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

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

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

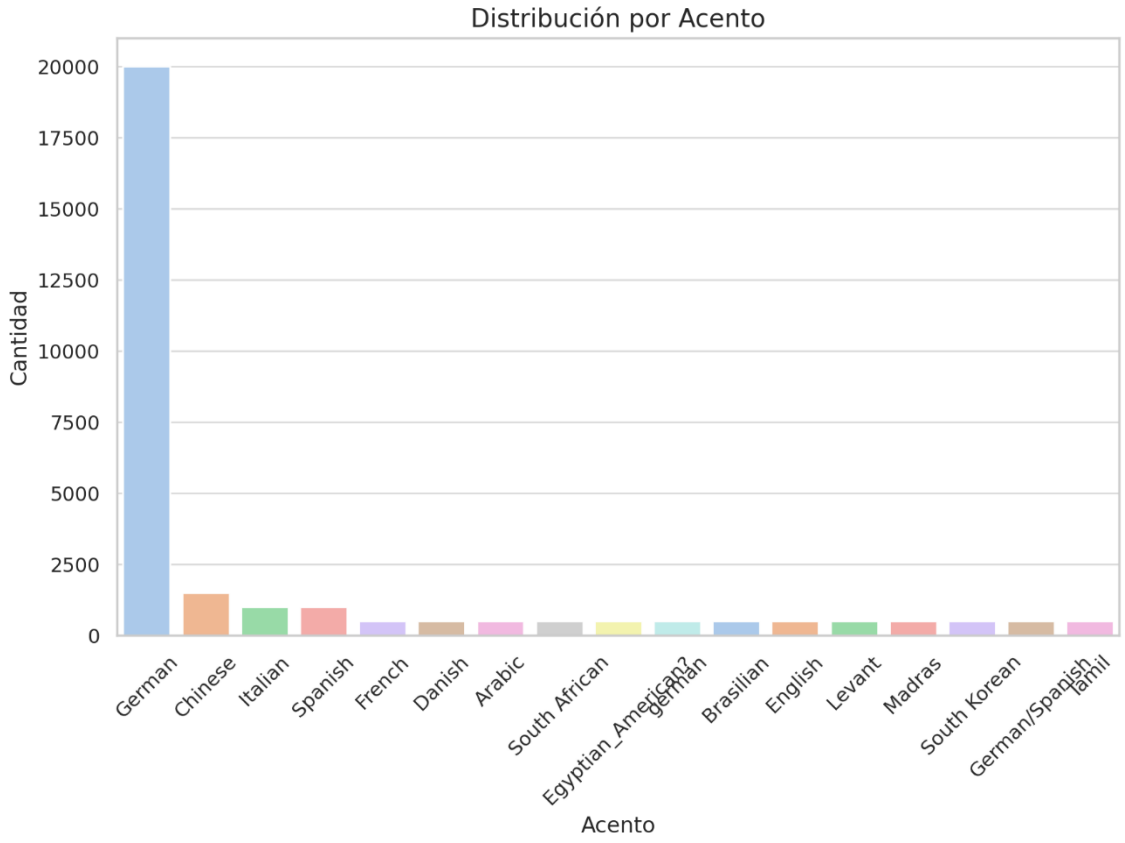
Estos datos los vamos a guardar en un .csv para tenerlos guardados y no tener que procesar los audios cada vez que queramos entrenar un modelo de inteligencia artificial.

El csv final con los datos es:

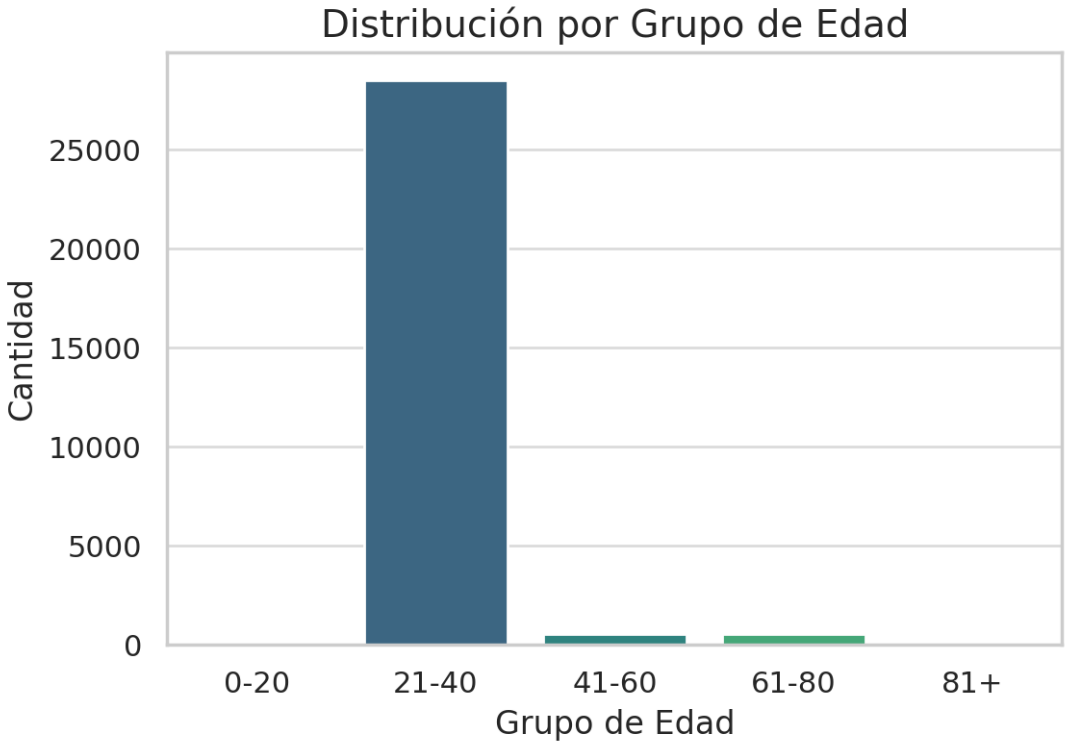
audio_features.csv

Análisis previo características generales

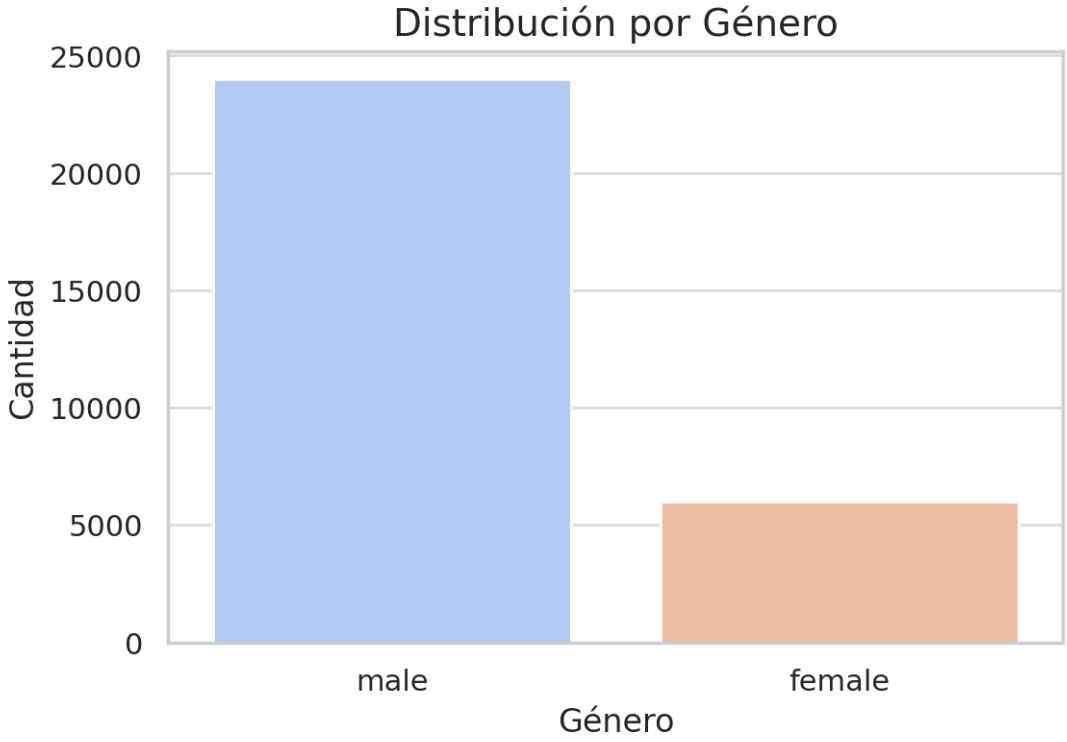
Acento



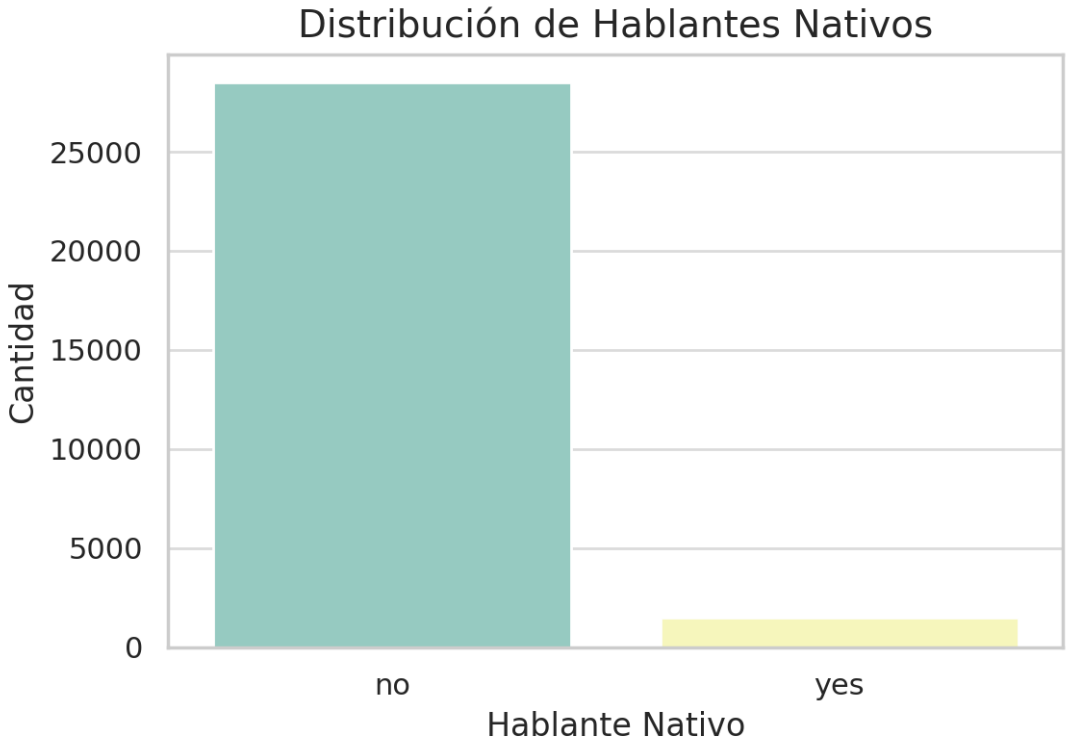
Edad



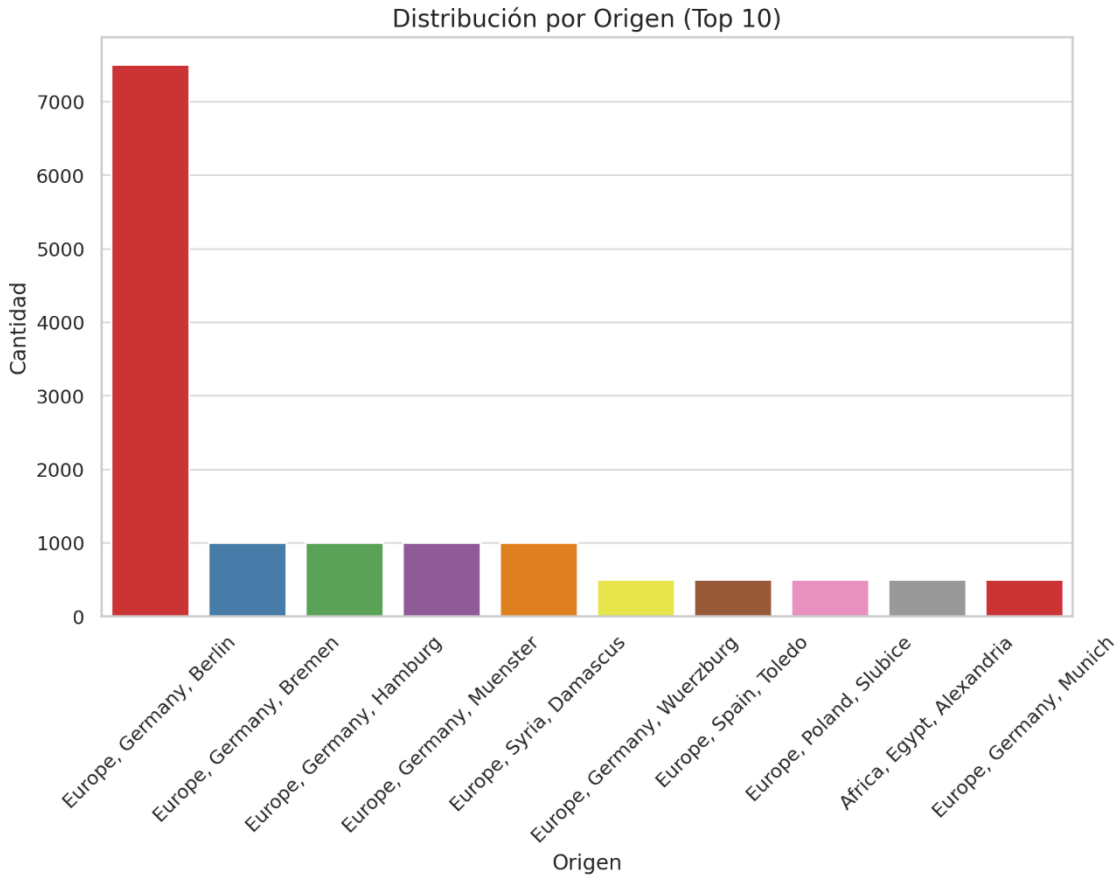
Genero



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).

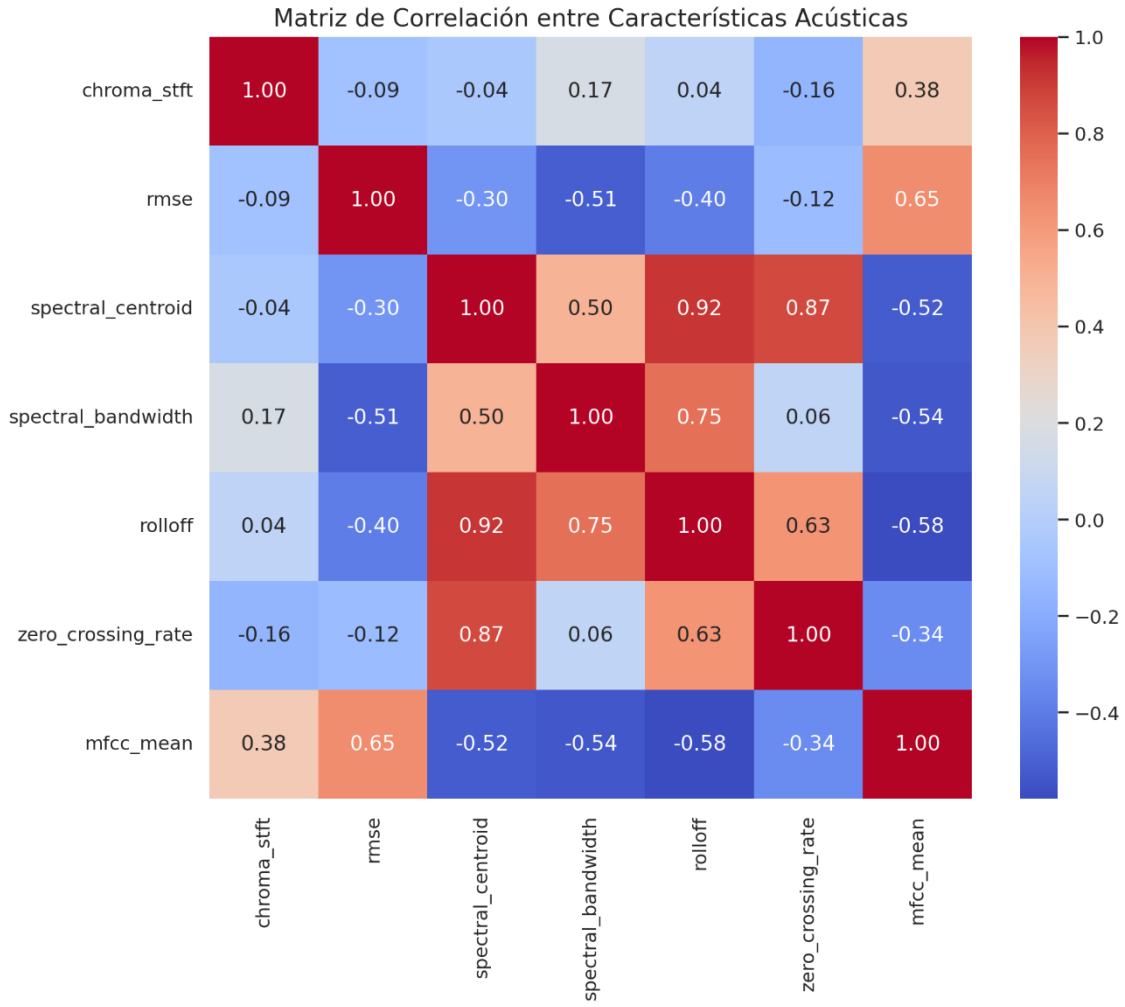
MFCC(Mean y Var)

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

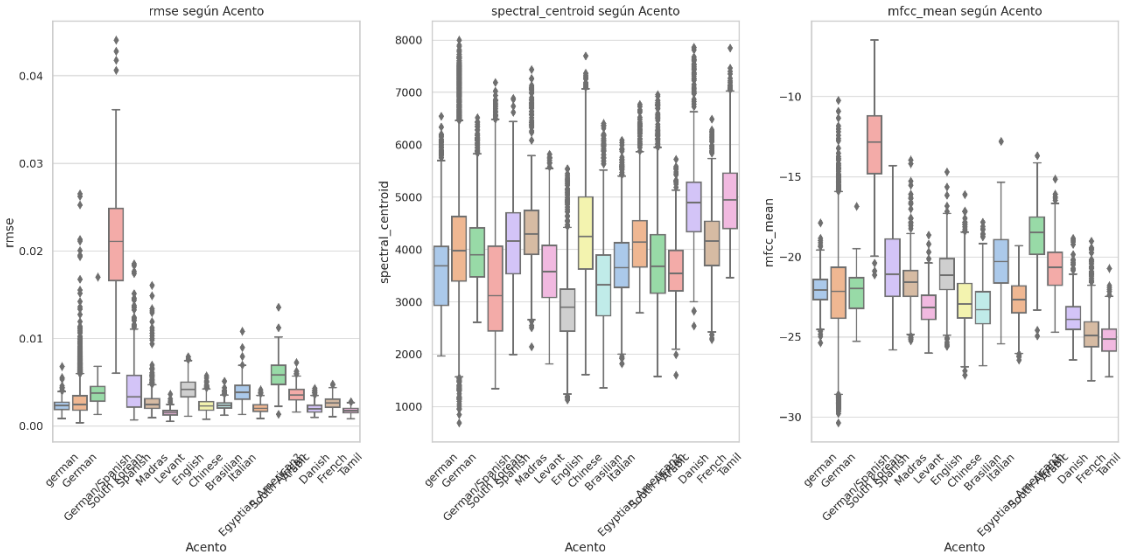
Se han tomado estas características porque la clasificación por género, acento, origen y edad requieren características que capturen propiedades del timbre, intensidad, y estructura espectral de la voz. MFCC y Spectral Centroid son ideales para capturar el timbre y las diferencias en el habla. RMS, Spectral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

Análisis características específicas

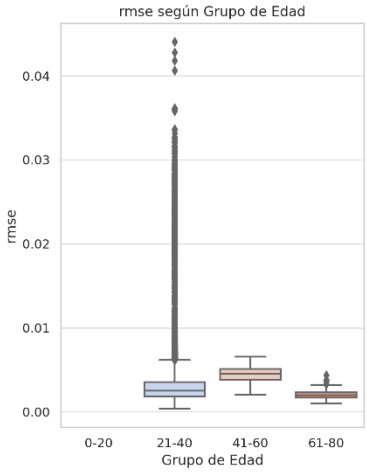
Correlación

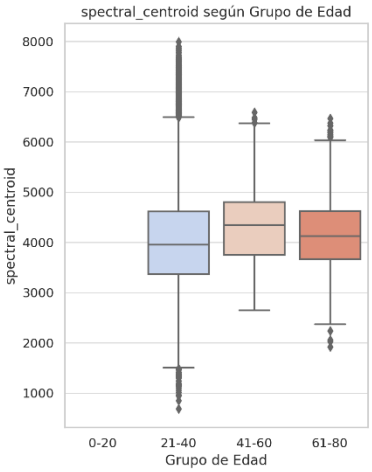


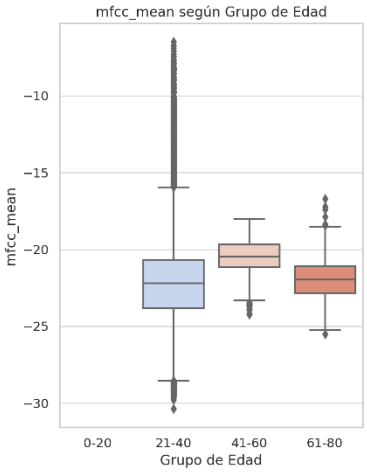
Media MFCC - Acento



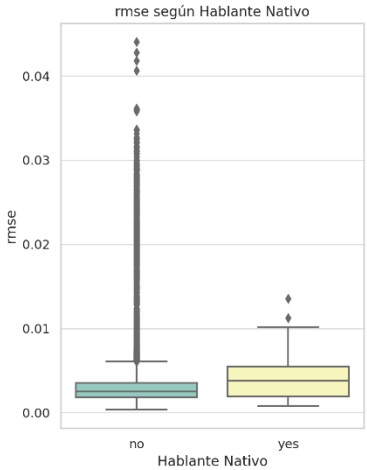
Media MFCC - Edad

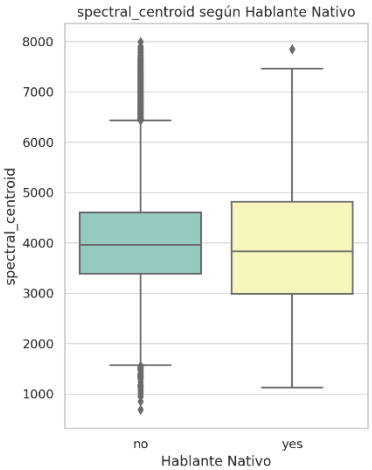


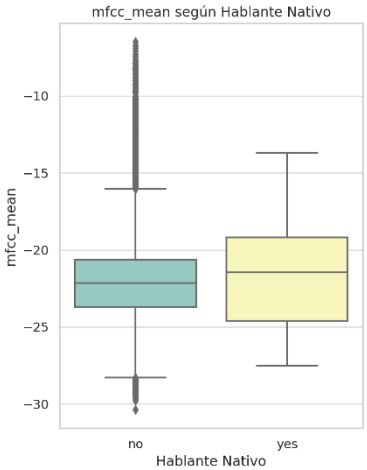




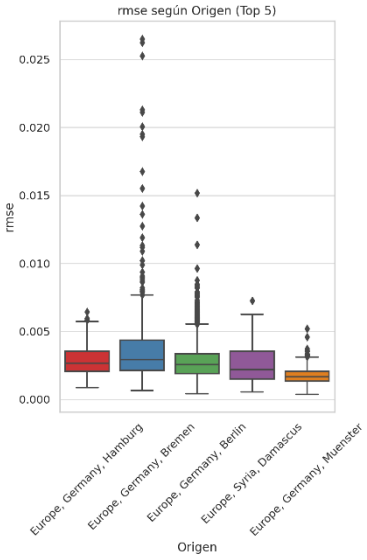
Media MFCC - Nativo

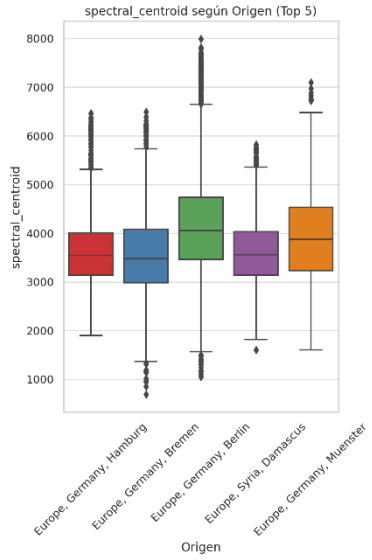


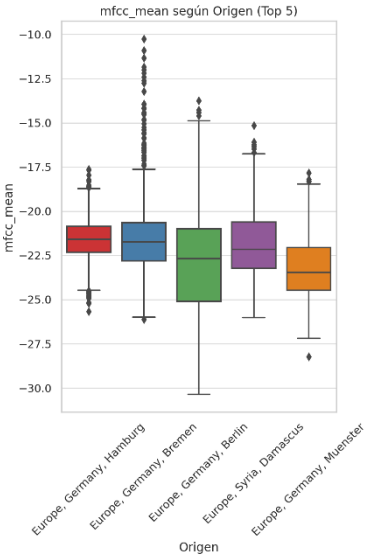




Media MFCC - Origen







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.

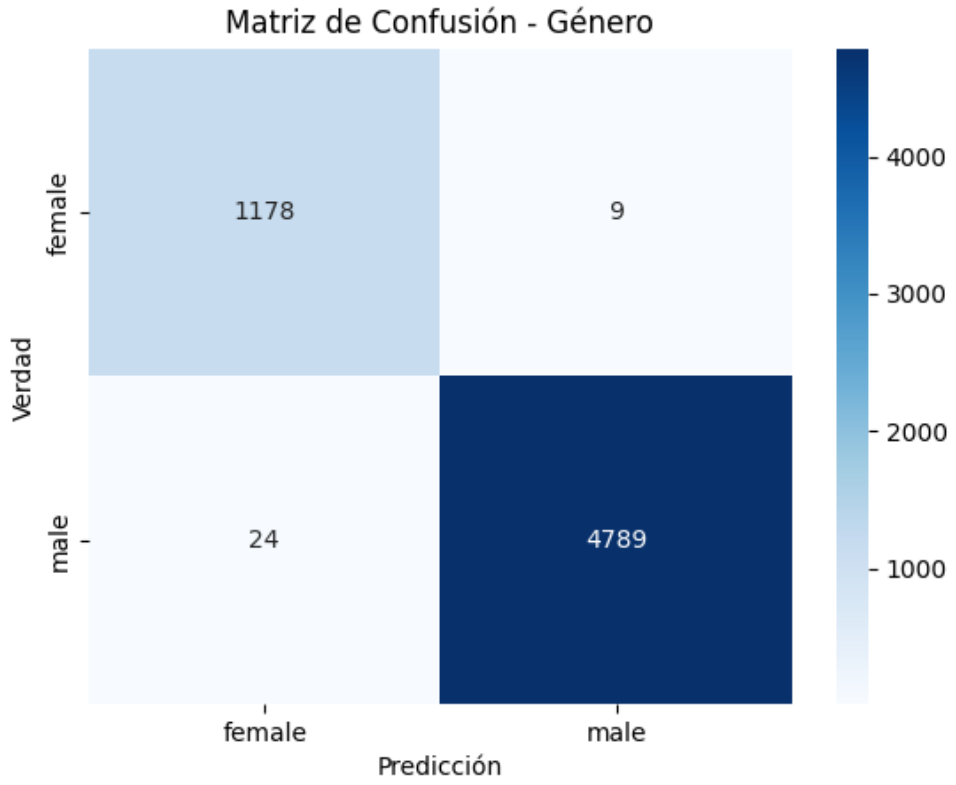
DNN

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

Resultados

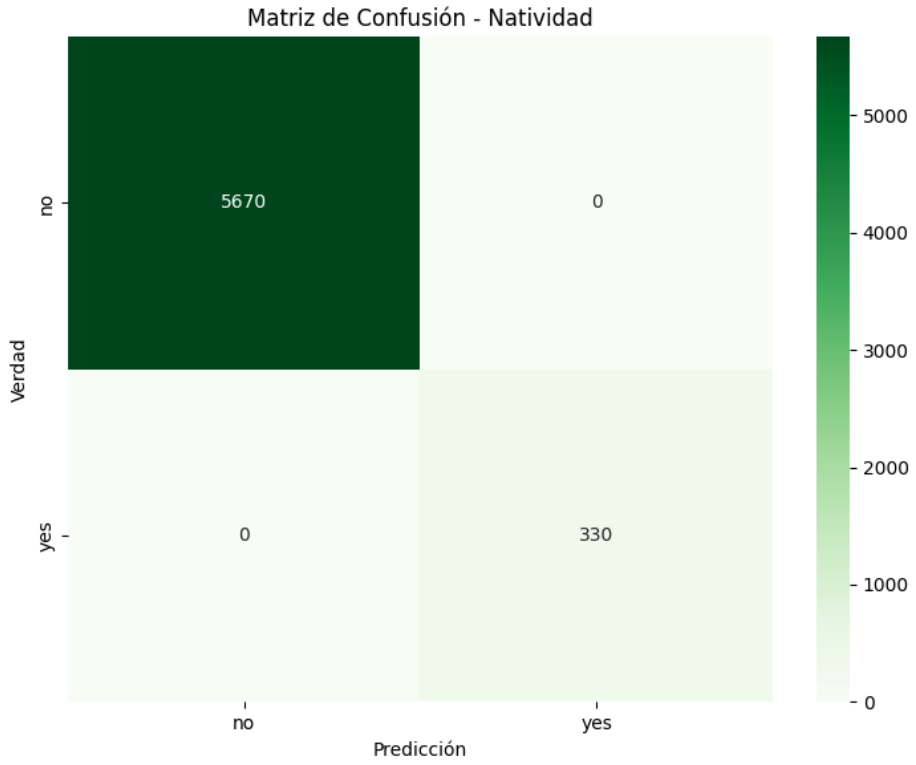
Decision Tree

Género




```
PS C:\Users\UMPO\Documents\Workspace\clasificacion_por_voz> python .\decision_tree_genero.py
Cargando datos...
Realizando validación cruzada...
Resultados de validación cruzada:
Accuracy promedio (validación): 0.99
F1-score promedio (validación): 0.99
Entrenando el modelo final...
Accuracy (prueba): 0.99

Reporte de clasificación:
              precision    recall  f1-score   support

      female       0.98      0.99      0.99      1187
        male       1.00      1.00      1.00      4813

    accuracy                           0.99      6000
   macro avg       0.99      0.99      0.99      6000
weighted avg       0.99      0.99      0.99      6000
```

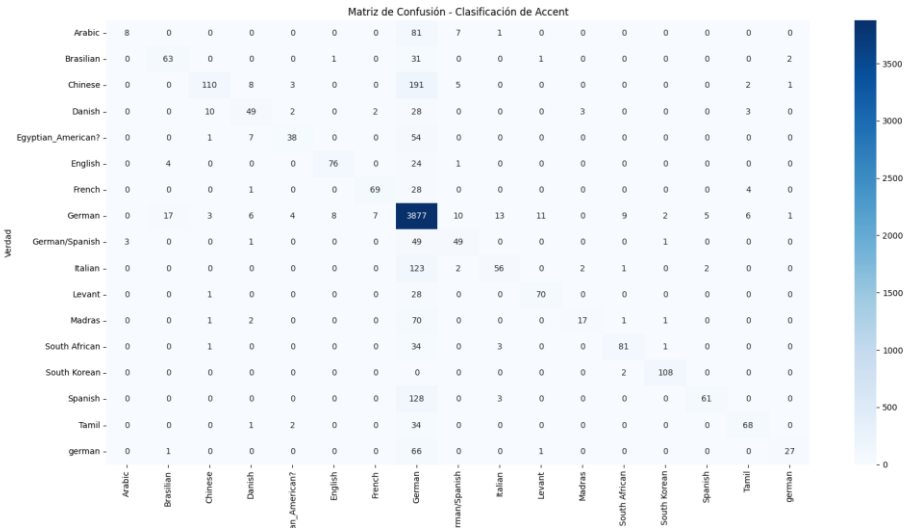
Nativo/No nativo




```
PS C:\Users\UMPO\Documents\Workspace\clasificacion_por_voz> python .\decision_tree_nativo.py
Cargando datos...
Entrenando modelo para clasificación de natividad...
Accuracy: 1.00

Reporte de clasificación:
              precision    recall  f1-score   support

          no       1.00      1.00      1.00      5670
         yes       1.00      1.00      1.00       330

    accuracy                           1.00      6000
   macro avg       1.00      1.00      1.00      6000
weighted avg       1.00      1.00      1.00      6000

PS C:\Users\UMPO\Documents\Workspace\clasificacion_por_voz>
```

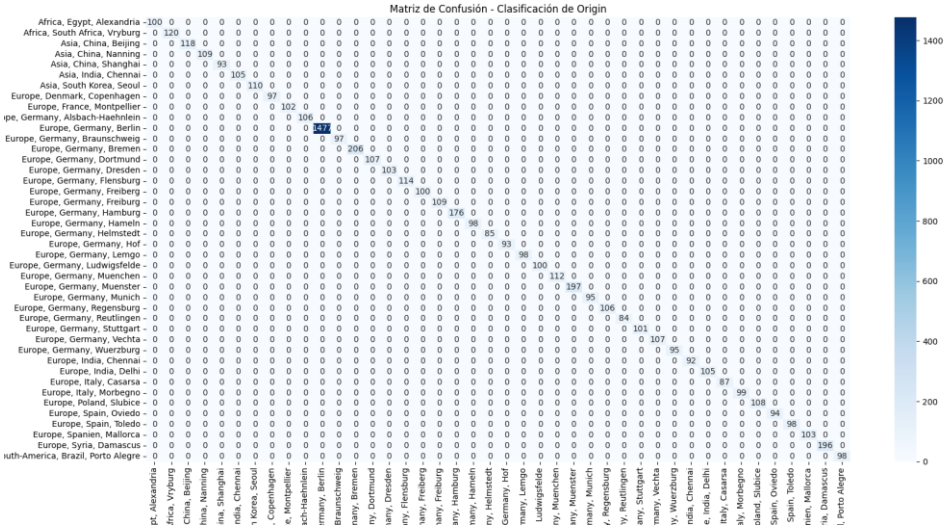
Máquina de Soporte Vectorial

Acento




```
PS C:\Users\UMPO\Documents\Workspace\clasificacion_por_voz> python .\svm_accent.py
Cargando datos...
Mejores hiperparámetros: {'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}
Accuracy promedio (validación cruzada): 0.79
F1-score promedio (validación cruzada): 0.75

Evaluación en conjunto de prueba:
Accuracy: 0.80

Reporte de clasificación:
                    precision    recall  f1-score   support

            Arabic       0.73      0.08      0.15        97
          Brasilian      0.74      0.64      0.69        98
            Chinese      0.87      0.34      0.49       320
             Danish      0.65      0.51      0.57        97
Egyptian_American?       0.78      0.38      0.51       100
            English      0.89      0.72      0.80       105
             French      0.88      0.68      0.77       102
             German      0.80      0.97      0.88      3979
     German/Spanish      0.66      0.48      0.55       103
            Italian      0.74      0.30      0.43       186
             Levant      0.84      0.71      0.77        99
             Madras      0.77      0.18      0.30        92
      South African      0.86      0.68      0.76       120
       South Korean      0.96      0.98      0.97       110
            Spanish      0.90      0.32      0.47       192
              Tamil      0.82      0.65      0.72       105
             german      0.87      0.28      0.43        95

           accuracy                           0.80      6000
          macro avg       0.81      0.52      0.60      6000
       weighted avg       0.81      0.80      0.77      6000
```

Origen



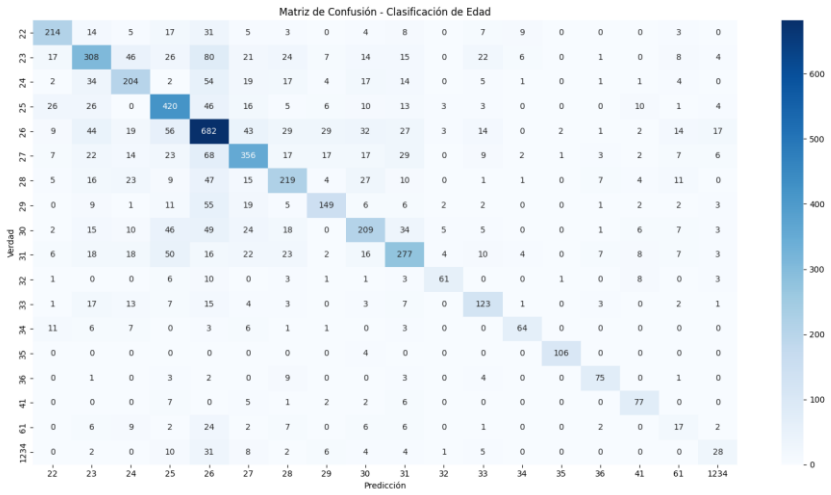

```
PS C:\Users\UMPO\Documents\Workspace\clasificacion_por_voz> python .\svm_origin.py
Cargando datos...
Mejores hiperparámetros: {'C': 0.1, 'gamma': 'scale', 'kernel': 'rbf'}
Accuracy promedio (validación cruzada): 1.00
F1-score promedio (validación cruzada): 1.00

Evaluación en conjunto de prueba:
Accuracy: 1.00
```

Reporte de clasificación:

	precision	recall	f1-score	support
Africa, Egypt, Alexandria	1.00	1.00	1.00	100
Africa, South Africa, Vryburg	1.00	1.00	1.00	120
Asia, China, Beijing	1.00	1.00	1.00	118
Asia, China, Nanning	1.00	1.00	1.00	109
Asia, China, Shanghai	1.00	1.00	1.00	93
Asia, India, Chennai	1.00	1.00	1.00	105
Asia, South Korea, Seoul	1.00	1.00	1.00	110
Europe, Denmark, Copenhagen	1.00	1.00	1.00	97
Europe, France, Montpellier	1.00	1.00	1.00	102
Europe, Germany, Alsbach-Haehnlein	1.00	1.00	1.00	106
Europe, Germany, Berlin	1.00	1.00	1.00	1477
Europe, Germany, Braunschweig	1.00	1.00	1.00	97
Europe, Germany, Bremen	1.00	1.00	1.00	206
Europe, Germany, Dortmund	1.00	1.00	1.00	107
Europe, Germany, Dresden	1.00	1.00	1.00	103
Europe, Germany, Flensburg	1.00	1.00	1.00	114
Europe, Germany, Freiberg	1.00	1.00	1.00	100
Europe, Germany, Freiburg	1.00	1.00	1.00	109
Europe, Germany, Hamburg	1.00	1.00	1.00	176
Europe, Germany, Hameln	1.00	1.00	1.00	98
Europe, Germany, Helmstedt	1.00	1.00	1.00	85
Europe, Germany, Hof	1.00	1.00	1.00	93
Europe, Germany, Lemgo	1.00	1.00	1.00	98
Europe, Germany, Ludwigsfelde	1.00	1.00	1.00	100
Europe, Germany, Muenchen	1.00	1.00	1.00	112
Europe, Germany, Muenster	1.00	1.00	1.00	197
Europe, Germany, Munich	1.00	1.00	1.00	95
Europe, Germany, Regensburg	1.00	1.00	1.00	106
Europe, Germany, Reutlingen	1.00	1.00	1.00	84
Europe, Germany, Stuttgart	1.00	1.00	1.00	101
Europe, Germany, Vechta	1.00	1.00	1.00	107
Europe, Germany, Wuerzburg	1.00	1.00	1.00	95
Europe, India, Chennai	1.00	1.00	1.00	92
Europe, India, Delhi	1.00	1.00	1.00	105
Europe, Italy, Casarsa	1.00	1.00	1.00	87
Europe, Italy, Morbegno	1.00	1.00	1.00	99
Europe, Poland, Slubice	1.00	1.00	1.00	108
Europe, Spain, Oviedo	1.00	1.00	1.00	94
Europe, Spain, Toledo	1.00	1.00	1.00	98
Europe, Spanien, Mallorca	1.00	1.00	1.00	103
Europe, Syria, Damascus	1.00	1.00	1.00	196
South-America, Brazil, Porto Alegre	1.00	1.00	1.00	98
accuracy			1.00	6000
macro avg	1.00	1.00	1.00	6000
weighted avg	1.00	1.00	1.00	6000

KNN

Edad

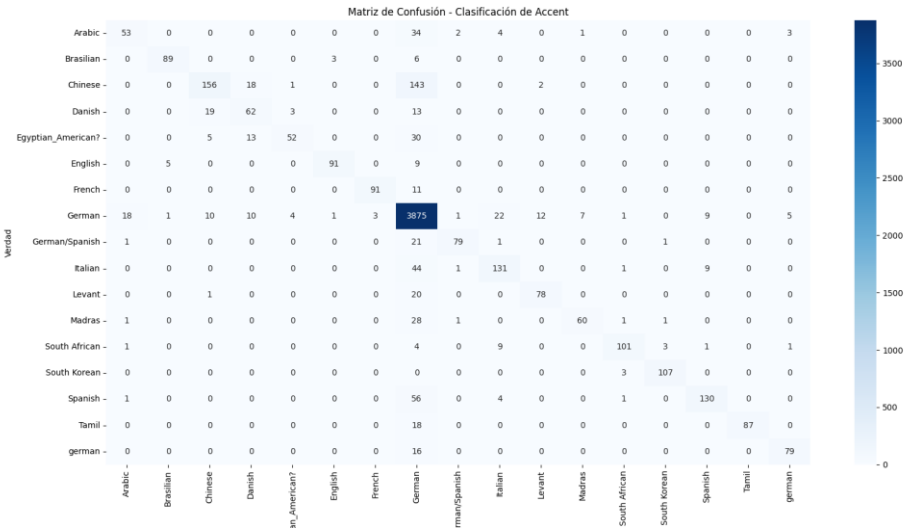



```
PS C:\Users\UMPO\Documents\Workspace\clasificacion_por_voz> python .\knn_edad.py
Cargando datos...
Mejores hiperparámetros: {'metric': 'manhattan', 'n_neighbors': 9, 'weights': 'distance'}
Accuracy promedio (validación cruzada): 0.58

Evaluación en conjunto de prueba:
Accuracy: 0.60
```

```
Reporte de clasificación:
              precision    recall  f1-score   support

          22       0.71      0.67      0.69       320
          23       0.57      0.51      0.54       599
          24       0.55      0.54      0.55       379
          25       0.60      0.71      0.65       589
          26       0.56      0.67      0.61      1023
          27       0.63      0.59      0.61       600
          28       0.57      0.55      0.56       399
          29       0.65      0.55      0.59       273
          30       0.56      0.48      0.52       434
          31       0.60      0.56      0.58       491
          32       0.77      0.62      0.69        98
          33       0.58      0.61      0.60       200
          34       0.73      0.63      0.67       102
          35       0.96      0.96      0.96       110
          36       0.74      0.77      0.75        98
          41       0.64      0.77      0.70       100
          61       0.20      0.20      0.20        84
        1234       0.38      0.28      0.32       101

    accuracy                           0.60      6000
   macro avg       0.61      0.59      0.60      6000
weighted avg       0.60      0.60      0.60      6000
```

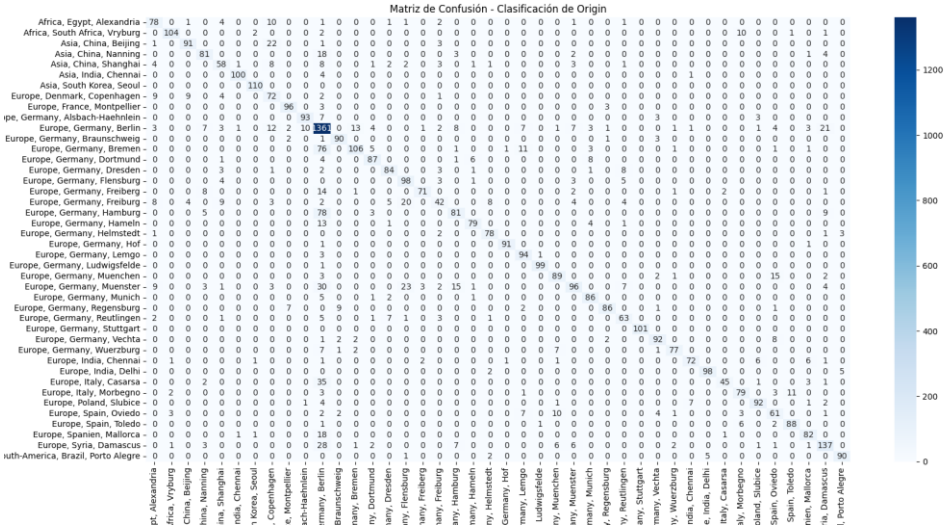
Redes neuronales

Acento




```
Pérdida en prueba: 0.3252
Precisión en prueba: 0.8868
188/188 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step

Reporte de Clasificación:
                    precision    recall  f1-score   support

            Arabic       0.71      0.55      0.62        97
         Brasilian       0.94      0.91      0.92        98
           Chinese       0.82      0.49      0.61       320
            Danish       0.60      0.64      0.62        97
Egyptian_American?       0.87      0.52      0.65       100
           English       0.96      0.87      0.91       105
            French       0.97      0.89      0.93       102
            German       0.90      0.97      0.93      3979
    German/Spanish       0.94      0.77      0.84       103
           Italian       0.77      0.70      0.73       186
            Levant       0.85      0.79      0.82        99
            Madras       0.88      0.65      0.75        92
      South African       0.94      0.84      0.89       120
       South Korean       0.96      0.97      0.96       110
            Spanish       0.87      0.68      0.76       192
              Tamil       1.00      0.83      0.91       105
            german       0.90      0.83      0.86        95

          accuracy                           0.89      6000
         macro avg       0.87      0.76      0.81      6000
      weighted avg       0.88      0.89      0.88      6000

PS C:\Users\UMPO\Documents\Workspace\clasificacion_por_voz>
```

Origen




```
Pérdida en prueba: 0.5795
Precisión en prueba: 0.8130
188/188 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step
```

Reporte de Clasificación:

	precision	recall	f1-score	support
Africa, Egypt, Alexandria	0.68	0.78	0.73	100
Africa, South Africa, Vryburg	0.94	0.87	0.90	120
Asia, China, Beijing	0.87	0.77	0.82	118
Asia, China, Nanning	0.74	0.74	0.74	109
Asia, China, Shanghai	0.66	0.62	0.64	93
Asia, India, Chennai	0.97	0.95	0.96	105
Asia, South Korea, Seoul	0.96	1.00	0.98	110
Europe, Denmark, Copenhagen	0.55	0.74	0.63	97
Europe, France, Montpellier	0.90	0.94	0.92	102
Europe, Germany, Alsbach-Haehnlein	0.89	0.88	0.89	106
Europe, Germany, Berlin	0.78	0.92	0.84	1477
Europe, Germany, Braunschweig	0.87	0.93	0.90	97
Europe, Germany, Bremen	0.85	0.51	0.64	206
Europe, Germany, Dortmund	0.84	0.81	0.82	107
Europe, Germany, Dresden	0.82	0.82	0.82	103
Europe, Germany, Flensburg	0.67	0.86	0.75	114
Europe, Germany, Freiberg	0.92	0.71	0.80	100
Europe, Germany, Freiburg	0.64	0.39	0.48	109
Europe, Germany, Hamburg	0.70	0.46	0.55	176
Europe, Germany, Hameln	0.88	0.81	0.84	98
Europe, Germany, Helmstedt	0.85	0.92	0.88	85
Europe, Germany, Hof	0.98	0.98	0.98	93
Europe, Germany, Lemgo	0.77	0.96	0.85	98
Europe, Germany, Ludwigsfelde	0.98	0.99	0.99	100
Europe, Germany, Muenchen	0.77	0.79	0.78	112
Europe, Germany, Muenster	0.77	0.49	0.60	197
Europe, Germany, Munich	0.82	0.91	0.86	95
Europe, Germany, Regensburg	0.92	0.81	0.86	106
Europe, Germany, Reutlingen	0.70	0.75	0.72	84
Europe, Germany, Stuttgart	1.00	1.00	1.00	101
Europe, Germany, Vechta	0.87	0.86	0.86	107
Europe, Germany, Wuerzburg	0.92	0.81	0.86	95
Europe, India, Chennai	0.89	0.78	0.83	92
Europe, India, Delhi	0.95	0.93	0.94	105
Europe, Italy, Casarsa	0.94	0.52	0.67	87
Europe, Italy, Morbegno	0.81	0.80	0.80	99
Europe, Poland, Slubice	0.88	0.85	0.87	108
Europe, Spain, Oviedo	0.64	0.65	0.64	94
Europe, Spain, Toledo	0.88	0.90	0.89	98
Europe, Spanien, Mallorca	0.83	0.80	0.81	103
Europe, Syria, Damascus	0.74	0.70	0.72	196
South-America, Brazil, Porto Alegre	0.92	0.92	0.92	98
accuracy			0.81	6000
macro avg	0.83	0.80	0.81	6000
weighted avg	0.82	0.81	0.81	6000