Universidad de Buenos Aires

Maestría en explotación de datos y descubrimiento del conocimiento

Aprendizaje Automático - 1° Cuatrimestre 2021

Trabajo Práctico N°2

Predicción de emociones en Audios

Integrantes:

- Adrian Marino
- Alejandro Szpak
- Claudio Collado

https://github.com/magistery-tps/aa-tp2
https://github.com/magistery-tps/aa-tp2/blob/master/notebooks/resolucion-tp-2.ipynb

1. Resumen

En la actualidad el incremento exponencial de la cantidad de datos generados y disponibles no es ajeno a los datos de tipo audio, viéndose una gran cantidad de aplicaciones de algoritmos de aprendizaje automático que utilizan como datos de entrada este tipo de formato.
Teniendo en cuenta lo anterior el presente Trabajo Práctico utiliza una base de datos de audios hablados y cantados sobre los cuales se extraen un conjunto de features predefinidos para luego utilizar diferentes algoritmos de ensambles, junto con dos estrategias de cross-validation , siendo el objetivo final la obtención de un modelo herramienta que permita predecir las emociones presentes en nuevos audios.

2. Introducción

El presente trabajo práctico tiene como objetivo principal la aplicación de conceptos y metodologías vistos en la segunda parte de la materia *Aprendizaje Automático* perteneciente a la *Maestría en explotación de datos y descubrimiento de conocimiento.*
Para su elaboración se parte de un conjunto de datos suministrados sobre los cuales se aplican algoritmos de ensambles de forma tal de obtener una herramienta de clasificación de la emoción presente en audios. A continuación se describe brevemente el contenido de las próximas secciones:

- En la **Sección 3** se describen las principales características y particularidades del conjunto de datos suministrados.
- En la **Sección 4** se describen las metodologías utilizadas sobre los datos y algoritmos.
- En la **Sección 5** se exponen los resultados obtenidos y el análisis correspondiente.
- En la **Sección 6** se enumeran las conclusiones.
- En la **Sección 7** se enumera la bibliografía general y lecturas utilizadas.
- En la **Sección 8** se presentan los Anexos complementarios al cuerpo principal.

3. Datos

Para el desarrollo del trabajo se utilizará el conjunto de datos *Ryerson Audiovisual Database of Emotional Speech and Song (RAVDESS).* Este conjunto de datos presenta las siguientes características principales:

- Contiene 24 actores profesionales (12 mujeres, 12 hombres) vocalizando dos declaraciones léxicamente emparejadas con un acento norteamericano neutral.
- El audio hablado incluye expresiones de calma, felicidad, tristeza, enojo, miedo, sorpresa y disgusto, mientras que el audio cantado contiene emociones tranquilas, felices, tristes, enojadas y temerosas.
- Cada expresión se produce en dos niveles de intensidad emocional (normal, fuerte) con una expresión neutra adicional.
- Todas las condiciones están disponibles en tres formatos de modalidad: Solo audio / Audio y video / Solo video (sin sonido).
- No hay archivos de audio cantado para el Actor_18.
- Por defecto el nombre de cada archivo contiene información predeterminada: Modalidad, canal vocal, emoción (target), intensidad emocional, declaración, número de repetición y el actor que corresponde. En este caso se analizó qué features era necesario utilizar para dar cumplimiento al contenido del TP y cuales no era posible utilizar por ser considerados que fueron generados en forma posterior a la grabación del audio (post-etiquetado) y que en audios nuevos a ingresar al modelo final no estarían presentes para su uso.

Para este trabajo se utilizó el conjunto de datos con formato solo Audio, el cual contiene:

- Archivos de audio hablado: 60 ensayos para 24 actores, totalizando 1440 instancias
- Archivo de audio cantado: 44 ensayos para 23 actores, totalizando 1012 instancias

4. Metodología

A lo largo del desarrollo del trabajo se realizaron en forma secuencial las siguientes actividades:

a) Obtención del dataset y preprocesamiento:
   i) A partir del enlace provisto se obtuvieron los archivos correspondiente a solo audio en las modalidades hablado y cantado. Luego se extrajeron el conjunto de atributos GeMAPS (Geneva Minimalistic Acoustic Parameter Set) utilizando la librería opensmile (python). El dataset en esta instancia contiene 100 columnas (features)
   ii) De los features obtenidos como parte del nombre del audio (modality, vocal channel, emotion,emotional intensity, statement,repetition y actor) nos quedamos con los siguientes:
      - vocal_chanel: Para poder separar los sets en speechs (hablado) y songs (cantado)
      - Emotion: Corresponde al Target
      - Actor: Utilizado para la estrategia de validación Leave-2-speakers out

iii) Se observa que no existen valores faltantes (NaN)

b) División del dataset: Realizado lo anterior se dispone del dataset completo en condiciones necesarias para avanzar con la aplicación de los modelos. Se realiza la división de este dataset en:
   i) Desarrollo (development) - 80%
   ii) Test - 20%.

   Se verificó que la proporción de las clases siga siendo la misma en ambos conjuntos.

c) Estrategias de cross-validation:
   i) 12-fold cross validation armando los folds de forma aleatoria.
   ii) Leave-2-speakers out: 12 folds conteniendo cada uno 2 actores distintos.

   En ambos casos se utilizó como métrica el accuracy y se probaron los siguientes modelos sin búsqueda de hiperparametros: Random Forest, AdaBoost, Gradient Boosting y XGBoost.

d) Para la mejor estrategia de división de datos de las analizadas anteriormente se comparó el desempeño de Random Forest, Ada Boost, Gradient Boosting y XGBoost.
   Como métrica de performance se decidió utilizar el accuracy: Esta elección se sustenta al verificar el balance de las diferentes clases (emociones a clasificar) ya que no se observaron grandes diferencias tanto a nivel general como para los grupos de sonidos cantados y hablados. Sumado a lo anterior y como siempre es recomendado se utilizaron a modo de complemento otro conjunto de métricas: Precisión, Recall y F1-Score, sin observarse en estos casos diferencias significativas que pudieran alertar una incorrecta selección de Accuracy como métrica principal.

e) Se realizó el análisis de la robustez de los modelos incorporando diferentes relaciones señal-ruido de ruido sintético (uniforme) a los audios disponibles.

f) Se probó con una Red Neuronal (Perceptrón Multicapa) y se compararon los resultados contra el mejor modelo obtenido en c)

5. Resultados

5.1 Estrategias de división de datos

5.1.1 12 folds cross-validation armando los folds de forma aleatoria.

En la Tabla N°1 se observan los valores de Accuracy obtenidos para la estrategia de 12-fold cross-validation armando los folds de forma aleatoria, para los 4 modelos considerados:

	Random Forest	AdaBoost	Gradient Boosting	XGBoost
Accuracy	65.98%	37.88%	63.85%	63.13%

Tabla N°1

Se observa que en este caso el mejor score se obtiene con Random Forest seguido por Gradient Boosting y XGBoost. Lejos de estos valores se ubican AdaBoost.

5.1.2 Leave-2-speakers out (LeaveOneGroupOut)

En la Tabla N°2 se observan los valores de Accuracy obtenidos para la estrategia Leave-2-speakers out, donde se tienen 12 folds conteniendo cada uno 2 actores distintos. Igual que en el caso anterior se observan los resultados para los 4 modelos considerados:

	Random Forest	AdaBoost	Gradient Boosting	XGBoost
Accuracy	52.71%	34.79%	51.77%	52.24%

Tabla N°2

Se observa nuevamente que el mejor score se obtiene con Random Forest seguido por XGBoost y Gradient Boosting . Con respecto a la estrategia anterior de 12 fold cross-validation armando los folds de forma aleatoria los Accuracy obtenidos en este caso para todos los modelos son menores.

Analizando los resultados obtenidos con los dos tipos de split mostrados anteriormente se obtiene una performance superior en todos los modelos utilizando 12 folds cross-validation (folds aleatorios) por sobre la estrategia Leave-2-speakers out. Sin embargo, al analizar de qué manera separan los datos ambas estrategias, consideramos que quizás la mejor performance de 12 folds cross-validation (folds aleatorios) puede estar dada por contener datos del mismo actor en los folds de train y test. Esto en principio no es lo ideal, dado que, inicialmente las características de voz nos parecen muy específicas de cada persona, por lo cual si intentamos predecir con datos de la misma persona con las que entrenamos, es esperable que tenga una buena performance.

En este punto nos pareció un factor importante analizar qué características tendría un dataset productivo, ya que estas características pueden determinar el éxito o fracaso del modelo. Para considerar esto, debemos generar datasets held-out (e idealmente de validación) lo más similares a los que el modelo se encontrará en entornos productivos y así poder probar que tan bien generaliza el modelo. En este caso de ejemplo, el modelo se entrena con 24 actores con la idea de generalizar hacia audios de otras personas no vistas, es por ello que si incluimos al mismo actor en train y test estaríamos alejándonos de lo que consideramos una realidad productiva del modelo. Teniendo esto en consideración y modo de análisis adicional probamos para un mismo algoritmo (Random Forest) las siguientes alternativas de split:

- Splits Random
- Splits Manuales

Finalmente se realizaron predicciones sobre un conjunto held-out (test) con actores no vistos en entrenamiento, de forma tal de tener valores de la métrica (accuracy) de comparación más realista. En la Tabla N°3 se observan los resultados obtenidos, los cuales permiten corroborar nuestro supuesto inicial: El Split Random tiene algo de overfitting ya que es más sensible a cambios cuando predecimos actores no vistos en entrenamiento. A partir de esto decidimos continuar el análisis posterior con la estrategia LeaveOneGroupOut, por ser esta más realista.

	Splits Random	**Splits Manuales**
Accuracy - Validation	43.11%	32.53%
Accuracy - Test	29.72%	38.50%

Tabla N°3

5.2 División de datos considerada más adecuada - Búsqueda de Hiperparametros

Considerando la división de datos elegida anteriormente (LeaveOneGroupOut) se realiza la búsqueda de hiperparametros para los 4 modelos y se compara el desempeño considerando la métrica accuracy. En la Tabla N°4 se observan los resultados obtenidos:

	Random Forest	**AdaBoost**	**Gradient Boosting**	**XGBoost**
Accuracy	38.69%	39.30%	54.78%	68.02%

Tabla N°4

Se observa que el mejor modelo corresponde a XGBoost. En Anexo 8.1 se observan las principales métricas y matrices de confusión para los casos analizados

5.2.1 Mejor Modelo - Audios hablados y cantados

Para el mejor modelo obtenido en 5.2 (XGBoost) se realiza la evaluación del mismo sobre la población de audios hablados y cantados. En la Tabla N°5 se observan los resultados obtenidos:

	Audios Hablados	**Audios Cantados**
Accuracy Mejor Modelo - XGBoost	63.19%	74.87%

Tabla N°5

En Anexo 8.2 se observan el resto de las métricas y matrices de confusión para los casos analizados

5.2.2 Mejor Modelo - Audios femeninos y masculinos

Para el mejor modelo obtenido en 5.2 (XGBoost) se realiza la evaluación del mismo sobre la población de audios femeninos y masculinos. En la Tabla N°6 se observan los resultados:

	Audios Femeninos	**Audios Masculinos**
Accuracy Mejor Modelo - XGBoost	68.33%	67.72%

Tabla N°6

En Anexo 8.3 se observan el resto de las métricas y matrices de confusión para los casos analizados

5.3 Robustez del modelo - Presencia de Ruido

Se analiza la robustez del modelo frente a perturbación en los audios debido a la inclusión de ruido artificial del tipo uniforme. El agregado del ruido se especifica por medio de la relación señal-ruido (SNR): Se utilizaron valores discretos de SNR desde valores muy pequeños (1e-300) en forma incremental llegando a valores de SNR igual a 16.
En Anexo 8.5 se observa los resultados obtenidos: En todos los casos se observa una reducción significativa del accuracy por inclusión de ruido.

5.4 Red Neuronal

Como análisis adicional se probó con una Red Neuronal clásica (fully connected) y no se observaron mejoras con respecto al mejor modelo de ensambles (XGBoost) obtenido en 5.2. En la Tabla N°7 se observan los resultados obtenidos:

	Red Neuronal	**XGBoost - Mejor Modelo del Ítem 5.2**
Accuracy	59.26%	68.02%

Tabla N°7

En Anexo 8.4 se observan el resto de las métricas y matriz de confusión para el caso analizado.

6. Conclusiones

- En este tipo de problema (y en general en los métodos supervisados) debemos considerar las características de los datos que queremos predecir, a la hora de armar nuestro modelo final, desde el split de los datos hasta la elección del algoritmo, métrica, hiperparametros y entrenamiento. En el caso de este TP nos parece algo crítico, la elección de un método de split que permita al modelo ser lo más generalista posible, para cumplir con el objetivo de generalizar un modelo a partir de voces de 24 actores, hacia otras personas. Si nuestro objetivo fuese predecir otros audios de esos mismos 24 actores, nuestra estrategia cambiaría radicalmente; por eso, tener bien en claro el objetivo es primordial para tomar decisiones en el modelado.
- Es importante planificar con qué variables vamos a contar en producción para que el modelo funcione acorde a lo esperado a la hora de predecir. En este ejemplo, contamos con datos de las variables incluidas en el archivo (el target y además emotion intensity, entre otras). Incluso algunos de estos datos correlacionan con el target, pero en este sentido es importante ver dónde y en qué momento se generaron y si contaremos con dichos datos en producción. De lo contrario, estaremos entrenando un modelo con variables que al momento de ponerlo en producción no vamos a tener, incurriendo en fallas y/o bajas fuertes de performance (por ejemplo si una de esas variables fuese altamente explicativa). Es por ello que, asumiendo que armamos nuestro modelo para recibir como input solo un archivo de audio, decidimos no usar dichas variables a la hora de entrenar.
- Con respecto al ruido, la baja de performance en test al incluirlo, nos hace ver que el modelo no es lo suficientemente robusto como para predecir de manera correcta ante cambios en el contexto en el cual fue grabado el audio. Una estrategia interesante a probar, fuera de alcance de este trabajo, es hacer oversampling con datos sintéticos en training, incluyendo ruido de distintos tipos (de forma aleatoria), con el objetivo de preparar al modelo para predecir con distintos ruidos ambiente, algo más similar a lo que se puede encontrar en entornos productivos. Investigando sobre el tema pudimos ver que esta técnica es muy utilizada en modelos de speech recognition.
- Con respecto a los resultados obtenidos con la Red Neuronal el valor de accuracy inferior al mejor modelo de ensambles (XGBoost) creemos que puede deberse a la cantidad limitada de datos de entrada del modelo, sabiendo que las redes neuronales maximizan la performance cuando se entrenan con muchos datos. Una alternativa fuera de alcance de este trabajo sería aplicar una técnica de “transfer learning” para redes neuronales cuando se cuenta con pocos datos. Dicha técnica consiste en aprovechar otros modelos entrenados con audios para la detección de palabras en sus capas más bajas (fonemas, palabras, etc.), importando la estructura de la red junto con sus matrices de pesos, luego eliminar la capa output, agregar algunas capas ocultas y una capa output con softmax como función de activación.

7. Bibliografía y Lecturas

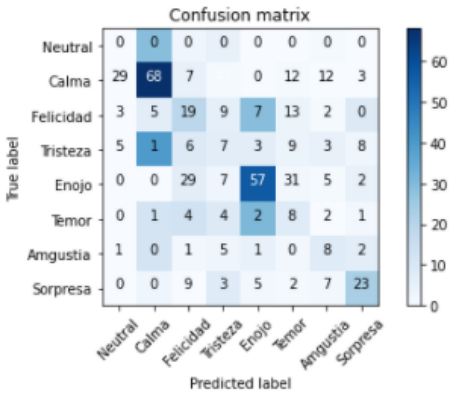
1. ***Machine Learning*** - Tom M. Mitchell (1997)
2. ***Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*** - Aurelien Geron (2017)
3. ***An Introduction to Statistical Learning*** - Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013)
4. ***Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*** - Morgan & Claypool (2010)
5. http://scott.fortmann-roe.com/docs/BiasVariance.html
6. ***The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing***
   https://www.computer.org/csdl/journal/ta/2016/02/07160715/13rRUypp569
7. ***Incorporating Noise Robustness in Speech Command Recognition by Noise Augmentation of Training Data (2020)***
   https://www.mdpi.com/1424-8220/20/8/2326/pdf

8. Anexos

8.1 Métricas y Matrices de confusión - Ítem 5.2

Classification Report:

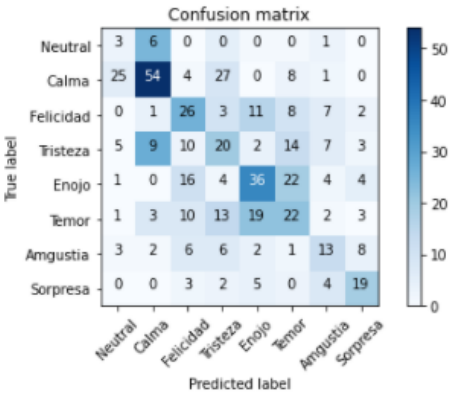
	precision	recall	f1-score	support
Neutral	0.00	0.00	0.00	38
Calma	0.40	0.91	0.55	75
Felicidad	0.33	0.25	0.29	75
Tristeza	0.17	0.09	0.12	75
Enojo	0.44	0.76	0.55	75
Temor	0.36	0.11	0.16	75
Amgustia	0.44	0.21	0.28	39
Sorpresa	0.47	0.59	0.52	39
accuracy			0.39	491
macro avg	0.33	0.36	0.31	491
weighted avg	0.33	0.39	0.32	491



Random Forest

Classification Report:

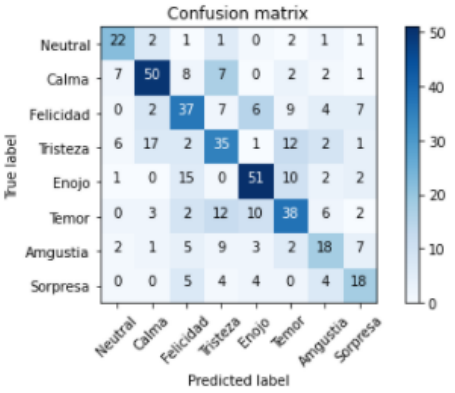
	precision	recall	f1-score	support
Neutral	0.30	0.08	0.12	38
Calma	0.45	0.72	0.56	75
Felicidad	0.45	0.35	0.39	75
Tristeza	0.29	0.27	0.28	75
Enojo	0.41	0.48	0.44	75
Temor	0.30	0.29	0.30	75
Amgustia	0.32	0.33	0.32	39
Sorpresa	0.58	0.49	0.53	39
accuracy			0.39	491
macro avg	0.39	0.38	0.37	491
weighted avg	0.38	0.39	0.38	491



AdaBoost

Classification Report:

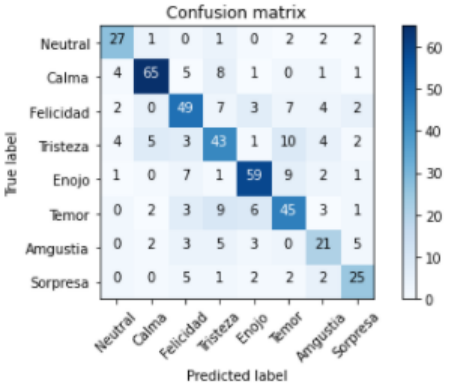
	precision	recall	f1-score	support
Neutral	0.73	0.58	0.65	38
Calma	0.65	0.67	0.66	75
Felicidad	0.51	0.49	0.50	75
Tristeza	0.46	0.47	0.46	75
Enojo	0.63	0.68	0.65	75
Temor	0.52	0.51	0.51	75
Amgustia	0.38	0.46	0.42	39
Sorpresa	0.51	0.46	0.49	39
accuracy			0.55	491
macro avg	0.55	0.54	0.54	491
weighted avg	0.55	0.55	0.55	491



Gradient Boosting

Classification Report:

	precision	recall	f1-score	support
Neutral	0.77	0.71	0.74	38
Calma	0.76	0.87	0.81	75
Felicidad	0.66	0.65	0.66	75
Tristeza	0.60	0.57	0.59	75
Enojo	0.74	0.79	0.76	75
Temor	0.65	0.60	0.63	75
Amgustia	0.54	0.54	0.54	39
Sorpresa	0.68	0.64	0.66	39
accuracy			0.68	491
macro avg	0.67	0.67	0.67	491
weighted avg	0.68	0.68	0.68	491

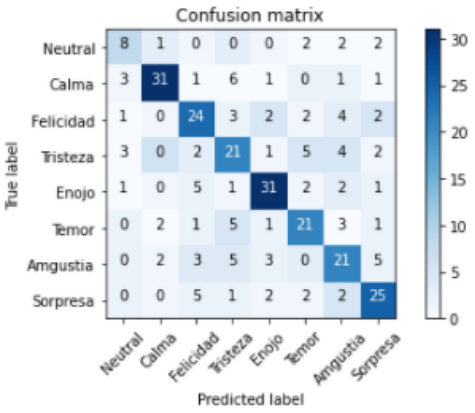


XGBoost

8.2 Métricas y Matrices de confusión - Ítem 5.2.1

Classification Report:

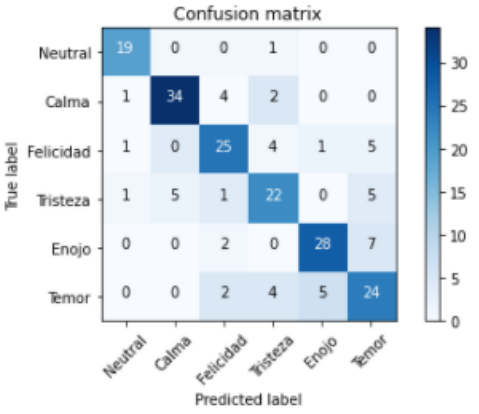
	precision	recall	f1-score	support
Neutral	0.53	0.50	0.52	16
Calma	0.70	0.86	0.78	36
Felicidad	0.63	0.59	0.61	41
Tristeza	0.55	0.50	0.53	42
Enojo	0.72	0.76	0.74	41
Temor	0.62	0.62	0.62	34
Amgustia	0.54	0.54	0.54	39
Sorpresa	0.68	0.64	0.66	39
accuracy			0.63	288
macro avg	0.62	0.62	0.62	288
weighted avg	0.63	0.63	0.63	288



Audios Hablados

Classification Report:

	precision	recall	f1-score	support
Neutral	0.95	0.86	0.90	22
Calma	0.83	0.87	0.85	39
Felicidad	0.69	0.74	0.71	34
Tristeza	0.65	0.67	0.66	33
Enojo	0.76	0.82	0.79	34
Temor	0.69	0.59	0.63	41
accuracy			0.75	203
macro avg	0.76	0.76	0.76	203
weighted avg	0.75	0.75	0.75	203

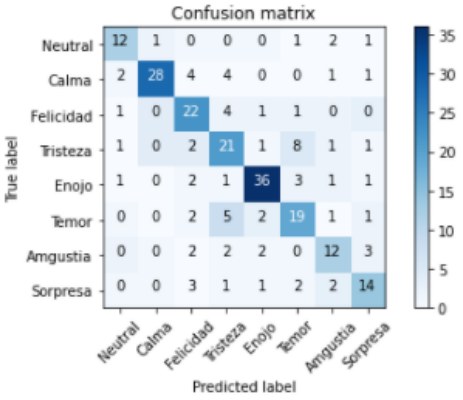


Audios Cantados

8.3 Métricas y Matrices de confusión - Ítem 5.2.2

Classification Report:

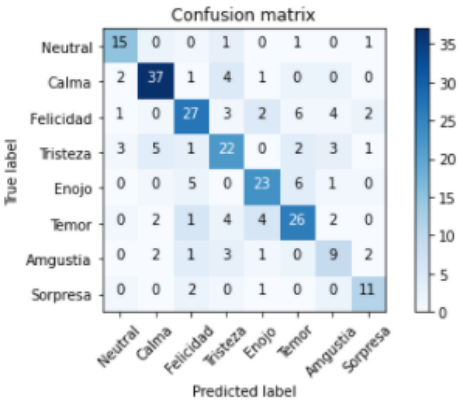
	precision	recall	f1-score	support
Neutral	0.71	0.71	0.71	17
Calma	0.70	0.97	0.81	29
Felicidad	0.76	0.59	0.67	37
Tristeza	0.60	0.55	0.58	38
Enojo	0.80	0.84	0.82	43
Temor	0.63	0.56	0.59	34
Amgustia	0.57	0.60	0.59	20
Sorpresa	0.61	0.64	0.62	22
accuracy			0.68	240
macro avg	0.67	0.68	0.67	240
weighted avg	0.68	0.68	0.68	240



Audios Femeninos

Classification Report:

	precision	recall	f1-score	support
Neutral	0.83	0.71	0.77	21
Calma	0.82	0.80	0.81	46
Felicidad	0.60	0.71	0.65	38
Tristeza	0.59	0.59	0.59	37
Enojo	0.66	0.72	0.69	32
Temor	0.67	0.63	0.65	41
Amgustia	0.50	0.47	0.49	19
Sorpresa	0.79	0.65	0.71	17
accuracy			0.68	251
macro avg	0.68	0.66	0.67	251
weighted avg	0.68	0.68	0.68	251

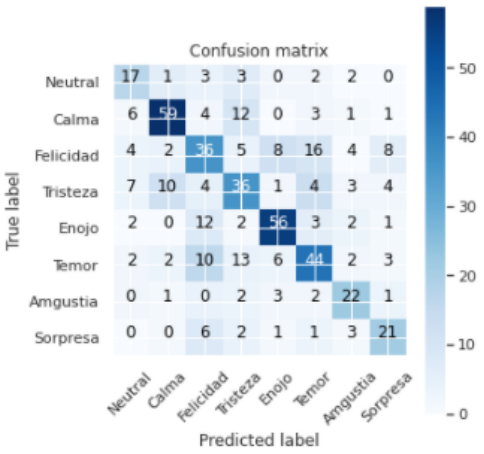


Audios Masculinos

8.4 Métricas y Matrices de confusión - Ítem 5.4

Classification Report:

	precision	recall	f1-score	support
Neutral	0.61	0.45	0.52	38
Calma	0.69	0.79	0.73	75
Felicidad	0.43	0.48	0.46	75
Tristeza	0.52	0.48	0.50	75
Enojo	0.72	0.75	0.73	75
Temor	0.54	0.59	0.56	75
Amgustia	0.71	0.56	0.63	39
Sorpresa	0.62	0.54	0.58	39
accuracy			0.59	491
macro avg	0.60	0.58	0.59	491
weighted avg	0.59	0.59	0.59	491



8.5 Robustez de los modelos frente a la perturbación por ruido - Ítem 5.3