Enfoque Dual para la detección de Emociones críticas en el Habla

DELFINA CHAVEZ BLASI[1] Y JULIETA GARCIA PEREYRA[2]

[1] Universidad de San Andrés, Ingeniería en Inteligencia Artificial , Buenos Aires, Argentina.
dchavezblasi@udesa.edu.ar

[2] Universidad de San Andrés, Ingeniería en Inteligencia Artificial , Buenos Aires, Argentina.
jgarciapereyra@udesa.edu.ar

***Abstract*—**Este estudio presenta una metodología para la predicción de siete emociones distintas a partir de características eGeMAPs extraídas de la voz con el dataset RAVDESS, utilizando modelos de aprendizaje automático. Dada la complejidad inherente del comportamiento humano, es esencial prestar atención a emociones que pueden conllevar riesgos significativos, como la tristeza, el enojo y el miedo, especialmente cuando se manifiestan con alta intensidad. Para abordar esta problemática, se desarrolló un modelo dual: el primer clasificador identifica la emoción, y si esta corresponde a tristeza, enojo o miedo, un segundo modelo determina si la intensidad de la emoción es normal o fuerte. Este enfoque facilita la identificación precisa del tipo de emoción y de su intensidad, aportando significativamente a la seguridad emocional en diversas aplicaciones.

Keywords—análisis del habla, aprendizaje automático, clasificación emocional, intensidad emocional.

I. Introduccion

Las emociones cumplen un rol fundamental en el desarrollo y funcionamiento integral de las personas. En el ámbito del procesamiento de señales de habla, la detección de emociones juega un papel crucial en diversas aplicaciones, desde la asistencia en salud mental hasta la mejora de la interacción humano-computadora. Este trabajo presenta un enfoque innovador para la predicción de siete emociones distintas a partir de características de la voz, utilizando modelos avanzados de aprendizaje automático. Es debido a la complejidad del comportamiento humano que ha de requerirse una atención especial a emociones que pueden representar un riesgo potencial: la tristeza, el enojo y el miedo, cuando se presentan en niveles extremos. Para abordar esta necesidad, desarrollamos un modelo dual a partir del dataset RAVDESS. En primer lugar, se desarrolló un clasificador que predice una emoción a partir del habla . Posteriormente, si la emoción predicha es alguna de las que pueden presentar un riesgo potencial en caso de ser muy intensas, un segundo modelo secuencial evalúa si la emoción tiene una intensidad normal o fuerte. Este enfoque tiene un impacto significativo, pues puede identificar situaciones que requieren intervención externa oportuna cuando una emoción intensa pone en peligro la vida de una persona.

II. Metodologia

A. Extracción y exploración de datos

El dataset RAVDESS incluye, por una parte, 1440 audios de 24 actores profesionales (12 mujeres, 12 hombres), vocalizando dos declaraciones léxicamente coincidentes en un acento neutral norteamericano. El habla incluye expresiones de calma, neutral, felicidad, tristeza, enojo, miedo, sorpresa y disgusto. Cada expresión se produce en dos niveles de intensidad emocional (normal, fuerte) excepto para el caso neutral que tiene una única intensidad. Por otra parte, hay 1012 audios de 23 actores (11 mujeres, 12 hombres) cantando las mismas expresiones anteriores pero sin las emociones de sorpresa y disgusto. En la Figura , se muestran las distribuciones de las emociones en los audios correspondientes a los datos.

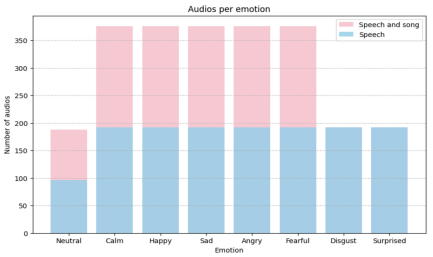

Figura 1. Cantidad de audios por cada emoción para el set de datos RAVDESS.

Como se observa en la Figura 1, no todas las emociones tienen la misma cantidad de audios. Se realizaron experimentos unificando las emociones de calma y neutral en una sola debido a su parecido, y utilizando únicamente los audios de discurso o juntando todos los audios, los de discurso y los de canciones. En cuanto a intensidad, las emociones de interés (miedo, enojo y tristeza) tienen un 50% de los audios etiquetados con intensidad normal y el otro 50% etiquetados con intensidad fuerte.

Es importante destacar que el conjunto de datos RAVDESS, si bien ofrece una variedad significativa de expresiones emocionales actuadas por actores profesionales, presenta ciertas limitaciones en su generalización a contextos fuera del ambiente controlado de la grabación. Esta limitación podría afectar la capacidad del modelo para capturar la diversidad natural de expresiones emocionales en situaciones cotidianas o no actuadas. No obstante, el dataset proporciona un fundamento básico y conceptual para el estudio y desarrollo de sistemas de reconocimiento de emociones en el habla.

A. Extracción de características (features)

Se utilizaron características acústicas extraídas mediante el conjunto de características eGeMAPS (Geneva Minimalistic Acoustic Parameter Set) utilizando el software openSMILE. Este conjunto incluye un total de 88 características acústicas que abarcan una variedad de aspectos del habla y son particularmente útiles para identificar emociones y sus intensidades. Entre los más importantes se destacan:

- *Frecuencia Fundamental (F0):* Indicador de la altura tonal de la voz, relevante para detectar variaciones en el tono que pueden reflejar diferentes estados emocionales.
- *Energía y Amplitud:* Parámetros como la energía de la señal de voz, que pueden indicar la intensidad y el volumen del habla.
- *Espectro de Frecuencia*: Análisis de componentes de frecuencia, como el espectro de Mel y los coeficientes cepstrales, que pueden capturar la calidad del timbre de la voz.
- *Medidas de Calidad de la Voz:* Parámetros como la tasa de vibrato y el jitter, que reflejan la estabilidad y la variación en la voz.

Con el fin de eliminar las características de menor relevancia, realizamos un modelo simple de Random Forest y exploramos la importancia de las features en la clasificación. La Figura 2 muestra los resultados.

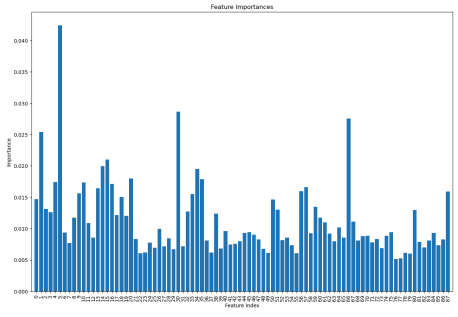

Figura 2. Importancia de cada característica de eGeMAPS en una clasificación con Random Forest. En el eje horizontal se encuentra el número de feature y en el eje vertical su importancia en la clasificación.

Dado que todas las features tienen una importancia significativa, decidimos no eliminar ninguna y utilizamos las 88 para todos los experimentos.

B. División de datos, normalización y validación cruzada

A la hora de evaluar el desempeño de los modelos en la clasificación, resulta de vital importancia dividir los datos para evitar correlaciones espurias y garantizar la capacidad de generalización del modelo a datos no vistos. En este trabajo, se implementaron tres enfoques diferentes de validación cruzada para asegurar una evaluación robusta:

1. *Por Actor:* Se asegura que los datos de entrenamiento y prueba no compartan actores: se selecciona un actor para validación y el modelo se entrena con los datos de todos los demás actores. Esto garantiza que el modelo sea capaz de generalizar a voces no vistas durante el entrenamiento.
2. *Por Declaración:* se entrena el modelo con los datos de una declaración específica y se valida con otra. Así, se evalúa la capacidad del modelo para generalizar a diferentes contenidos del habla.
3. *Por Actor y Declaración*: se elige un actor y una declaración para validación, y el modelo se entrena con los datos de todos los demás actores y declaraciones. Esto crea un escenario donde ni el actor ni la declaración utilizados para la validación han sido vistos durante el entrenamiento, proporcionando una evaluación más rigurosa del modelo.

Adicionalmente, se exploran dos tipos de normalización:

1. *Normalización Global*: Se normalizan las características acústicas globalmente, utilizando la media y la desviación estándar de cada característica (feature) calculadas a partir de todo el conjunto de datos de entrenamiento. Se asegura que todas las características tengan una media de 0 y una desviación estándar de 1, facilitando el proceso de aprendizaje de algunos modelos.
2. *Normalización por Actor:* En este enfoque, se normalizan las características de manera individual para cada actor. Es decir, se calcula la media y la desviación estándar de las características para cada actor y luego se normalizan los datos de cada actor por separado. Esto permite evaluar si la normalización específica por actor mejora la precisión del modelo al considerar las variaciones individuales en la voz de cada actor.

Es importante subrayar que la normalización por actor se empleó con fines experimentales y no es viable en producción porque no se tendría acceso previo a los datos de un usuario desconocido para calcular sus parámetros de normalización.. Este método fue utilizado para explorar el impacto de las variaciones individuales de los actores en la clasificación de emociones. El objetivo era determinar si esta normalización podría simplificar la tarea del modelo al reducir la variabilidad interindividual en las características acústicas.

D. Modelos Implementados

En la primera fase del proyecto, cuyo objetivo era desarrollar modelos para la predicción de emociones, se implementaron varios algoritmos de aprendizaje automático, incluyendo una Red Neuronal densamente conectada de Múltiples Capas (MLP) , un Random Forest y un Gradient Boosting. Para la segunda fase, enfocada en la detección de la intensidad emocional, se desarrollaron modelos adicionales como otra MLP, Gradient Boosting y una Regresión Logística. Para evaluar el desempeño de cada modelo, se utilizaron las siguientes métricas sobre los grupos de validación: la matriz de confusión y la accuracy. La elección de estas métricas se basa en sus capacidades para proporcionar una evaluación integral del modelo. En la primera fase de predicción de emociones, la accuracy permite obtener una visión general de la efectividad del modelo. Asimismo, la matriz de confusión proporciona una visualización detallada que garantiza que nuestras emociones de interés en la segunda fase (tristeza, enojo y miedo) sean predichas con un grado aceptable de certeza. Luego de analizar los resultados obtenidos por cada modelo, optamos por un modelo final compuesto por un Gradient Boosting y una Regresión Logística. La salida del modelo de Gradient Boosting se utiliza como entrada para la

Regresión Logística. A este enfoque lo denominamos el 'Modelo Dual' y se explica en la sección III C.

II. Resultados y Discusión

A. Predicción de Emociones

Comenzamos analizando las posibles correlaciones espurias presentes en el dataset. Para simplificar y reducir el costo computacional, se realizó el análisis utilizando únicamente audios de 'speech'. El dataset fue inicialmente dividido en un conjunto de entrenamiento y otro de prueba, asegurando que ningún actor estuviera presente en ambos conjuntos. Esto resultó en un conjunto de entrenamiento con 22 actores y un conjunto de prueba con 2 actores. Posteriormente, se utilizó el conjunto de entrenamiento para realizar validación cruzada utilizando los 3 métodos de K-folds explicados en la sección *2C*. Se evaluaron las métricas de todos los modelos bajo estos tres tipos de validación cruzada, observando comportamientos similares en todos ellos.

Es relevante mencionar que el dataset original contiene ocho emociones: calma, neutral, felicidad, tristeza, enojo, miedo, sorpresa y disgusto. Al analizar la matriz de confusión generada por el modelo Random Forest, notamos que el modelo solía confundir 'Neutral' y 'Calma' reiteradas veces. Esta confusión se debe a la similitud entre los audios asociados a estas emociones. Por consiguiente, se optó por unificarlas en una sola categoría. Todos los resultados presentados en este trabajo se basan en el dataset con estas emociones unificadas. A continuación, analizaremos estos resultados centrando la atención en el modelo de Random Forest (las emociones se balancean tomando 'balanced_subsamples'. No se normalizan los folds).

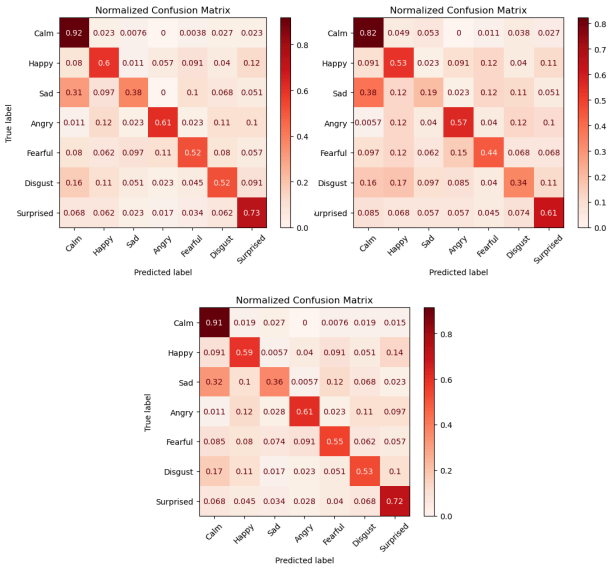

Figura 3. Matrices de Confusión del Random Forest. Arriba a la izquierda, corresponde al cross-validation realizado por declaración, con un accuracy de 0.6303. Arriba a la derecha, corresponde a grupos por actor, con un accuracy de 0.5212. La de abajo corresponde a grupos por actor y declaración, con un accuracy de 0.6295.

Como se puede observar en la Figura 3, el accuracy más alto corresponde a la matriz de confusión generada por la división por declaración. Esto se debe a que al dividir por declaración, el modelo puede entrenar y predecir utilizando información específica del discurso(actor). En comparación, la segunda matriz a la derecha, que utiliza la división por actor donde el modelo solo considera la información de la declaración, muestra un accuracy más bajo. Este hallazgo sugiere que las correlaciones espurias en el dataset están relacionadas con las características individuales de los actores.. Así, al eliminar esta información, el rendimiento del modelo disminuye, ya que deja de utilizar datos relevantes asociados a cada actor.

La evaluación del modelo usando tanto la división por declaración como por actor representa un enfoque riguroso y robusto. El modelo se evalúa sin usar información específica de la declaración ni del actor de validación, lo que asegura una evaluación independiente de su capacidad de generalización. Aunque podría esperarse una precisión menor debido a la exclusión de datos críticos sobre el actor y la declaración, los resultados muestran que el modelo se adapta bien a estas condiciones más estrictas, demostrando su capacidad para captar patrones generales aplicables a diferentes discursos y actores.

A continuación, decidimos explorar el objetivo planteado en la sección *2c* con respecto a la normalización por actor. El propósito fue evaluar el impacto de las variaciones individuales entre actores en la clasificación de emociones y determinar si esta forma de normalización podría simplificar la tarea del modelo al reducir la variabilidad interindividual en las características acústicas. Presentamos a continuación la matriz de confusión.

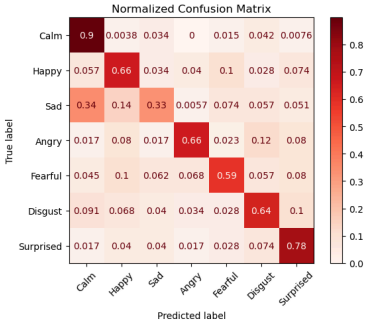

Figura 4. Matriz de confusión del modelo Random Forest evaluado con validación cruzada por discurso y por actor, utilizando normalización por actor. Accuracy = 0.6674.

Los resultados de la Figura 4 indican que la normalización por actor es, efectivamente, preferible a la normalización global (la cual mostró un desempeño equivalente al de la tercera matriz en la Figura 3, sin normalización. Esto se debe a que este modelo de aprendizaje automático es robusto a estos cambios). La normalización por actor permite al modelo considerar las variaciones individuales en la voz de cada actor, evitando así las correlaciones espurias que pueden surgir con la normalización global. Por ejemplo, la normalización global podría llevar al modelo a asociar tonos graves con hombres y tonos agudos con mujeres, lo cual podría influir erróneamente en las predicciones basadas únicamente en el género. No obstante, como se mencionó anteriormente, este tipo de normalización es inviable en la

práctica y se llevó a cabo únicamente con fines experimentales y teóricos.También se desarrollaron tres modelos para definir cuál sería el más adecuado: MLP, Random Forest y Gradient Boosting. Para llevar a cabo este experimento, se utilizó el dataset que contiene tanto speech, como song. A su vez tiene las emociones de 'Neutral' y 'Calma' unificadas, se evaluó su performance sobre el cross validation con folds por actor y por declaración y normalización global.

TABLA I.

Modelos	**Accuracy: Predicción de emociones**	
	K-Folds	***Test Set***
MLP	0.6047	0.5433
Random Forest	0.6649	0.6442
Gradient Boosting	0.6898	0.6490

Como se puede observar en la Tabla I, la MLP obtuvo los peores resultados. Esto puede deberse, en parte, a la escasez de datos con los que se trabajó. Consideramos realizar una ampliación de datos (data augmentation), pero descartamos esta idea ya que temíamos que esto podría eliminar las características únicas de cada emoción, lo cual podría confundir aún más a la red. También probamos con oversampling, pero la diferencia en los resultados fue mínima en comparación con el dataset original. Por lo tanto, decidimos mantener el dataset sin modificaciones.

Observamos que los métodos relacionados con Random Forest y Gradient Boosting son superiores a la red neuronal debido a que son modelos más simples y su rendimiento se beneficia del ensamblaje, lo que reduce significativamente el error de predicción. Entre Random Forest y Gradient Boosting, se destaca que Gradient Boosting es ligeramente superior, gracias a su capacidad para manejar mejor datos desequilibrados, reducir el sesgo del modelo mediante construcción secuencial, y capturar relaciones complejas en los datos de manera más detallada.

Para finalizar esta sección, presentamos la matriz de confusión del modelo Gradient Boosting evaluado sobre el conjunto de prueba.

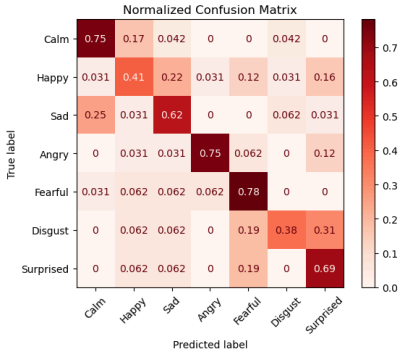

Figura 5. Matriz de confusión del Gradient Boosting sobre el set de prueba, utilizando el dataset de 'speech' y 'song', con emociones unificadas y normalización global.

Es importante destacar que las emociones que el modelo predice con mayor precisión son aquellas que son de interés para la siguiente sección.

B. *Predicción de Intensidad*

Una vez realizada la clasificación de la emoción, se experimentó con diversos modelos para la clasificación binaria de su intensidad. El principal desafío dentro de esta tarea fue encontrar un modelo que lograra captar las relaciones complejas entre los features para determinar una emoción es intensa o no y, en simultáneo, que no memorice los datos y sea capaz de generalizar. Como solo se entrenó este modelo para tres emociones (tristeza, enojo y miedo), la cantidad de datos disponibles se redujo a 1032 audios (de discurso y canción), lo cual dificultó el entrenamiento y evaluación de los modelos. Adicionalmente, se llevaron a cabo experimentos para determinar la mejor forma de representar las características de entrada a los modelos. En particular, se evaluó si el rendimiento del modelo mejoraba al incluir, además de las características eGeMAPS, la emoción real mediante un encoding one-hot, ya que previo a la clasificación de la intensidad se obtuvo un modelo que clasifica la emoción y se puede utilizar como característica de entrada adicional. La idea detrás de esta estrategia es proporcionar al modelo una probabilidad condicional: la probabilidad de que una emoción sea intensa dado que ya se ha clasificado la emoción base. Se llevaron a cabo pruebas con una Regresión Logística, un Gradient Boosting y una MLP. Todos sus hiperparámetros se tunearon utilizando validación cruzada por actor y por declaración al mismo tiempo para eliminar la mayor cantidad de correlaciones espurias en los datos.

TABLA II.

Modelos	**Accuracy: Predicción de Intensidad**			
	K-Folds		***Test Set***	
	con emoción	***sin emoción***	***con emoción***	***sin emoción***
MLP	0.85	0.80	0.76	0.68
Gradient Boosting	0.87	0.82	0.77	0.67
Regresión Logística	0.85	0.80	0.72	0.65

Con los resultados obtenidos en la Tabla II se verifica la hipótesis de que, efectivamente, todos los modelos mejoran al ser entrenados con la emoción como entrada. Esto se debe a que la inclusión de la emoción como característica adicional proporciona información contextual relevante que ayuda al modelo a diferenciar mejor entre emociones intensas y no intensas. Al conocer la emoción base, el modelo puede hacer una predicción más informada sobre la intensidad de la emoción, aprovechando la estructura condicional de las relaciones entre emociones y su intensidad ya que un enojo intenso no necesariamente tiene características similares a una tristeza profunda. Esta mejora en el rendimiento valida la estrategia de utilizar un encoding one-hot de la emoción clasificada previamente, lo cual es consistente con la teoría de que proporcionar información adicional y relevante puede mejorar la capacidad de

generalización y precisión de los modelos de clasificación de emociones, especialmente para este caso particular en donde no contamos con un gran volumen de datos. También resulta importante destacar que, nuevamente, al normalizar a todos los grupos por actor, la accuracy fue significativamente mayor: sobre los grupos de la validación cruzada y entrenando con la emoción como entrada del modelo, la regresión logística alcanzó una accuracy de 0.89, mientras que el Gradient Boosting llegó a 0.88 y la MLP a 0.87. Esto nos muestra que la individualidad y expresión del actor influye no solo en la identificación de la emoción sino también en su intensidad, ya que cada persona tiene un estilo único de expresión emocional.

Ahora bien, como los resultados de accuracy fueron similares para los modelos en la validación, decidimos verificar qué sucedía con las matrices de confusión. Para el objetivo del trabajo, resulta más importante identificar con éxito las emociones intensas y no resulta tan relevante identificar emociones no intensas. Sin embargo, hay un trade-off, ya que, dependiendo de la aplicación, una emoción potencialmente peligrosa puede requerir alguna intervención que conlleva un costo asociado. Por lo tanto, si bien queremos identificar la mayor cantidad de casos positivos con éxito, no queremos tener demasiados falsos positivos. Para abordar este desafío, la regresión logística de scikit-learn se destacó entre los modelos evaluados debido a su capacidad y simplicidad para asignar diferentes costos a cada clase. Esta característica permite ajustar el modelo para que penalice más los errores en la identificación de emociones intensas (positivos) en comparación con los errores en la identificación de emociones no intensas (negativos). De esta manera, podemos equilibrar el objetivo de maximizar la detección de emociones intensas y, al mismo tiempo, minimizar los falsos positivos. Se utilizó la ponderación de clases (class weight) con un peso de 1 para las emociones normales y un peso de 2.6 para las emociones fuertes. Así, el modelo se entrenó con una mayor penalización para los falsos negativos de emociones intensas, lo que incentivó una mejor identificación de estas emociones. Al mismo tiempo, se buscó un balance que no incrementara excesivamente el número de falsos positivos, evitando intervenciones innecesarias. El resultado de la matriz de confusión sobre los grupos de validación y el set de test se muestran a la izquierda y a la derecha respectivamente de la Figura a continuación:

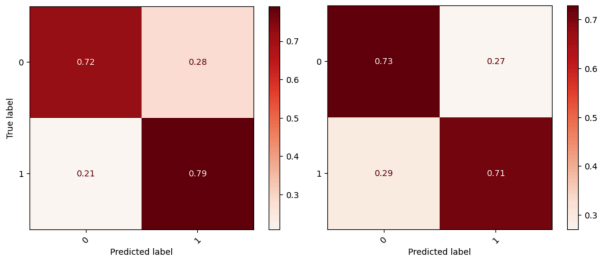

Figura 6. A la izquierda se encuentra la matriz de confusión que muestra los resultados sobre los grupos de validación, mientras que a la derecha se observan los resultados sobre el conjunto de pruebas de la Regresión Logística tuneada con los siguientes hiperparámetros: {'C': 2, 'solver': 'lbfgs', 'max_iter': 500, 'class_weight': {1: 1, 2: 2.6}}

Como se observa en la Figura 6, en ambos casos el modelo es capaz de identificar más del 70% de las emociones intensas con éxito y el porcentaje de falsos positivos no supera el 30%. Decidimos elegir este como nuestro modelo final para la predicción de intensidad ya que se distingue por su simplicidad y rapidez, sin comprometer su desempeño. Tanto la MLP como el Gradient Boosting, presentan una complejidad computacional mayor y un tiempo de entrenamiento significativamente más largo. Esto, sumado a la capacidad de la regresión logística para balancear el costo entre clases, la convierte en la opción más adecuada para nuestra aplicación.

C. Modelo Dual: emoción e intensidad

Al definir los dos modelos más convenientes para la identificación de emociones en el habla y la predicción de la intensidad de ciertas emociones que tienen el potencial de volverse peligrosas, se creó un modelo dual que integra estos dos componentes de manera secuencial. El primer componente del modelo se encarga de clasificar la emoción a partir de las características acústicas extraídas de los audios utilizando el conjunto de características eGeMAPS. Una vez determinada la emoción, el segundo componente del modelo se activa si la emoción identificada pertenece a las categorías de tristeza, enojo o miedo, que, en casos extremos, pueden representar un peligro para las personas y es necesario poder intervenir lo antes posible. Con ello se obtiene una predicción binaria para identificar si esa emoción es potencialmente crítica. A continuación se presenta un esquema del modelo.

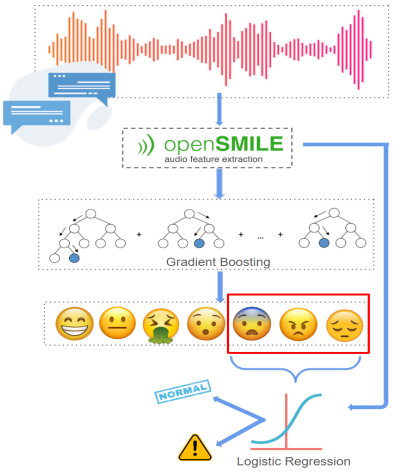

Figura 7. Esquema del Modelo Dual. Comienza con un archivo en formato .wav, extrae eGeMAPs con el módulo de openSMILE para luego enviar las 88 características a un Gradient Boosting que clasifica la emoción. En el caso de predecir susto, tristeza o enojo se le envía la emoción codificada como one-hot y las características eGeMAPs a una Regresión Logística que realiza una clasificación binaria para identificar la intensidad.

Las métricas de la predicción de emociones del modelo ya fueron ilustradas previamente en la Figura 5 . Ahora bien, a la hora de predecir la intensidad, hay que tener en cuenta que si la predicción de la emoción falla muy probablemente la predicción de la intensidad también y no será útil. Por este motivo, los resultados finales sobre el set de test obtenidos al predecir la intensidad con la Regresión Logística y la salida del Gradient Boosting se ilustran a continuación.

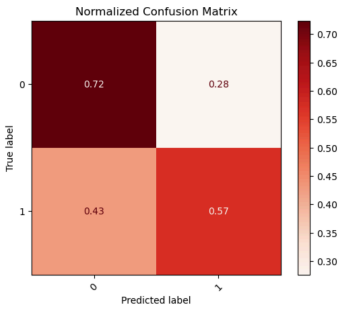

Figura 8. Matriz de confusión sobre el set de test de la predicción de la intensidad luego de la predicción de la emoción para tristeza, enojo y miedo.

La matriz de confusión final sobre el conjunto de prueba, que mostró una precisión general del modelo dual de 0.65, refleja tanto los éxitos como las limitaciones del enfoque. La precisión relativamente alta en la identificación de emociones intensas y la capacidad del modelo para manejar falsos positivos dentro de límites aceptables son logros significativos. Sin embargo, el modelo también evidencia la necesidad de más datos y mayor diversidad en los actores para mejorar su capacidad de generalización.

Estos resultados subrayan la importancia de continuar investigando y desarrollando modelos que no solo identifiquen con precisión las emociones, sino que también sean capaces de evaluar su intensidad con mayor exactitud. La individualidad y expresión del actor influyen significativamente en la identificación de emociones y su intensidad, lo que sugiere que la obtención de más datos de una variedad más amplia de actores es esencial para mejorar aún más el rendimiento del modelo.

IV. Conclusiones

Se puede concluir que realizar validación cruzada agrupando los folds por actor y por declaración es la estrategia óptima para ajustar parámetros y evaluar el rendimiento del modelo. Esta metodología permite mitigar correlaciones espurias al utilizar al actor tanto para entrenamiento como para validación, asegurando una generalización efectiva del modelo en futuras aplicaciones. En cuanto a los modelos desarrollados, encontramos que las redes neuronales presentan una complejidad considerable en relación con la cantidad limitada de datos disponibles. Por otro lado, los árboles de decisión y sus variantes ensambladas demostraron un desempeño robusto y consistente en nuestras pruebas. Tanto al momento de clasificar la emoción como al momento de identificar su intensidad, la normalización por actor mejoró significativamente la evaluación de los modelos. No obstante, esta estrategia introduce un sesgo potencial en el modelo, ya que la normalización por actor puede limitar la capacidad del modelo para generalizar a datos de actores no vistos previamente. La normalización por actor puede reducir la variabilidad individual y mejorar la precisión de los modelos en la clasificación de la intensidad emocional dentro de los datos conocidos, pero no necesariamente refleja la diversidad de un conjunto de datos más amplio. Entre los modelos desarrollados para la predicción de emociones, se seleccionó Gradient Boosting como el más destacado. Este modelo ha demostrado ser especialmente efectivo en la predicción de nuestras emociones de interés. Para la clasificación de la intensidad, se optó por la Regresión Logística por su simplicidad y buen accuracy asi como también por su flexibilidad para ajustar los pesos de las clases y encontrar un óptimo entre verdaderos y falsos positivos. Con ambos modelos se logró crear un Modelo Dual que logra identificar la emoción y la intensidad con un 0.65 de accuracy en el set de test.

III. Trabajo futuro

Futuras investigaciones deben centrarse en ampliar el conjunto de datos y en explorar nuevos enfoques de modelado que puedan capturar de manera más efectiva la complejidad de las emociones humanas. En un escenario ideal, se debería contar con una mayor cantidad de datos provenientes de una variedad más amplia de actores para entrenar y validar los modelos, asegurando así una mejor generalización y una mayor robustez frente a variaciones individuales.. Asimismo, se podrían recolectar datos con diversos niveles de intensidad emocional para proporcionar un conjunto de datos más completo y detallado que refleje la gama completa de expresiones emocionales humanas. Otra línea de investigación futura podría incluir el uso de redes neuronales profundas, como redes convolucionales y redes recurrentes, para capturar características más complejas y aprender representaciones más abstractas de las emociones a partir de grandes volúmenes de datos. Además, se podrían combinar características de audio con datos visuales para mejorar la identificación y clasificación de emociones. La integración de diferentes fuentes de información puede proporcionar una visión más completa y precisa del estado emocional de una persona, mejorando así la capacidad del modelo para reconocer y evaluar emociones en situaciones del mundo real.

IV. MATERIAL BIBLIOGRAFICO

Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391