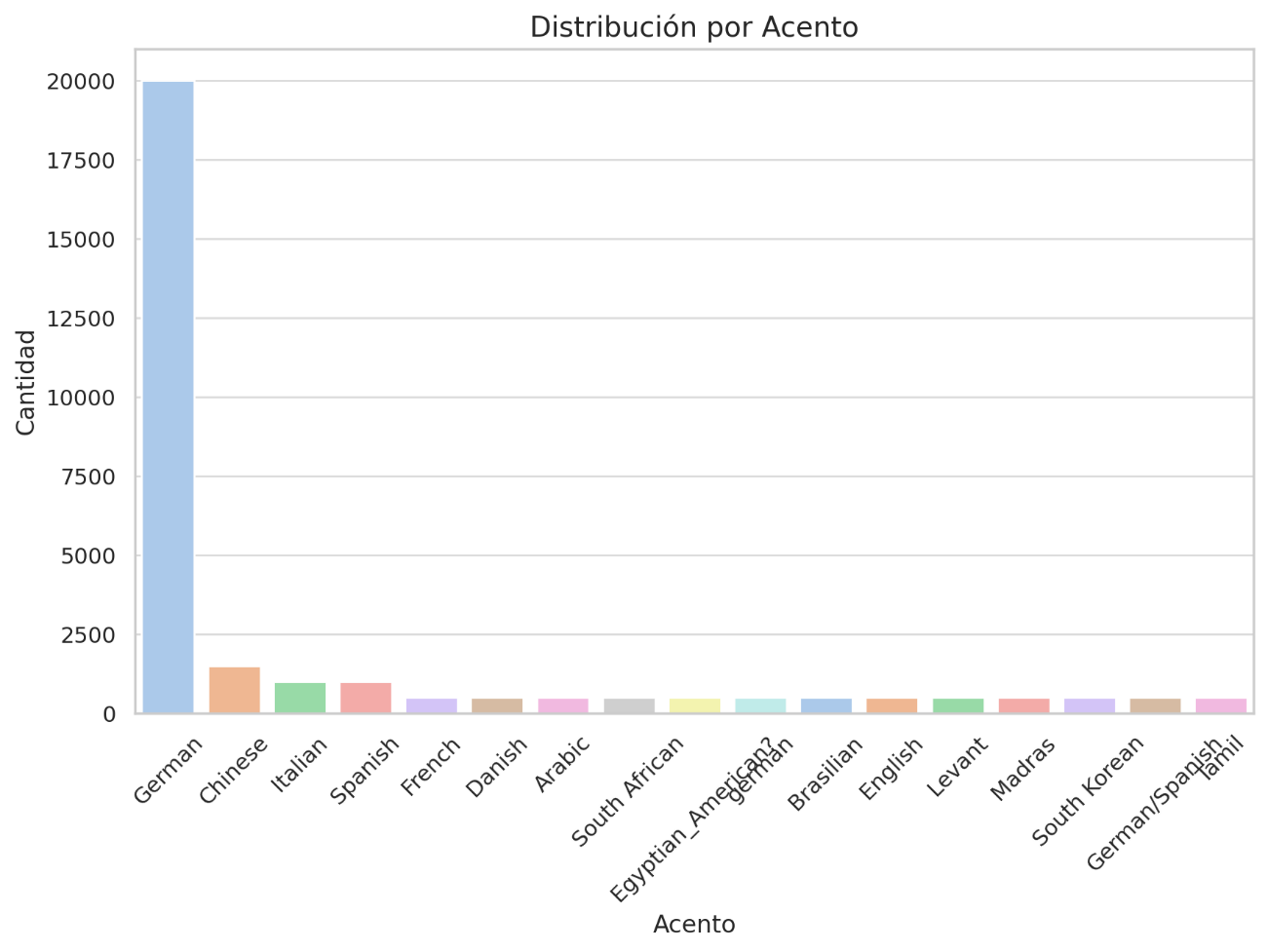
Estos datos los vamos a guardar en un .csv para tenerlos guardados y no tener que procesar los audios cada vez que queramos entrenar un modelo de inteligencia artificial.

El csv final con los datos es:

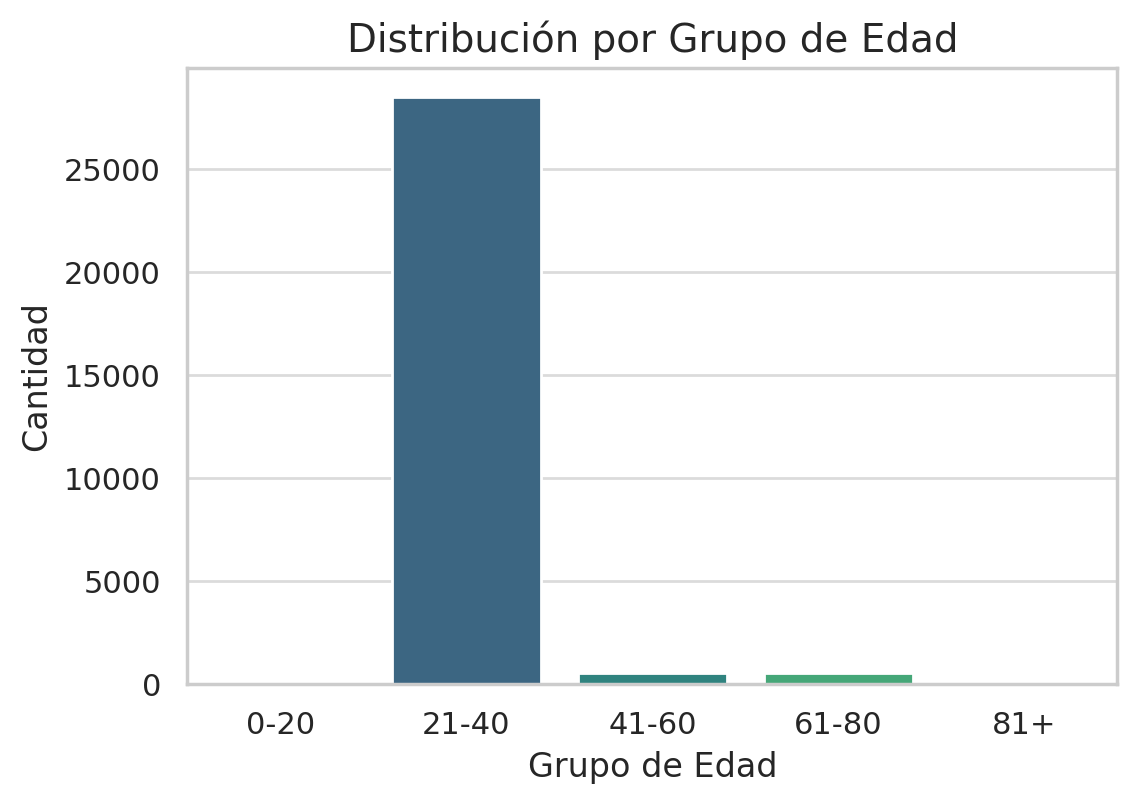
<audio_features.csv>

### Análisis previo características generales

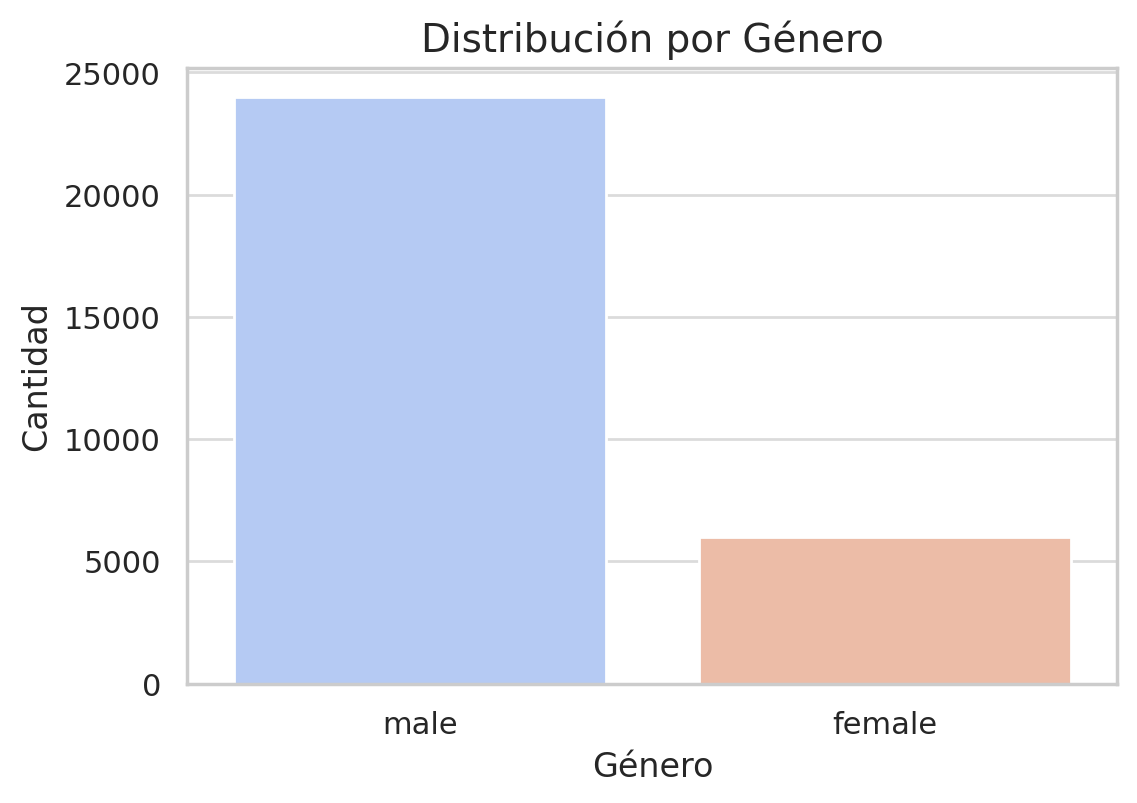
#### Acento



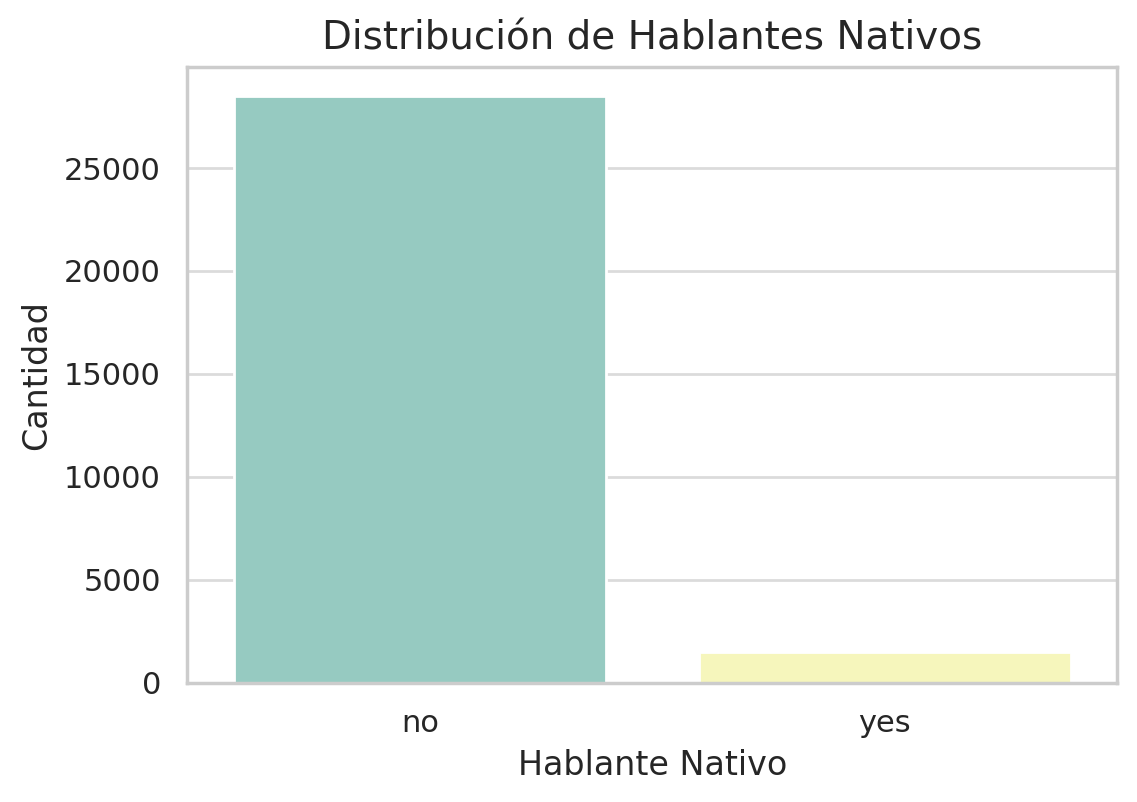
#### Edad



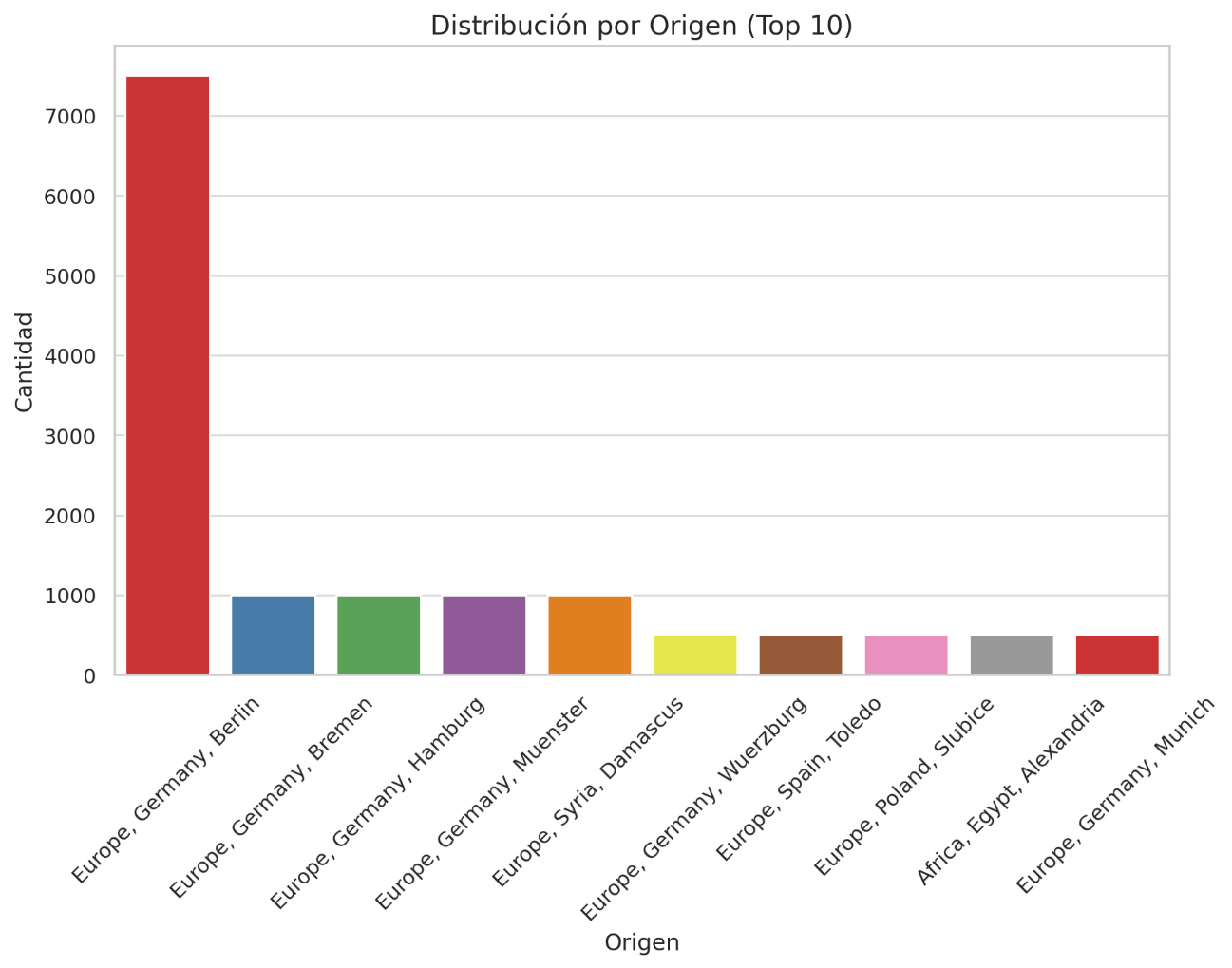
#### Genero



#### Nativos



#### Origen



### Características voz seleccionadas

#### 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).

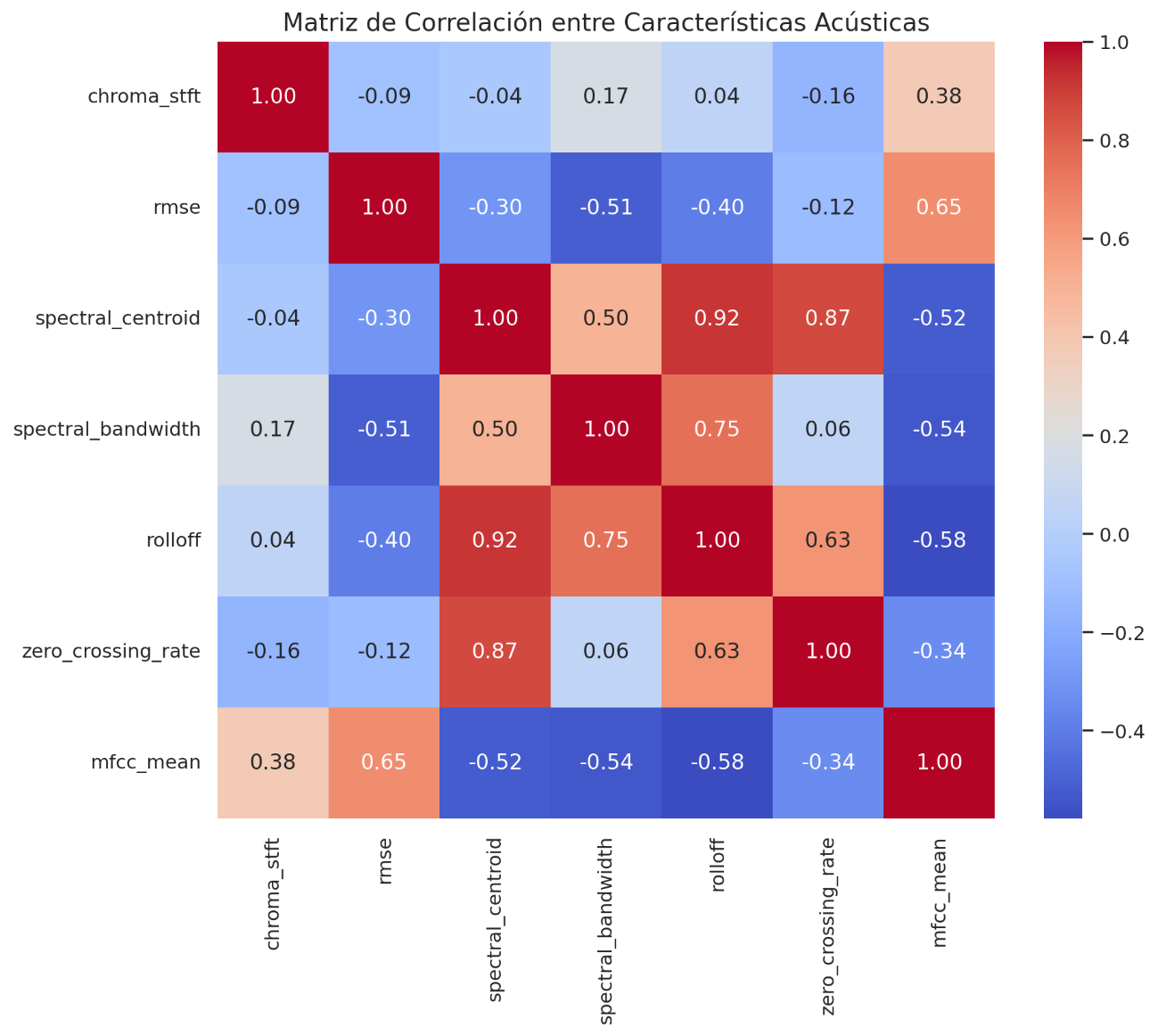
#### MFCC(Mean y Var)

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

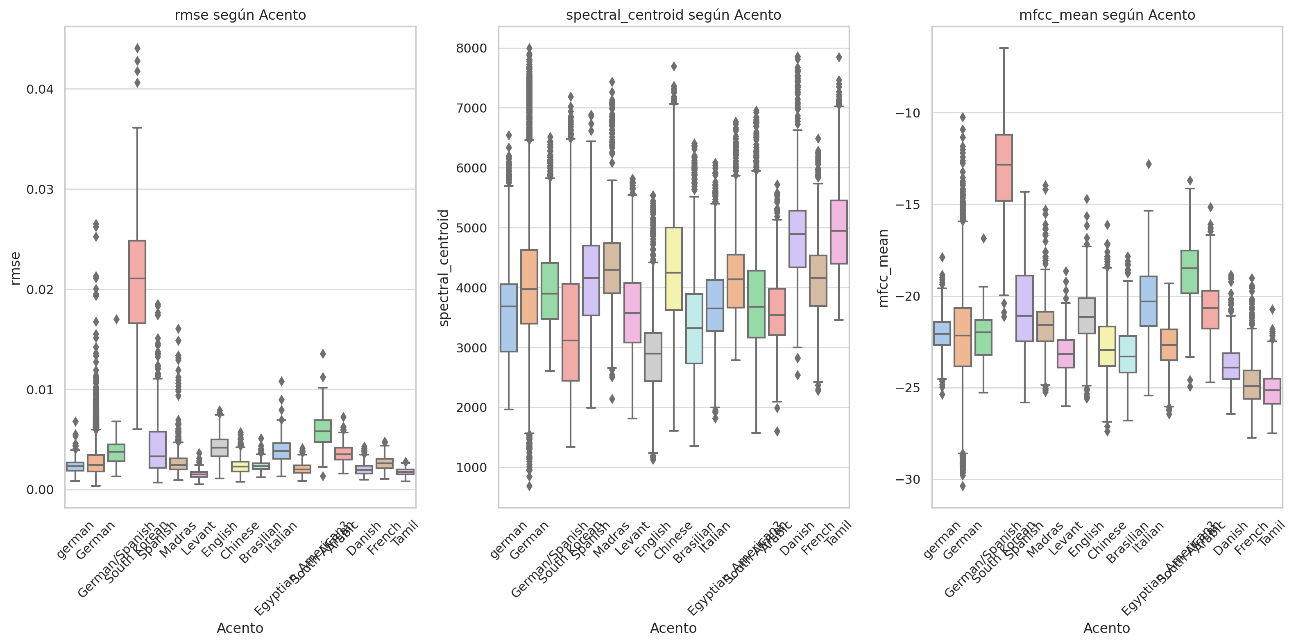
Se han tomado estas características porque la clasificación por género, acento, origen y edad requieren características que capturen propiedades del timbre, intensidad, y estructura espectral de la voz. MFCC y Spectral Centroid son ideales para capturar el timbre y las diferencias en el habla. RMS, Spectral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

### Análisis características específicas

#### Correlación



#### Media MFCC - Acento



#### Media MFCC - Edad

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

#### Media MFCC - Nativo

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

#### Media MFCC - Origen

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

#### Análisis resultados

##### Acento

Se observa que ciertos acentos tienen un rango más amplio de valores para características como rmse, spectral\_centroid y mfcc\_mean.

Por ejemplo, acentos como el alemán (german) muestran una mayor variabilidad en spectral\_centroid en comparación con otros acentos. Esto podría deberse a las características fonéticas del idioma o la forma de pronunciación.

Acentos con menor variabilidad podrían reflejar consistencia en la pronunciación entre hablantes de esa región.

##### Edad

La variable rmse (Root Mean Square Energy) muestra valores ligeramente más altos en el grupo de edad 21-40. Esto podría estar relacionado con una mayor energía vocal en adultos jóvenes en comparación con otros grupos.

spectral\_centroid tiende a ser más consistente en los grupos mayores (como 41-60 y 61-80), lo que podría reflejar una menor variabilidad en las frecuencias vocales con la edad.

mfcc\_mean no muestra diferencias significativas entre los grupos de edad, aunque podría haber ligeras variaciones en los extremos de edad.

##### Nativo

Los hablantes nativos presentan una mayor consistencia en las características como rmse y mfcc\_mean. Esto es esperable, ya que los hablantes nativos suelen tener un mayor control y fluidez en su idioma.

En contraste, los hablantes no nativos presentan una mayor dispersión en spectral\_centroid, posiblemente debido a variaciones en el aprendizaje y la pronunciación del idioma.

##### Origen(Top 5)

Hay diferencias notables en características como rmse y spectral\_centroid entre los orígenes analizados. Esto podría reflejar las influencias culturales y lingüísticas de cada región.

Algunos orígenes tienen una mayor dispersión (e.g., Europe, Germany), lo que podría deberse a la diversidad interna en términos de dialectos y acentos.

Orígenes más homogéneos muestran menor dispersión, lo que podría indicar una pronunciación más uniforme entre los hablantes.

##### Generales

Las características acústicas (rmse, spectral\_centroid, mfcc\_mean) son sensibles a factores como el acento, la edad y si el hablante es nativo o no. Esto sugiere que estas variables podrían ser útiles para clasificar o identificar hablantes según su perfil demográfico.

La correlación observada entre ciertas características acústicas indica que podrían estar relacionadas entre sí, lo que refuerza la idea de que las características acústicas no son independientes, sino que forman un conjunto interconectado.

### 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

Finalmente, como las redes neuronales sirven para cualquier clasificación se han utilizado para el acento y el origen.

# Resultados

## Decision Tree

### Género

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Imagen que contiene Texto

Descripción generada automáticamente

### Nativo/No nativo

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Texto

Descripción generada automáticamente

## Máquina de Soporte Vectorial

### Acento

Imagen que contiene Gráfico

Descripción generada automáticamente

Texto

Descripción generada automáticamente

### Origen

Tabla

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente con confianza media

## KNN

### Edad

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente con confianza baja

## Redes neuronales

### Acento

Gráfico

Descripción generada automáticamente

Texto

Descripción generada automáticamente con confianza baja

### Origen

Tabla

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Imagen que contiene Texto

Descripción generada automáticamente