Maestría en

CIENCIA DE DATOS Y MÁQUINAS DE APRENDIZAJE CON MENCIÓN EN INTELIGENCIA ARTIFICIAL

Trabajo previo a la obtención de título de Magister en Ciencia De Datos y Máquinas de Aprendizaje con Mención en Inteligencia Artificial

AUTOR/ES:

ANDRADE SANCHEZ MARIA AUGUSTA

AREVALO MUÑOZ ROSA ELENA

MASABANDA VINOCUNA CARLA MARIA

NARVAEZ GOMEZ GIOVANNY

TUTOR/ES:

Paulina Vizcaíno Imacaña

Alejandro Cortés López

TEMA
DETECCIÓN DE EMOCIONES EN VOZ

Quito - Ecuador

Certificación de autoría

Nosotros, **María Augusta Andrade Sánchez, Rosa Elena Arévalo Muñoz, Carla María Masabanda Vinocuna, Giovanny Alberto Narváez Gómez**, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

Firmado electrónicamente por:
MARIA AUGUSTA ANDRADE SANCHEZ
Validar únicamente con FirmaEC

Firma del graduando
María Augusta Andrade Sánchez

Firmado electrónicamente por:
ROSA ELENA AREVALO MUÑOZ
Validar únicamente con FirmaEC

Firma del graduando
Rosa Elena Arévalo Muñoz

Firmado electrónicamente por:
CARLA MARIA MASABANDA VINOCUNA
Validar únicamente con FirmaEC

Firma del graduando
Carla María Masabanda Vinocuna

Firmado electrónicamente por:
GIOVANNY ALBERTO NARVAEZ GOMEZ
Validar únicamente con FirmaEC

Firma del graduando
Giovanny Alberto Narváez Gómez

Autorización de Derechos de Propiedad Intelectual

Nosotros, **María Augusta Andrade Sánchez, Rosa Elena Arévalo Muñoz, Carla María Masabanda Vinocuna, Giovanny Alberto Narváez Gómez**, en calidad de autores del trabajo de investigación titulado ***Detección de Emociones en Voz***, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, 8 de julio de 2025

Firmado electrónicamente por:
MARIA AUGUSTA ANDRADE SANCHEZ
Validar únicamente con FirmaEC

Firma del graduando
María Augusta Andrade Sánchez

Firmado electrónicamente por:
ROSA ELENA AREVALO MUÑOZ
Validar únicamente con FirmaEC

Firma del graduando
Rosa Elena Arévalo Muñoz

Firmado electrónicamente por:
CARLA MARIA MASABANDA VINOCUNA
Validar únicamente con FirmaEC

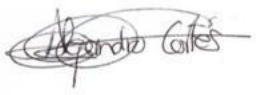
Firma del graduando
Carla María Masabanda Vinocuna

Firmado electrónicamente por:
GIOVANNY ALBERTO NARVAEZ GOMEZ
Validar únicamente con FirmaEC

Firma del graduando
Giovanny Alberto Narváez Gómez

Aprobación de dirección y coordinación del programa

Nosotros, Alejandro Cortés e Ivan Reyes, declaramos que: María Augusta Andrade Sánchez, Rosa Elena Arévalo Muñoz, Carla María Masabanda Vinocuna, Giovanny Alberto Narváez Gómez son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés López

Director de la

Maestría en Ciencia de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial

Iván Reyes Chacón

Coordinador/a de la

Maestría en Ciencia de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial

DEDICATORIA

Este proyecto está dedicado a mi familia, por ser mi mayor alegría e inspiración. A mis compañeros y profesores, por acompañarme en este camino de aprendizaje.

María Augusta

Este proyecto está dedicado a mis queridos hijos que son una fuente inagotable de inspiración y superación en mi vida y a mi amado esposo, por sus palabras de aliento que son motivación en mi vida, por su amor y paciencia que son el consuelo en mis momentos de dificultad.

Rosa Elena

Este proyecto está dedicado a mis padres: a mi madre, quien desde hace veintisiete años guía mi camino desde el cielo, y a mi padre, mi compañero en las buenas y en las malas en esta aventura llamada vida.

Carla

Este proyecto está dedicado a mi familia, por su amor incondicional, apoyo silencioso pero constante, por recordarme siempre la importancia de seguir adelante, su ejemplo de esfuerzo y su fe en mí.

Giovanny

AGRADECIMIENTOS

Deseo expresar mi más profundo agradecimiento a todas las personas que, de una u otra forma, han sido parte fundamental en la realización de este Trabajo de Fin de Máster.

En primer lugar, agradezco a mi familia, cuyo apoyo constante y palabras de aliento han sido mi mayor fuente de fortaleza a lo largo de este proceso. A mis compañeros de clase, gracias por compartir esta experiencia académica con entusiasmo, colaboración y amistad. También quiero extender mi agradecimiento a los profesores del máster, que con su guía y compromiso han aportado valiosas orientaciones que enriquecieron el aprendizaje y fortalecieron los fundamentos de este proyecto. A todos, gracias por haber formado parte de este logro académico.

A todos, gracias por haber formado parte de este logro.

María Augusta

Quiero expresar mi sincero agradecimiento a los profesores de la maestría, por compartir con generosidad sus conocimientos y experiencias a lo largo de este proceso formativo. En cada clase se evidenció el dominio que tienen sobre sus respectivas áreas, lo cual enriqueció significativamente mi aprendizaje. Asimismo, agradezco a la universidad por brindar el espacio académico, los recursos y el entorno propicio para la consecución de este objetivo académico.

Carla

Deseo expresar mi gratitud profunda a los profesores y coordinadores, por compartir sus conocimientos y motivar a mirar más allá de los datos, entendiendo que detrás de cada análisis hay una realidad que merece ser comprendida y transformada.

A mis compañeras de proyecto, por su colaboración, compromiso y amistad que hizo más llevadero este proceso. Gracias por las horas compartidas entre ideas, líneas de código, discusiones técnicas. Este proyecto final es también reflejo del trabajo colectivo y del valor de aprender en comunidad.

Giovanny

Quiero expresar mi más sincero agradecimiento a mis profesores de la maestría, quienes generosamente compartieron sus conocimientos conmigo a lo largo de este enriquecedor proceso.

Como fiel creyente agradezco a Dios, porque sin su bendición diaria y su guía constante no sería quien soy hoy. A mi familia, que ha sido el motor principal para culminar este proyecto; su amor y aliento han sido fundamentales en cada paso del camino.

A mis padres, cuyas palabras motivadoras me acompañaron en cada momento difícil, gracias por su fe inquebrantable en mí. A mis compañeros, quienes, de diversas maneras, compartieron vivencias valiosas que hicieron posible nuestra culminación en la maestría.

Finalmente, mi más profundo agradecimiento a mi compañero de vida, quien ha sido mi respaldo constante y mi pilar en esta etapa, sin su comprensión, este logro no habría sido posible.

Rosa Elena

RESUMEN

Este trabajo aborda el problema de detección de emociones en voz a través de su clasificación en seis categorías: enojado, desagrado, miedo, tristeza, felicidad y neutral. Los análisis se realizaron con un conjunto de datos balanceado de grabaciones de voz etiquetadas por emoción, de los cuales se extrajeron características acústicas utilizando la biblioteca Librosa.

Se aplicaron distintos modelos de clasificación, desde algoritmos tradicionales como Random Forest, SVM y XGBoost, hasta los basados en redes neuronales como MLP, CNN. Las métricas de evaluación calculadas como el accuracy, precision, recall, F1-score y balanced accuracy alcanzaron valores de alrededor del 76%, indicando en general un desempeño equilibrado.

El análisis por clase evidenció que la emoción "enojado" fue la mejor clasificada por todos los modelos, con un F1-score máximo del 85% en el modelo basado en MLP, lo cual sugiere que sus características vocales son más diferenciables. Por el contrario, las emociones como "desagrado", "tristeza" y "miedo" presentaron valores de F1 más bajos. Las matrices de confusión mostraron patrones recurrentes de error, especialmente entre emociones de tono bajo o activación similar, como triste-neutral y miedo-triste. Los mejores resultados se obtuvieron con modelos basados en redes neuronales profundas, particularmente MLP y CNN, lo que demuestra la efectividad de estas arquitecturas para capturar patrones en señales de voz. En conclusión, el sistema desarrollado muestra un desempeño aceptable para tareas de reconocimiento emocional por voz, aunque aún enfrenta desafíos en la discriminación de emociones de baja intensidad.

Palabras Claves: Reconocimiento de emociones, Deep learning, Modelos de clasificación, Procesamiento de audio.

ABSTRACT

This work addresses the problem of voice emotion detection by classifying them into six categories: anger, disgust, fear, sadness, happiness, and neutrality. The analyses were performed on a balanced dataset of emotion-labeled voice recordings, from which acoustic features were extracted using the Librosa library.

Different classification models were applied, ranging from traditional algorithms such as Random Forest, SVM, and XGBoost, to neural network-based models such as MLP and CNN. Evaluation metrics such as accuracy, precision, recall, F1 score, and balanced accuracy achieved values of around 76%, indicating overall balanced performance.

The class analysis showed that the "angry" emotion was the best classified by all models, with a maximum F1 score of 85% in the MLP-based model, suggesting that its vocal characteristics are more distinguishable. In contrast, emotions such as "disgust," "sadness," and "fear" had lower F1 values. The confusion matrices showed recurring error patterns, especially between emotions of low pitch or similar activation, such as sad-neutral and fear-sad.

The best results were obtained with models based on deep neural networks, particularly MLP and CNN, demonstrating the effectiveness of these architectures in capturing patterns in voice signals. In conclusion, the developed system shows acceptable performance for voice emotion recognition tasks, although it still faces challenges in discriminating low-intensity emotions.

Keywords: Emotion recognition, Deep learning, Classification models, Audio processing

TABLA DE CONTENIDOS

CAPITULO 1 5

INTRODUCCION 5
Definición Del Proyecto 6
Justificación E Importancia de trabajo de Investigación 6
Alcance 7
Objetivos 8
Objetivo General 8
Objetivos Específicos 8

CAPITULO 2 10

REVISIÓN DE LITERATURA 10
Estado del Arte 10
Marco Teórico 19

CAPITULO 3 27

DESARROLLO 27
Desarrollo Del Trabajo 27
Marco Teórico 39
Análisis de Hiperparámetros de los Modelos de Aprendizaje 39
Interpretabilidad de Modelos - Características Principales 53

CAPITULO 4 54

ANÁLISIS DE RESULTADOS 54
Pruebas de Concepto 54
Análisis de Resultados 78
Análisis De Resultados - Comparación De Modelos 78
Análisis Del Rendimiento Del Modelo 82
Principales Características – Interpretabilidad 83
Principales Características En Orden De Importancia 84

CAPITULO 5 86

CONCLUSIONES Y RECOMENDACIONES 86
Conclusiones 86
Recomendaciones 87

REFERENCIAS 89

LISTA DE TABLAS

Tabla 1 Métodos de detección de anomalías *22*
Tabla 2 Aplicaciones de la detección de anomalías *23*
Tabla 3 Estructura de carpetas del proyecto *27*
Tabla 4 Características o features extraídos *33*
Tabla 5 Características vs emociones *35*
Tabla 6 Métricas de los modelos entrenados *79*
Tabla 7 Análisis de F1 y matriz de confusión *79*
Tabla 8 Interpretación por clase *85*

LISTA DE FIGURAS

Figura 1 *Características acústicas y Clasificación de emociones en audio* 19
Figura 2 *Estructura de procesamiento del set de datos a través de la función run_pipeline* 30
Figura 3 *Distribución del data set* 31
Figura 4 *Total de emociones luego del aumento de datos* 32
Figura 5 *Métricas modelo* random_forest_model.pkl 54
Figura 6 *Métricas por clase* 54
Figura 7 *Matriz de confusión* 55
Figura 8 *Métricas modelo svm_best_model.pkl* 56
Figura 9 *Métricas por clase* 56
Figura 10 *Matriz de confusión* 57
Figura 11 *Métricas modelo xgboost_best.pkl* 58
Figura 12 *Métricas por clase* 58
Figura 13 *Matriz de confusión* 59
Figura 14 *Métricas modelo rn_model.pkl* 60
Figura 15 *Métricas por clase* 60
Figura 16 *Matriz de confusión* 61
Figura 17 *Métricas de evaluación* 62
Figura 18 *Métricas por clase* 62
Figura 19 *Matriz de confusión* 63
Figura 20 *Métricas de evaluación modelo MLP* 64
Figura 21 *Métricas por clase* 64
Figura 22 *Matriz de confusión* 65
Figura 23 *Gráfico de pérdida* 66
Figura 24 *Métricas Modelo CNN 1D* 66
Figura 25 *Métricas por clase del modelo CNN 1D* 67
Figura 26 *Matriz de Confusión del modelo CNN 1D* 68
Figura 27 *Gráfico de precisión y Pérdida durante el entrenamiento Modelo CNN 1D* 69
Figura 28 *Métricas Modelo CNN 1D L2* 69
Figura 29 *Métricas por clase del modelo CNN 1D L2* 70
Figura 30 *Matriz de Confusión modelo CNN 1D L2* 71
Figura 31 *Gráfico de precisión y Pérdida durante el entrenamiento Modelo CNN 1D L2* 72
Figura 32 *Métricas Modelo CNN - LSTM* 72
Figura 33 *Matriz por clase del modelo CNN - LSTM* 73
Figura 34 *Matriz de Confusión modelo CNN - LSTM* 74
Figura 35 *Gráfico de precisión y Pérdida durante el entrenamiento Modelo CNN - LSTM* 75
Figura 36 *Métricas Modelo CNN - LSTM L2* 75
Figura 37 *Matriz por clase del modelo CNN - LSTM L2* 76
Figura 38 *Matriz de Confusión modelo CNN - LSTM L2* 77

Figura 39 *Gráfico de precisión y Pérdida durante el entrenamiento Modelo CNN - LSTM L2* 78
Figura 40 *Gráfica de Pérdida durante el entrenamiento* 82
Figura 41 *Gráfica de Pérdida durante el entrenamiento* 84

CAPITULO 1

INTRODUCCION

La detección de emociones a partir del habla es un área de la Inteligencia Artificial con un gran potencial en diversas aplicaciones. La capacidad de un sistema para interpretar las emociones expresadas en la voz puede revolucionar varios puntos que se mencionan a continuación:

Interacción Humano-Computadora.

Un estudio de MIT (Massachusetts Institute of Technology) encontró que los sistemas que pueden interpretar emociones humanas mejoran significativamente la satisfacción del usuario. Al integrar la detección de emociones, las interacciones pueden volverse un 30% más efectivas.

Atención Médica.

La Organización Mundial de la Salud (OMS) ha señalado que la salud mental es una prioridad global. La detección temprana de emociones a través del habla puede ayudar a identificar problemas de salud mental, lo que podría reducir el costo de la atención médica en un 30%.

Marketing Y La Seguridad.

Un estudio de Gartner indica que el 70% de las decisiones de compra se basan en emociones. Las empresas que utilizan análisis de emociones en sus estrategias de marketing pueden aumentar sus tasas de conversión en un 20%.

Este proyecto busca abordar estos problemas desarrollando un sistema de detección de emociones en voz que utilice técnicas avanzadas de aprendizaje automático y procesamiento de señales. Al centrarse en un conjunto limitado de emociones básicas y utilizar datasets etiquetados, se espera mejorar la precisión y eficacia en la identificación de emociones, contribuyendo a una mejor interacción humano-computadora y a aplicaciones prácticas en diversas áreas.

Definición Del Proyecto

El presente trabajo tiene como objetivo realizar el reconocimiento de emociones por voz utilizando técnicas de Machine Learning y Deep Learning. El proyecto realiza y ejecuta varios modelos de Ciencia de Datos, yendo desde modelos menos complejos como es RandomForest a modelos más complejos de redes neuronales profundas para identificar las emociones como miedo, felicidad, desagrado, tristeza, enojo y neutral. Para ello, se emplean bases de datos públicas etiquetadas y se han entrenado los modelos, para esto se extrajo las características de los audios y se aumentó datos sintéticos. El enfoque se centra en evaluar la precisión de los diferentes modelos propuesto y su capacidad para generalizar ante nuevos datos. El proyecto no aborda el reconocimiento multimodal ni el análisis en tiempo real, y se limita a emociones expresadas en inglés en entornos controlados. Este trabajo busca contribuir al avance de sistemas inteligentes con sensibilidad emocional, con posibles aplicaciones en áreas como salud, educación o experiencia de usuario.

Justificación E Importancia de trabajo de Investigación

La comunicación humana no se limita a las palabras; las emociones juegan un papel fundamental en la transmisión de mensajes. Actualmente, la mayoría de las interacciones con sistemas informáticos no consideran el componente emocional, lo que limita la naturalidad y la eficacia de la comunicación. Un sistema capaz de detectar emociones en la voz podría mejorar significativamente estas interacciones, permitiendo a las máquinas comprender mejor las necesidades y los estados emocionales de los usuarios.

En el ámbito tecnológico el uso de técnicas avanzadas de procesamiento de señales y aprendizaje automático en la detección de emociones es un área de investigación en crecimiento. Este proyecto contribuirá al avance de la tecnología en este campo, ofreciendo un enfoque innovador para mejorar la interacción humano-computadora. Con el auge del deep learning y transformers aplicados a

audio, es posible mejorar la precisión del reconocimiento de emociones en comparación con enfoques tradicionales.

En el ámbito de la atención médica, la detección de emociones podría auxiliar en la identificación temprana de trastornos mentales, ayudando a los profesionales a brindar un diagnóstico más preciso y oportuno. El reconocimiento de emociones en voz puede contribuir al bienestar emocional al aplicarse en entornos como terapia en línea y análisis de llamadas de emergencia. En marketing, la identificación de las emociones de los clientes frente a un producto o servicio puede optimizar las estrategias publicitarias. En seguridad, un sistema de detección de emociones en voz podría ayudar a identificar posibles amenazas o comportamientos sospechosos.

Existen diversos métodos para la detección de emociones en el habla, pero la precisión de estos sistemas sigue siendo un desafío. Este proyecto busca contribuir al avance en esta área, proponiendo una solución que combine técnicas de procesamiento de señales y aprendizaje automático para obtener una mayor precisión en la identificación de emociones.

Alcance

El sistema se centrará en la detección de emociones a partir de grabaciones de voz en entornos controlados.

No se abordarán otras modalidades de detección emocional, como el análisis facial o de texto.

La recolección de datos se limitará a conjuntos de datos públicos y grabaciones personalizadas en un entorno específico.

Se analizarán audios de hombres y mujeres adultos.

La efectividad del modelo dependerá de la calidad y variedad de los datos de entrenamiento. Las bases de datos utilizadas pueden no representar toda la gama de emociones o variaciones de acento, idioma o contexto.

El modelo estará limitado a las emociones y características presentes en las bases de datos de entrenamiento. Su capacidad para generalizar a otros contextos o a datos de voz en tiempo real podría ser limitada.

Las emociones humanas son complejas y a menudo se mezclan, por lo que la precisión en la clasificación de emociones puede no ser perfecta, especialmente en emociones complejas o sutiles.

Objetivos

Objetivo General

Este proyecto se centrará en el desarrollo de un sistema de detección de emociones en voz que sea capaz de distinguir entre un conjunto predefinido de emociones básicas (alegría, tristeza, enojo, miedo).

Existen datasets etiquetados que permiten el entrenamiento de modelos sin necesidad de recopilar información propia, facilitando la implementación del estudio. El sistema se entrenará con un conjunto de datos de voz etiquetados, y se evaluará su desempeño mediante métricas de precisión, recall y F1-score.

Objetivos Específicos

El alcance del proyecto incluye las siguientes etapas:

- Recopilación y preprocesamiento de datos: Selección y limpieza de un conjunto de datos de voz etiquetados con las emociones objetivo (alegría, tristeza, enojo, miedo).
- Extracción de características: Análisis de características acústicas del habla, tales como tono de voz, intensidad, ritmo y timbre, utilizando técnicas de procesamiento de señal.
- Diseño e implementación del modelo de clasificación: Entrenamiento de un modelo de aprendizaje automático (ej. redes neuronales, máquinas de soporte vectorial) utilizando las características extraídas.

- Evaluación del sistema: Medición de la precisión del sistema utilizando datos de prueba y análisis de resultados.
- Documentación del sistema: Elaboración de un informe técnico que detalle la metodología utilizada, los resultados obtenidos y las limitaciones del sistema.
- El proyecto no pretende cubrir la totalidad de las emociones humanas, ni abordar todos los posibles escenarios de aplicación. Se centrará en un conjunto limitado de emociones básicas y en la creación de un prototipo funcional que sirva como base para futuros desarrollos.

CAPITULO 2

REVISIÓN DE LITERATURA

Estado del Arte

Definición Y Contexto Del Reconocimiento De Emociones En Voz. El vínculo entre la emoción y la voz ha sido objeto de múltiples estudios desde diferentes disciplinas. Tal como señaló Lope de Vega, “Mal puede tener la voz tranquila quien tiene el corazón temblando” (Sagredo, s.f.), lo cual anticipa una verdad ampliamente documentada: las emociones influyen directamente en la producción vocal. Los movimientos musculares del aparato respiratorio y de la laringe se ven afectados por los estados emocionales, modificando así el tono, el ritmo y la intensidad de la voz.

Las emociones básicas presentan características vocales prototípicas. Por ejemplo, la alegría se asocia a un tono alto y ritmo rápido; la tristeza a un tono grave y ritmo lento; el miedo a una entonación monótona y débil; y la cólera a inflexiones bruscas e intensidad fuerte. La entonación —frecuentemente definida como la “melodía” del habla— varía no solo en oraciones completas, sino también en palabras o sílabas aisladas, y permite expresar una amplia gama de actitudes como sorpresa, duda, afirmación o rechazo.

La voz es un reflejo del estado general del individuo, incluyendo dimensiones físicas, emocionales y sociales. Profesionales como médicos, psicólogos, pedagogos y músicos analizan tanto la voz como los elementos paraverbales (suspiros, silencios, ritmo respiratorio) y corporales (postura, gestos) para evaluar el estado de salud de una persona. Una voz sana se caracteriza por su claridad, vitalidad, pureza de timbre y ausencia de tensión o esfuerzo. En contraste, ciertas patologías como la depresión o la esquizofrenia se manifiestan en voces monótonas, lentas y con timbre alterado.

Diversos estudios demuestran que los oyentes emiten juicios sobre la personalidad, el estado emocional e incluso la profesión del hablante basándose únicamente en su voz. Estos juicios se ven

influenciados por estereotipos sonoros, como ocurre con las voces de sacerdotes o actores, que son fácilmente reconocidas o asociadas incluso de forma errónea pero consistente.

Finalmente, la voz humana también funciona como una “huella vocal” individual, utilizada incluso en procedimientos judiciales mediante análisis de sonogramas. En el ámbito médico, esta herramienta tiene un potencial diagnóstico importante gracias al análisis computarizado de patrones vocales anómalos. La intensidad emocional puede alterar profundamente el funcionamiento de los órganos fonadores: desde una voz entrecortada por espasmos del diafragma, hasta una voz enronquecida o un bloqueo completo del habla, conocido como “nudo en la garganta”.

Fayek, H. M., Lech, M., & Cavedon, L. (2017) en su investigación abordan sobre “Evaluating deep learning architectures for Speech Emotion Recognition”, describen que, en los últimos años se ha usado la inteligencia artificial especialmente las redes neuronales profundas, para mejorar el reconocimiento de emociones en la voz. Este estudio compara diferentes tipos de redes para ver cuál identifica mejor las emociones al hablar. Para esto, se usa grabaciones de diálogos emocionales y se prueba métodos que analizan la voz por partes pequeñas llamadas “marcos”. Se demuestra que las redes neuronales convolucionales (ConvNets) son las más efectivas para distinguir emociones, sobre todo cuando se usan técnicas que evitan errores comunes como el sobreajuste. Este trabajo ayuda a desarrollar sistemas que entienden mejor cómo se siente una persona por su forma de hablar, lo que puede aplicarse en asistentes virtuales o herramientas para mejorar la comunicación.

Correa L. (2019) presenta como trabajo de tesis “Reconocimiento automático de emociones en audio y video usando Machine Learning”, desarrolla un modelo capaz de identificar emociones humanas como alegría, tristeza o sorpresa a partir de grabaciones de audio y video, usando algoritmos de redes neuronales. Para ello, se utilizó la base de datos RAVDESS y se implementaron tres clasificadores: uno que reconoce emociones por la expresión facial en video, otro por el tono de voz en audio, y un tercero

que combina ambos. El modelo final alcanzó precisiones de hasta 97% en video y 91% al combinar audio y video. La tesis demuestra que es posible detectar emociones automáticamente y con alta precisión, lo que abre la puerta a aplicaciones en atención al cliente, seguridad y robótica social, aunque se identifican limitaciones como la necesidad de que la persona mire de frente a la cámara y hable en entornos sin ruido. Se concluye que con bases de datos más diversas y mejoras en los modelos, estos sistemas podrían adaptarse a aplicaciones reales.

Livingstone S. & Russo F. (2018) presentan su estudio "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)" el estudio presenta la creación y validación de la base de datos RAVDESS, un conjunto de grabaciones de audio y video que muestra expresiones emocionales dinámicas en inglés norteamericano, realizadas por 24 actores profesionales. Incluye 7356 grabaciones de habla y canto con emociones como alegría, tristeza, enojo, miedo, sorpresa y disgusto, producidas a dos niveles de intensidad y en diferentes modalidades (audio-video, solo audio, solo video).

El objetivo es proporcionar un recurso validado para investigaciones sobre reconocimiento de emociones en voz y rostro. La base fue evaluada por cientos de participantes que calificaron precisión, intensidad y autenticidad de las emociones, demostrando alta validez y fiabilidad en las expresiones capturadas. RAVDESS es gratuita para la comunidad científica y destaca por ser una de las pocas bases con expresiones emocionales cantadas validadas, lo que la hace útil para estudios en neurociencia, psicología, psiquiatría, audiología y tecnologías como el reconocimiento automático de emociones.

López, F. (2019). Reconocimiento de emociones en la voz mediante aprendizaje profundo, este trabajo busca mejorar el reconocimiento de emociones en la voz cuando se enfrenta a hablantes desconocidos, ya que los modelos suelen perder precisión fuera de sus datos de entrenamiento. Para ello, se entrenaron clasificadores basados en transformers con modelos preentrenados (HuBERT y Wav2Vec2) en seis datasets (CAFE, CREMA-D, EMOFILM, RAVDESS, SAVEE y TESS), logrando precisiones

muy altas (hasta 100 %) al evaluar en los mismos datos. Sin embargo, al probar estos modelos con datasets no vistos, la precisión bajó drásticamente (por ejemplo, a 33-39 % en CREMA-D). Al combinar varios datasets en el entrenamiento, la precisión frente a hablantes desconocidos mejoró notablemente, alcanzando un 54,76 % en CREMA-D, demostrando que agregar variedad en los datos de entrenamiento incrementa la capacidad de generalización de los modelos.

Khalil R., Jones E., Babar M., Jan T., Zafar M. & Alhussain T. (2017), el artículo “Speech Emotion Recognition using Deep Learning Techniques”, revisa las técnicas de aprendizaje profundo aplicadas al reconocimiento de emociones en la voz, destacando métodos como DBM, RNN, DBN, CNN y AE. Estas técnicas facilitan el entrenamiento de modelos y son eficientes gracias al uso de pesos compartidos. Sin embargo, presentan limitaciones como arquitecturas internas complejas, menor eficiencia con datos temporales variables y riesgos de sobreaprendizaje. El estudio sirve para evaluar el desempeño actual y señala posibles mejoras para futuros sistemas de reconocimiento emocional.

Aplicaciones Actuales En Distintas Áreas. Las redes neuronales se pueden aplicar en varias áreas como:

Salud Mental: El reconocimiento de emociones en la voz se ha convertido en una herramienta clave en el ámbito de la salud mental, permitiendo detectar signos tempranos de depresión, estrés o ansiedad a través de análisis de voz, expresiones faciales y señales no verbales, a través de la aplicación de redes neuronales para mejorar la evaluación emocional de pacientes en contextos clínicos. Se desarrollan chatbots y robots terapéuticos capaces de identificar emociones como ira, tristeza o miedo y responder empáticamente mediante técnicas como la respiración guiada o el refuerzo positivo. Estos dispositivos no sustituyen al terapeuta, pero complementan su labor, ofreciendo apoyo continuo, monitoreo emocional y estimulación personalizada, especialmente útil en personas con autismo, demencia o aislamiento social. Si bien aún existen desafíos éticos y técnicos, su integración en centros

de salud mental, hogares o residencias representa un avance prometedor para ampliar el acceso y la eficacia del cuidado psicológico.

Asistentes virtuales: Los asistentes personales basados en inteligencia artificial (IA), como Siri y Alexa, han evolucionado desde herramientas que ejecutan comandos básicos hacia sistemas inteligentes que anticipan las necesidades del usuario, comprenden sus emociones y ofrecen respuestas personalizadas. Esta evolución se sustenta en tecnologías avanzadas de aprendizaje automático, procesamiento de lenguaje natural e inteligencia emocional artificial, que permiten a los asistentes reconocer estados emocionales mediante el análisis de voz y texto, y ajustar sus interacciones para mejorar la experiencia del usuario.

Educación personalizada: El reconocimiento emocional mediante la voz puede ser utilizado en plataformas de aprendizaje digital para detectar frustración, desinterés o entusiasmo en los estudiantes. Esto permite adaptar los contenidos, cambiar el ritmo de las actividades o activar alertas para intervención docente, mejorando el compromiso y la eficacia del aprendizaje.

Seguridad y monitoreo en tiempo real: Sistemas de seguridad pueden incorporar análisis emocionales en la voz para detectar signos de estrés, agresividad o pánico en llamadas de emergencia, interrogatorios o espacios públicos. Esto permite respuestas más rápidas y precisas ante posibles situaciones de riesgo o violencia.

El reconocimiento de emociones en la voz representa una tecnología emergente con un potencial transformador en múltiples sectores. Su capacidad para interpretar matices afectivos en la comunicación oral permite no solo optimizar la interacción humano-máquina, sino también fortalecer procesos de acompañamiento emocional, detección de crisis y personalización de servicios. Desde la salud mental y la educación hasta la seguridad, el entretenimiento y la atención al cliente, esta tecnología abre nuevas posibilidades para desarrollar sistemas más empáticos, adaptativos y eficaces, sin

embargo; debe coexistir el equilibrio entre la innovación tecnológica y la responsabilidad ética, mediante políticas claras que garanticen la protección de la información personal y la transparencia en su uso, asegurando así la confianza y el bienestar de los usuarios.

Desafíos Generales Y Complejidades Del Área. El análisis de emociones a través de la voz constituye un gran desafío debido a las complejidades del área, tanto en la variabilidad de emociones como en la calidad de los datos que se obtienen para el entrenamiento de los diferentes modelos.

Uno de los desafíos que se debe abordar es la variabilidad emocional las cuales pueden cambiar significativamente entre individuos y culturas, complicando su identificación precisa. Ha sido materia de estudio la relación entre la cultura y la forma en que las personas expresan sus emociones. Las normas culturales y expectativas sociales determinan qué emociones se consideran apropiadas o tienen un contexto positivo y cómo deben expresarse y de la misma manera, se identifican emociones consideradas negativas y que deben ser en ocasiones reprimidas. Cada cultura posee formas únicas de manifestar sentimientos, lo que genera diferencias significativas en la comunicación emocional. Estas variaciones pueden enriquecer las relaciones interculturales o generar interpretaciones incorrectas.

Otro de los desafíos que se debe enfrentar es la calidad de los datos, es decir las bases de datos disponibles suelen ser limitadas en tamaño y diversidad, afectando la generalización de los modelos. Factores como el ruido de fondo y la calidad de la grabación pueden interferir con la precisión del reconocimiento de emociones en voz. En ocasiones es necesario recurrir a la generación de data sintética (o datos sintéticos), que no es más que la información generada artificialmente mediante algoritmos, modelos estadísticos o inteligencia artificial, en lugar de ser recopilada directamente del mundo real. Su objetivo principal es simular datos reales sin comprometer la privacidad o requerir grandes esfuerzos de recolección. Para el caso específico de análisis de audios, se recurren a transformaciones de las señales con el fin de aumentar un conjunto de datos (data augmentation),

generando versiones modificadas de los datos originales para mejorar la robustez de modelos de aprendizaje automático. Entre las transformaciones más comunes están el agregar ruido aleatorio a la señal de audio original para simular imperfecciones o variaciones naturales, estirar o comprimir la duración del audio sin alterar el tono, modificando su velocidad, desplazar temporalmente la señal de audio hacia adelante o hacia atrás de manera aleatoria y cambiar el tono del audio sin alterar su duración, desplazando las frecuencias.

Finalmente; otra de las aristas que se debe considerar es la parte regulatoria, ética y la privacidad de los datos. Este es un tema analizado y discutido a nivel mundial como parte de la regulación del uso de inteligencia artificial y dentro de los estudios proyectos realizados para reconocimiento de emociones en voz se plantean preocupaciones sobre la privacidad y el consentimiento informado de los usuarios. En el Ecuador se presentaron tres proyectos de ley los cuales abordan distintos enfoques: regulatorio, de desarrollo tecnológico y de protección de derechos infantiles frente a la IA.

El Proyecto de Ley Orgánica de Regulación y Promoción de la IA propone un marco integral inspirado en la UE, clasificando sistemas según niveles de riesgo y creando una Autoridad Nacional de Control de IA. Establece 16 principios rectores y se aplicará en complemento con otras leyes, priorizando la protección de derechos fundamentales.

El Proyecto de Ley para el Fomento y Desarrollo de la IA busca impulsar la adopción e investigación en IA con incentivos económicos y educativos. Ambos proyectos comparten principios como privacidad y transparencia, y contemplan un “sandbox” para tecnologías de alto riesgo.

El Proyecto de Ley de Aprovechamiento Digital e IA para Niños y Adolescentes regula el uso seguro de IA y plataformas digitales enfocadas en menores. Crea la ANSIA y prohíbe ciertas tecnologías

que impliquen manipulación, clasificación injustificada o reconocimiento facial en tiempo real, salvo casos excepcionales.

Extracción De Características Acústicas. En el reconocimiento de emociones en la voz, cada característica aporta información diferente sobre cómo se expresa una emoción (como alegría, tristeza, ira, miedo, etc.) a través del tono, ritmo, energía y timbre vocal. Las características o descriptores de audio, generales y ampliamente utilizada en el análisis de señales de audio y música, especialmente en tareas como el reconocimiento de género musical, clasificación de sonidos, análisis de emociones, y reconocimiento de voz son las que se describen a continuación:

Zero-Crossing Rate (ZCR). Cuántas veces la señal cruza el eje horizontal (valor cero). Bueno para detectar sonidos percutivos o distinguir entre voz y ruido. Indica la actividad de la señal, útil para diferenciar entre emociones activas vs. pasivas. Por ejemplo, una voz alegre o enojada tendrá mayor ZCR por la energía y la velocidad del habla, mientras que una voz triste tendrá un ZCR más bajo. (Open IA 2025)

La Transformada de Fourier de Corto Plazo (STFT). Permite analizar cómo cambian las frecuencias de una señal a lo largo del tiempo. La STFT divide la señal en pequeñas ventanas temporales y calcula la transformada de Fourier en cada una. Esto genera una representación tiempo-frecuencia que muestra qué frecuencias están presentes en cada instante del audio. Cómo se puede utilizar en el reconocimiento de emociones, por ejemplo, en la alegría se suele observar un espectro con mayor energía en frecuencias medias-altas, con patrones rítmicos más rápidos y variaciones tonales frecuentes, la STFT mostrará bandas con energía cambiante y dinámica; por el contrario, la tristeza concentra la energía en frecuencias más bajas y con menor variabilidad temporal. La STFT mostrará una señal más homogénea y estable en el tiempo-frecuencia. (Open IA 2025)

Coeficientes Cepstrales en las Frecuencias de Mel (MFCC). Representa la envolvente espectral de una señal. Convierte el espectro de frecuencia a una escala mel (más cercana a cómo los humanos percibimos el sonido), luego aplica una transformada logarítmica y otra transformada discreta del coseno (DCT). Muy utilizado en reconocimiento de voz, clasificación de música y análisis de timbre. Captura el timbre de la voz, que cambia con la emoción. Ejemplo: Una voz enojada puede tener un timbre más tenso y una voz triste suele ser más suave y apagada.

RMS (Root Mean Square Energy). Mide la energía promedio de la señal en un segmento. Indica la intensidad o volumen percibido de una señal. Mide la intensidad o volumen percibido de la voz. Ejemplo: La ira o la alegría pueden tener mayor energía que el miedo o la tristeza, que tienden a ser más apagados.

Espectrograma de mel. Las frecuencias han sido convertidas a la escala Mel, que simula cómo el oído humano percibe el tono (más sensible a cambios en frecuencias bajas). Es un mapa de energía: frecuencia Mel × tiempo. Si se analiza una voz con alegría se puede mostrar un espectrograma de Mel con energía más dispersa y variaciones rápidas en el tiempo, especialmente en las bandas medias-altas, mientras que la energía más concentrada y uniforme en frecuencias bajas y menos variación temporal representa una voz con tristeza. La ira puede manifestarse con picos de energía en frecuencias medias-altas y patrones de vibración más tensos.

Figura 1

Características acústicas y Clasificación de emociones en audio

Marco Teórico

Aspectos Psicológicos Y Lingüísticos.

Categorías Básicas: Se puede clasificar las emociones como: alegría, tristeza, enojo y miedo; estas categorías son generales y se pueden expresar de manera parecida en diferentes culturas.

Dimensiones Afectivas: Se pueden describir las emociones mediante dimensiones como:

Valencia: Describe la calidad apropiada de los sentimientos, es decir, si es Positivo (Alegría, amor, satisfacción) o Negativo (Tristeza, ira, miedo)

Activación: Indica el nivel de excitación asociado con una emoción, puede cambiar desde un estado de baja activación (Serenidad, tristeza, relajación) a un estado de alta activación (Euforia, ansiedad, entusiasmo).

El modelo que aplica estas dos dimensiones puede visualizar y clasificar las emociones de manera más efectiva.

Expresión Vocal De Emociones: Son teorías sobre cómo se manifiestan en la voz.

Teorías Sobre La Manifestación En La Voz: La vocalización de las emociones se manifiesta en cómo varían los parámetros acústicos como el tono, la intensidad y el ritmo en respuesta a distintos estados emocionales.

Influencia Cultural Y Lingüística En La Expresión Emocional: La manera en que las emociones se expresan verbalmente puede cambiar según la cultura y el idioma, lo que influye en la interpretación y en el reconocimiento de emociones en el habla.

Características Acústicas Relevantes Para El Reconocimiento De Emociones.

Parámetros De Voz

Prosodia: Son todas las variaciones tanto en el tono como en el ritmo que pueden mostrar diferentes emociones.

Frecuencia fundamental: La emoción expresada altura del sonido puede asociarse con la altura del sonido.

Energía: El nivel de emoción puede manifestarse por la intensidad del habla.

MFCC (Mel-frequency cepstral coefficients): Sirve para tomar características acústicas importantes y es aplicado en el procesamiento de señales.

Jitter y shimmer: Mediciones de variabilidad en la frecuencia y amplitud, respectivamente, pueden ser indicadores de estrés o emoción.

Técnicas De Preprocesamiento Y Extracción De Características.

Filtrado y normalización: Procedimientos de limpieza y estandarización de los datos de audio antes de la obtención de características.

Extracción de características acústicas: Uso de algoritmos para trasformar señales de audio en interpretaciones numéricas que pueden ser aplicadas en modelos de aprendizaje automático.

Innovaciones En Embeddings Y Representación Es Acústicas. Embeddings acústicos: Técnicas recientes que permiten describir características acústicas de forma más eficiente, optimizando la capacidad del modelo para reconocer emociones.

Fundamentos Técnicos De Modelos De Aprendizaje Automático Y Profundo. El aprendizaje supervisado involucra entrenar modelos aplicando datos etiquetados para que puedan predecir resultados en datos no vistos.

"Su objetivo es aprender una función que mapee las entradas a las salidas deseadas, de forma que podrá hacer predicciones de datos no vistos en el futuro." (López, 2025)

Algoritmos comunes en la clasificación emocional. SVM (Máquinas de Soporte Vectorial): Son conjunto de métodos de aprendizaje supervisado utilizados para la clasificación y regresión de datos en espacios de alta dimensión con la intención de separar las distintas clases de datos de la manera más eficiente posible.

k-NN (k-Nearest Neighbors): Un método de aprendizaje supervisado simple que clasifica un dato en función de la mayoría de sus k vecinos más cercanos. Este algoritmo se puede aplicar en problemas de relación entre las características y las clases no es lineal.

HMM (Modelos Ocultos de Markov): Eficientes para modelar secuencias temporales y datos de audio. Son generalmente utilizados en el análisis de datos secuenciales, como el reconocimiento de voz y la bioinformática. Estos modelos permiten derivar estados ocultos a partir de observaciones visibles.

GMM (Modelos de Mezcla Gaussiana): Utilizados para modelar la distribución de características acústicas y son utilizados en varias aplicaciones como:

- Agrupamiento (Clustering): Es una técnica de aprendizaje no supervisado con la intensión de agrupar un conjunto de clusters con el objetivo de que éstos sean más semejantes entre sí que con los otros grupos.

- Reconocimiento de patrones: Es una técnica importante en el estudio de datos, se utiliza en varias aplicaciones incluyendo el análisis de imágenes como :
    - Análisis de Imágenes: Implica sacar información valiosa de imágenes digitales.
    - Aplicaciones: Se puede aplicar en diferentes estudios como:
    - Radiografías, resonancias magnéticas
    - Vehículos autónomos para la detección de objetos.
    - Fotografías para el reconocimiento de escenas
    - Técnicas utilizadas:
    - Segmentación de Imágenes: Para el análisis se realiza una división de varios objetos dentro de una imagen.
    - Clasificación: Etiquetar las imágenes sustentadas en patrones reconocidos.

Detección de anomalías: Es un proceso importante en el análisis de datos que se utiliza para identificar datos atípicos o anormales en amplios conjuntos de datos. A continuación, se detallan sus métodos y aplicaciones:

Tabla 1

Métodos de detección de anomalías

Metódo	Aplicación
Métodos Estádisticos	Z-Score:Utiliza la media y la desviación estándar para identificar puntos que se desvían más allá de un cierto umbral. Pruebas de hipótesis: Evaluar si un punto de datos pertenece a una distribución esperada.

Aprendizaje Automático	Modelos Clustering: Algortimos como K-means pueden identificar puntos que no perteneces a ningún grupoe consideran como anomalías. Redes Neuronales: Modelos como autoencoders puede aplicar representaciones de datos normales y detectar desviaciones.
Métodos basados en Distancia	K-vecinos más cercanos K-NN: Detecta anomalías tomando en consideración la distancia de los vecinos más cercanos.

Tabla 2

Aplicaciones de la detección de anomalías

Aplicaciones	Uso
Seguridad Informática	Reconocer fraudes y actividades dudosas en transacciones financieras.
Mantenimiento Predictivo	Identificación de fallos en maquinarias a partir de información de sensores.
Salud	Determinar patrones inusuales en datos médicos, como resultados de pruebas.
Control de calidad	Detección de productos en mal estado en líneas de producción.

Introducción a redes neuronales, CNN, RNN, LSTM, Transformers y su aplicación en audio. CNN (Redes Neuronales Convolucionales): Esta red es práctica para el procesamiento de datos en forma de imágenes y audio. Tiene algunas características como:

Estructura Jerárquica: Las CNN están creada para reconocer patrones jerárquicos, es decir, pueden conocer características simples en las primeras capas y características más complejas en las capas posteriores.

Convoluciones: Intervienen operaciones de convolución para obtener características de los datos de entrada, que facilita reconocer patrones locales.

Pooling: Ejecutar capas de pooling (submuestreo) para disminuir la dimensionalidad de los datos y mantener las características más significativas.

RNN (Redes Neuronales Recurrentes): Son redes sólidas para manejar secuencias de datos, como el audio y el lenguaje. A continuación, algunas características:

Conexiones Recurrentes: Las RNN tienen conexiones que facilitan que la información responda.

Manejo de Secuencias: Son idóneos para procesar secuencias de longitud variable, lo que las hace apropiadas para datos como texto, audio y video.

Memoria a Corto Plazo: Las RNN pueden retener información reciente, pero pueden tener complicaciones para mantener información a largo plazo a causa de problema del desvanecimiento del gradiente.

LSTM (Memoria a Largo Plazo): Una variante de RNN que lleva mejor las dependencias a largo plazo en los datos.

Transformers: Modelos que han modificado el procesamiento del lenguaje y que pueden utilizarse al análisis de audio.

Interpretabilidad de Modelos - Características Principales. De acuerdo con Molnar (2022), cuanto más interpretable sea un modelo de aprendizaje automático, más fácil será comprender y explicar por qué se tomaron ciertas decisiones o cómo se llegaron a determinadas predicciones. Un

modelo es más interpretable que otro si sus decisiones son más fáciles de entender para el humano y por lo tanto genera confianza sobre los resultados.

Podemos definir algunos objetivos al utilizar los métodos de interpretabilidad, entre ellos tenemos que nos permite mejorar el modelo. La interpretabilidad ayuda a depurar el modelo al identificar cuándo comete errores. Algunos errores pueden ser tan simples como la codificación incorrecta de la característica objetivo o errores en los cuales las características contradicen el conocimiento o dominio que se tiene sobre el problema o análisis que se realiza. Quizás se haya cambiado las clases objetivo y una característica que se sabe que es importante no es utilizada por el modelo. Es posible que se haya cometido un error en el procesamiento de datos o en la ingeniería de características.

Otro uso de la interpretabilidad es justificar el modelo y las predicciones. El aprendizaje automático interpretable ayuda a justificar el modelo y sus predicciones ante otras personas o entidades. Las partes interesadas de un sistema de aprendizaje automático pueden ser los creadores que construyen el sistema y entrenan el modelo, los operadores interactúan directamente con el sistema, los ejecutores toman decisiones basadas en los resultados, los sujetos de decisión se ven afectados por las decisiones, los auditores auditan e investigan el sistema. Estas partes interesadas buscan una justificación del modelo y sus predicciones, y pueden requerir diferentes tipos de justificación.

SHAP (Explicaciones Aditivas de Shapley) es un método para explicar predicciones individuales. SHAP calcula los valores de Shapley a partir de la teoría de juegos de coalición, como se explicó en el capítulo sobre los valores de Shapley. El objetivo de SHAP es explicar la predicción de una instancia calculando la contribución de cada característica a la predicción. Los valores de las características de una instancia de datos actúan como participantes en una coalición. Los valores de Shapley nos indican cómo distribuir equitativamente la predicción entre las características.

Una limitacion que se puede considerar en SHAP es que el KernelSHAP es lento. Esto hace que su uso sea poco práctico cuando se desea calcular valores de Shapley para muchas instancias. Además, todos los métodos SHAP globales, como la importancia de las características SHAP, requieren calcular valores de Shapley para muchas instancias.

CAPITULO 3

DESARROLLO

Desarrollo Del Trabajo

El script desarrollado en Google Colab realiza un proceso integral para el análisis y clasificación de emociones a partir de datos de audio, empleando diversos modelos de aprendizaje automático. A continuación, se detallan las etapas del procedimiento:

Configuración Inicial Y Carga De Módulos

Con el objetivo de estructurar y hacer el proyecto más legible, fácil de ejecutar y entender, se creó una estructura de directorios que se encuentran en github. La estructura del proyecto final contiene tres subcarpetas en la carpeta ProyectoFinalG5:

Tabla 3

Estructura de carpetas del proyecto

Carpeta/Archivo	Descripción
/src	Funciones de carga de datos, extracción de features, generación de datos sintéticos, división de dataset. Adicionalmente, se almacenan los archivos .pkl de los set de datos train, val y test.
/models	Funciones con los modelos que se entrenaron. Se almacenan los modelos en formato .pkl
/results	Funciones que permite analizar las principales características.

main.ipybn	Programa principal que contiene la lógica de ejecución y llamada a las funciones
README.md	Documentación general del proyecto

Nota: Se importa la estructura del proyecto desde el repositorio de github.

Descarga Y Carga De Conjuntos De Datos Desde Kaggle. Se importan las bibliotecas necesarias, tales como sys, pandas, numpy, kagglehub, shutil, os y joblib. Además, se importan funciones personalizadas desde los directorios src (incluyendo load_datanew, split_dataset, split_data, balanceo_data) y models (como randomforest, svm, xgboost, rn, cnn y mlp), así como shap_graph desde el directorio results.

Se utiliza la función kagglehub.dataset_download para descargar cuatro conjuntos de datos de audio emocional disponibles en Kaggle: RAVDESS, TESS, CREMA-D y SAVEE. Se definen las rutas locales correspondientes a los directorios que contienen los archivos de audio de cada dataset descargado.

Con fines didácticos y educativos se pueden encontrar audios grabados de forma profesional y que se encuentran debidamente clasificados de acuerdo con la emoción que representan. Entre los conjuntos de datos encontrados tenemos:

“RAVDESS contiene 1440 archivos: 60 ensayos por actor x 24 actores = 1440. El RAVDESS contiene 24 actores profesionales (12 mujeres, 12 hombres) que vocalizan dos frases con la misma léxica con acento norteamericano neutro. Las emociones del habla incluyen calma, alegría, tristeza, enojo, miedo, sorpresa y disgusto. Cada expresión se produce en dos niveles de intensidad emocional (normal, fuerte), con una expresión neutra adicional.” (Livingstone, 2018)

CREMA-D es un conjunto de datos de 7442 clips originales de 91 actores. Estos clips pertenecían a 48 actores y 43 actrices de entre 20 y 74 años, provenientes de diversas razas y etnias (afroamericanos,

asiáticos, caucásicos, hispanos y no especificados). Los actores hablaron a partir de una selección de 12 frases. Las frases se presentaron utilizando una de seis emociones diferentes (ira, disgusto, miedo, felicidad, neutral y tristeza) y cuatro niveles de emoción diferentes (bajo, medio, alto y no especificado). (Lok, 2019)

TESS es un conjunto de datos donde dos actrices (de 26 y 64 años) pronunciaron un conjunto de 200 palabras objetivo en la frase portadora "Say the word _" y se grabaron las palabras representando cada una de siete emociones (ira, asco, miedo, felicidad, sorpresa agradable, tristeza y neutralidad). Hay 2800 puntos de datos (archivos de audio) en total. El conjunto de datos está organizado de tal manera que cada una de las dos actrices y sus emociones se encuentran en su propia carpeta. Dentro de esta, se puede encontrar el archivo de audio con las 200 palabras objetivo. El formato del archivo de audio es WAV. (Lok, 2019)

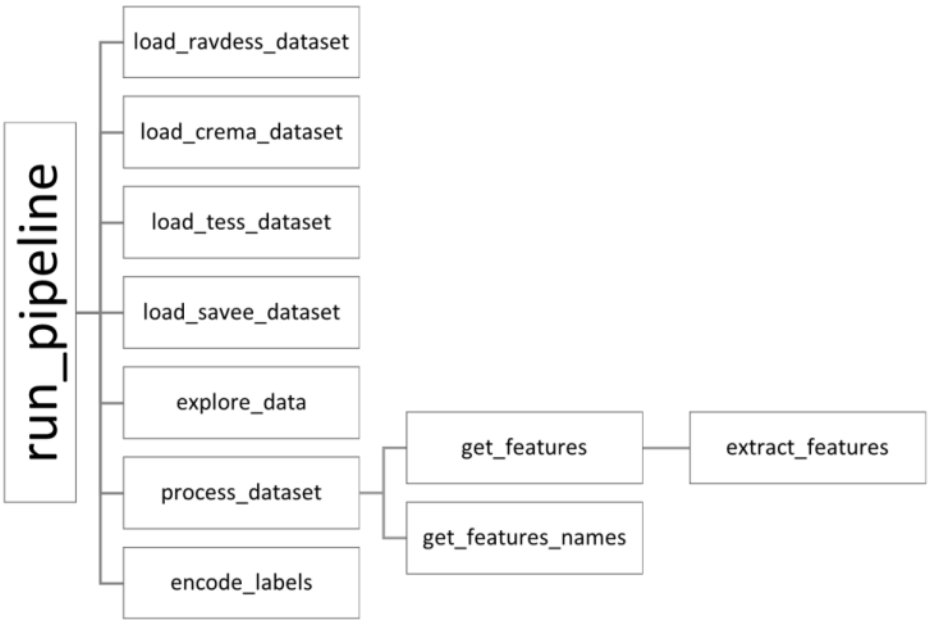
La base de datos SAVEE se registró a partir de cuatro hablantes nativos de inglés (identificados como DC, JE, JK, KL), estudiantes de posgrado e investigadores de la Universidad de Surrey, de entre 27 y 31 años. Las emociones se describieron psicológicamente en categorías discretas: ira, asco, miedo, felicidad, tristeza y sorpresa. También se añadió una categoría neutra para obtener grabaciones de 7 categorías de emociones. El material textual consistió en 15 oraciones TIMIT por emoción: 3 comunes, 2 específicas de la emoción y 10 genéricas, diferentes para cada emoción y fonéticamente equilibradas. Los 3 comunes y 2 × 6 = 12 específicas de la emoción se registraron como neutras, obteniendo 30 oraciones neutras. Esto resultó en un total de 120 enunciados por hablante. (Lok, 2019)

Ejecución del pipeline de procesamiento de datos. La función un_pipeline (importada desde src.load_datanew) es la que realiza todo el procesamiento del set de datos, es decir; carga los audios de los datasets especificados, extrae características relevantes (como coeficientes MFCC, ZCR, Chroma, RMS

y Mel), crea la data sintética, asigna los nombres a los features y combina los datos en un conjunto de características (X) junto con sus etiquetas correspondientes (Y).

Figura 2

Estructura de procesamiento del set de datos a través de la función run_pipeline

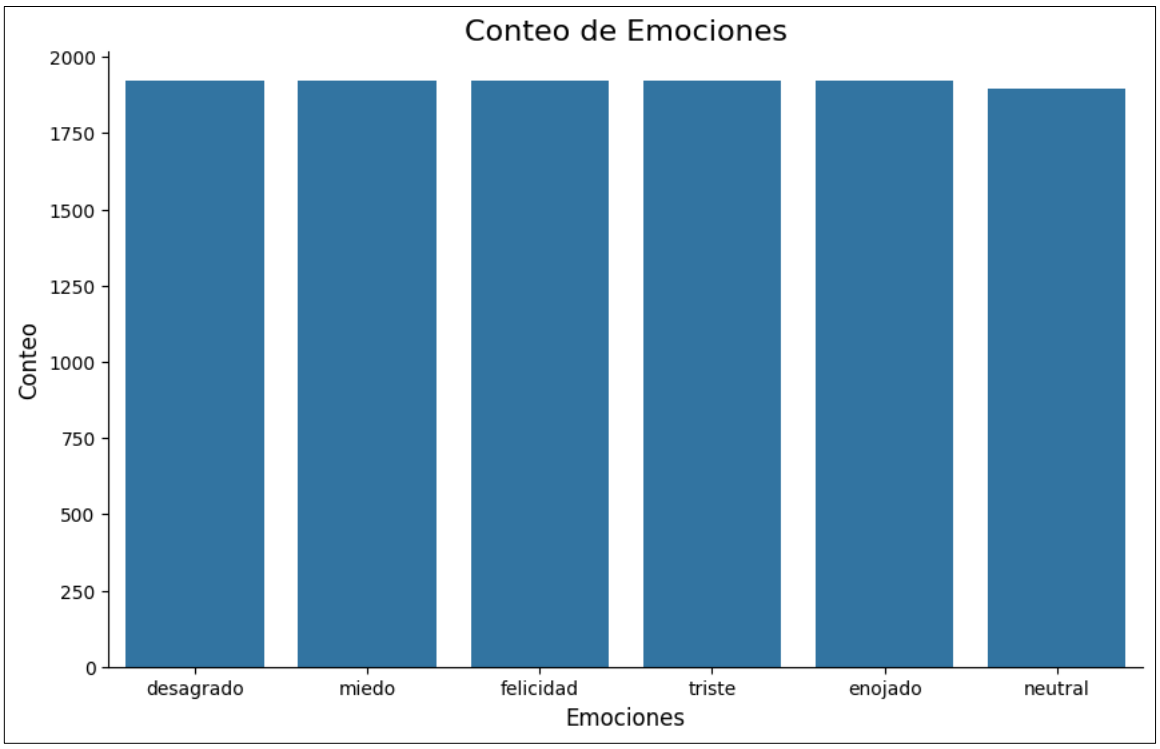


Se imprimen las dimensiones de estos conjuntos para verificar la correcta extracción, y se visualizan las primeras filas mediante DataFrames de Pandas.

Adicionalmente, se realiza un análisis exploratorio de los datos, en los cuales se puede visualizar que el problema de clasificación tiene seis categorías: desagrado, miedo, felicidad, triste, enojado, neutral y que el conjunto de datos se encuentra balanceado.

Figura 3

Distribución del data set



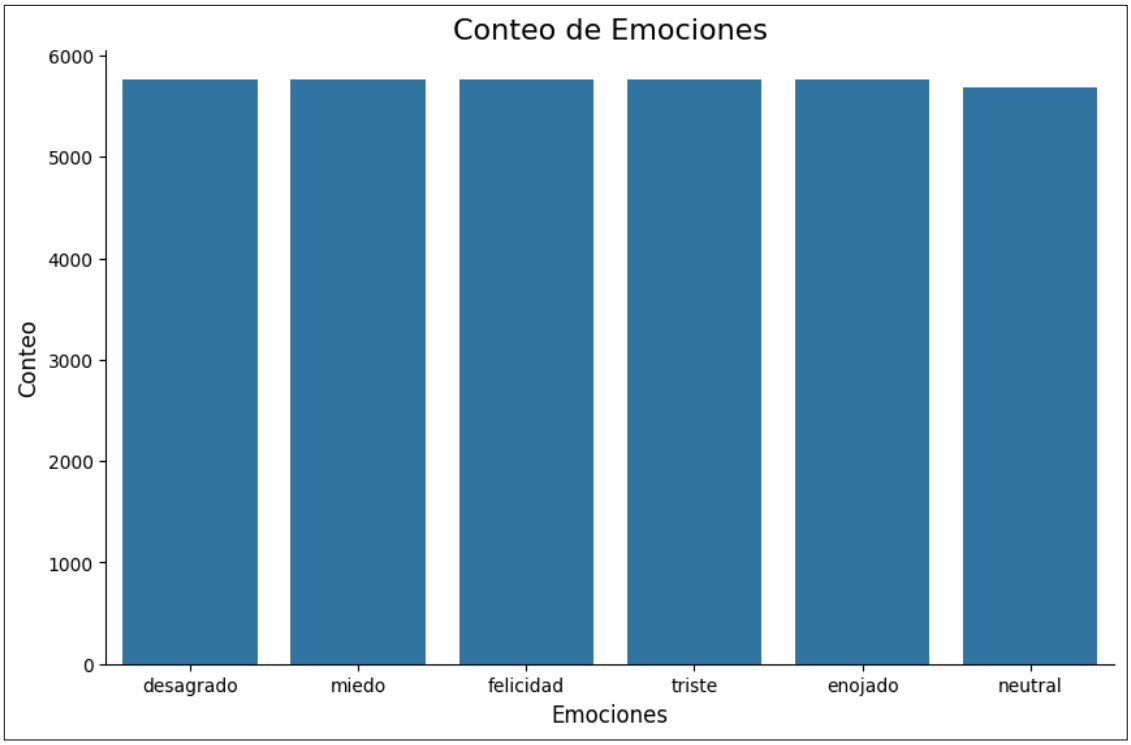
Aumento De Datos En Señales De Audio. Se realiza el aumento de datos para incrementar la diversidad del conjunto de entrenamiento mediante la generación de nuevas muestras sintéticas a partir de los datos existentes. Para este caso del procesamiento de señales de audio, se implementa mediante la introducción de perturbaciones controladas, como la inyección de ruido, el desplazamiento temporal, la modificación del tono (pitch shifting) y el ajuste de la velocidad de reproducción. (Ko, 2015).

El propósito de estas transformaciones es mejorar la capacidad de generalización del modelo, haciéndolo más robusto frente a variaciones comunes en los datos de entrada. Para preservar la validez del proceso de aprendizaje supervisado, es fundamental que las modificaciones aplicadas no alteren la

etiqueta original de la muestra, de modo que las instancias generadas conserven la misma clase emocional que el audio original. (Courville, 2016).

Figura 4

Total de emociones luego del aumento de datos



Al ejecutar el aumento de datos se obtiene un total de 34534 audios lo que significa que cada audio se triplico.

En la siguiente tabla se muestran las características que se extraen y que forman parte de las variables utilizadas para el estudio. En total se extrajeron 364 features de los audios con el objetivo de analizar dichas características en el entrenamiento de los modelos.

Tabla 4

Características o features extraídos

Característica	Definición
Zero-Crossing Rate (ZCR).	Extrae 2 características: media y desviación estándar de esa única banda, total 2 características Alto en emociones con energía/agitación (ira, miedo, alegría) Bajo en emociones más planas o lentas (tristeza, calma)
La Transformada de Fourier de Corto Plazo (STFT)	Extrae 12 medias y 12 desviaciones estándar, total 24 características Variación alta en emociones expresivas (alegría, ira) Variación baja en emociones planas (tristeza)
Coeficientes Cepstrales en las Frecuencias de Mel (MFCC)	Extrae 40 medias y 40 desviaciones estándar, total 80 características MFCC 0, más alto en ira, alegría; más bajo en tristeza MFCC 1–2, alegría = más agudos; tristeza = más graves MFCC 3–6, cambios fuertes en alegría, miedo, ira MFCC 7–13, aumentan en sorpresa, miedo, ira

	MFCC 14–40, más activos en emociones con expresión intensa (ira, miedo)
RMS (Root Mean Square Energy)	Extrae 2 características: media y desviación estándar, total 2 características Alta en emociones intensas (ira, alegría) Baja en tristeza
Espectrograma de mel.	Extraes 128 medias y 128 desviaciones estándar, total 256 características Banda 1–20, 0–400 Hz, Tono base (pitch), tristeza Banda 20–60, 400–2,000 Hz, Voz nasal, tensión, miedo Banda 60–128, 2,000–~11,000 Hz, Fricción vocal, sorpresa

Según una síntesis generada por ChatGPT (OpenAI, 2025), las características acústicas como ZCR, RMS, MFCC, STFT, etc. muestran patrones distintos según la emoción expresada y en resumen las posibles correlaciones entre las características extraídas y las emociones son las siguientes:

Tabla 5

Características vs emociones

Característica	Ira	Miedo	Felicidad	Tristeza	Desagrado	Neutral
ZCR_mean	Alta	Alta	Alta	Baja	Media	Media
ZCR_std	Alta	Alta	Alta	Baja	Media	Baja
RMS_mean	Alta	Media	Alta	Baja	Media	Media
RMS_std	Alta	Alta	Alta	Baja	Media	Baja
MFCC (1–13)	Muy activos	Detallados, agudos	Bien definidos	Planos, atenuados	Timbre apagado	Estables
Mel (bajas / altas)	Alta energía en agudos	Pico en medios-altos	Energía en medios-altos	Predominio en graves	Concentrada en medios	Distribución balanceada
Chroma_mean	Alta	Media	Alta	Baja	Media	Media
Chroma_std	Alta	Alta	Alta	Baja	Media	Baja

Descripción vocal	Voz proyectada, agresiva, con fluctuaciones fuertes	Voz aguda, temblorosa, entrecortada	Voz brillante, expresiva y enérgica	Voz suave, monótona, sin cambios significativos	Voz tensa, con tono plano, menos expresiva que ira o miedo	Voz neutra, clara, sin intención emocional marcada

Balanceo De Clases. De manera inicial existía un desbalanceo en la data cargada, especialmente en la categoría sorprendida. Se utilizó SMOTE (Synthetic Minority Over-sampling Technique) es una técnica empleada en aprendizaje automático, especialmente en clasificación, para tratar el problema del desbalance de clases en datasets, sin embargo; los resultados que se obtuvieron fue una disminución de las métricas en los diferentes modelos.

Finalmente, para el balanceo de clases, se unificaron las emociones neutral y calma en una sola categoría denominada neutral. Además, la emoción sorprendida fue excluida del análisis debido a la baja cantidad de registros disponibles. En consecuencia, el análisis de emociones por audio se centró en las siguientes categorías: desagrado, enojado, felicidad, miedo, neutral y triste.

División Del Conjunto De Datos. La función prepare_datasets2 (importada desde src.split_dataset) segmenta el conjunto total (X, Y) en subconjuntos de entrenamiento (x_train, y_train), validación (x_val, y_val) y prueba (x_test, y_test). Estos subconjuntos se guardan en archivos .pkl. Posteriormente, se llama a check_train("train.pkl") para inspeccionar el conjunto de entrenamiento y confirmar la posible existencia de desbalance en las clases.

Una vez que se analizó el set de datos y se realizó todo el procesamiento y división, como una recomendación de mejores prácticas se acogió el almacenamiento de dichos datos en archivos .pkl (Pickle), que tiene como ventajas la velocidad y eficiencia pues guarda los datos en un formato binario que es rápido de leer y escribir en comparación con archivos de texto o CSV. Esto acelera mucho la carga y guardado de datos durante el desarrollo y la experimentación.

Otra ventaja es la preservación de estructuras complejas, los archivos .pkl pueden almacenar casi cualquier objeto de Python, no solo datos tabulares planos. Esto incluye listas, diccionarios, arreglos de NumPy, DataFrames de pandas, y estructuras más complejas que pueden ser difíciles de guardar o reconstruir con otros formatos.

Si requerimos utilizar el mismo set de datos en diferentes entrenamientos, pueden ser reutilizados porque permite cargarlos directamente sin tener que hacer el proceso de división cada vez, garantizando reproducibilidad y ahorro de tiempo.

Configuración De Git Y Carga De Archivos Al Repositorio. Con el objetivo de almacenar los datos de train, test y validación, se realiza la integración con github.

Se configura la identidad del usuario en Git (user.name y user.email), y se establecen variables para el token, usuario y repositorio de GitHub (se advierte que incluir tokens directamente en código público representa un riesgo de seguridad). El directorio de trabajo cambia al repositorio local clonado en Colab (/content/ProyectoFinalG5). Se configura la URL remota para autenticar con el token, se realiza un git pull para sincronizar cambios remotos (permitiendo historias no relacionadas con --allow-unrelated-histories), se agregan archivos específicos (src/train.pkl, src/val.pkl, src/test.pkl, src/class_labels.npy) al área de preparación (staging), se crea un commit con mensaje descriptivo y se suben los cambios con git push origin main --force (el uso de --force sobrescribe el historial remoto y debe manejarse con precaución).

Carga Y Verificación De Datos De Entrenamiento. Se carga el conjunto de entrenamiento desde src/train.pkl mediante joblib.load y se imprime la dimensión del conjunto de etiquetas (y_train) para confirmar su correcta preparación y balanceo.

Entrenamiento De Modelos De Aprendizaje Automático. Se ejecutan diversas funciones de entrenamiento de modelos importados desde el directorio models:

- run_random_forest(): entrena y evalúa un modelo Random Forest.
- run_svm(): entrena y evalúa un modelo Support Vector Machine (SVM).
- run_xgboost(): entrena y evalúa un modelo XGBoost.
- run_rn(): entrena y evalúa un modelo de una Red Neuronal Simple
- ejecutar_modelo_cnn(): entrena y evalúa un modelo de múltiples capas convolucionales con regularización L2 para mitigar el sobreajuste.
- run_mlp(): entrena y evalúa una red neuronal profunda del tipo perceptrón multicapa (MLP).
- modelo_cnn_1D (): entrena y evalúa un modelo CNN de una dimensión.
- modelo_cnn_1DL2 (): entrena y evalúa un modelo CNN de una dimensión con regularización L2.
- entrenar_modelo_cnn_lstm(): entrena y evalúa un modelo combinado CNN-LSTM
- entrenar_modelo_cnn_lstmL2(): entrena y evalúa un modelo combinado CNN-LSTM con regularización L2.

Análisis De Importancia De Características Mediante SHAP. Como se mencionó anteriormente, para explicar por qué se tomaron ciertas decisiones o cómo se llegaron a determinadas predicciones, es

importante la interpretación de los resultados entendiendo cómo las características o features aportan en los resultados del modelo, para ello se utilizó el método SHAP.

Se carga el modelo seleccionado como óptimo usando joblib.load. Luego, se llama a la función generate_shap_outputs (importada desde results.shap_graph), la cual recibe un modelo, un subconjunto de prueba y los nombres de características para generar explicaciones SHAP que visualizan la contribución de cada variable a las predicciones. Los resultados se exportan en formatos gráficos y/o CSV.

Marco Teórico

Análisis de Hiperparámetros de los Modelos de Aprendizaje

Los hiperparámetros son configuraciones importantes que se definen antes del entrenamiento del modelo y afectan notablemente su rendimiento. Al contrario de los parámetros que se ajustan durante el entrenamiento, la selección apropiada de hiperparámetros, como la tasa de aprendizaje y el número de capas, puede definir si un modelo se entrena bien o sufre de sobreajuste.

El análisis de hiperparámetros involucra a parte de la identificación de los más relevantes, también el mejoramiento de su combinación para incrementar el rendimiento del modelo. A continuación, se incluye la descripción de los hiperparámetros de los modelos que se realizó en este proyecto:

Modelo Random Forest:

Se implementa un flujo completo para entrenar y evaluar un modelo de clasificación usando la función run_random_forest. Primero, se gestiona los datos divididos en conjuntos de entrenamiento, validación y prueba, decodificando las etiquetas one-hot a formato numérico.

Luego, realiza una búsqueda exhaustiva de hiperparámetros mediante GridSearchCV con validación cruzada para optimizar el modelo.

Una vez seleccionado el mejor modelo, lo entrena con los datos combinados de entrenamiento y validación, y lo evalúa en el conjunto de prueba independiente usando métricas como accuracy, precision, recall y F1-score. Finalmente, el modelo entrenado se guarda en disco usando joblib para su posterior reutilización.

Este pipeline garantiza un entrenamiento robusto, evitando el sobreajuste y proporcionando una estimación imparcial del rendimiento en datos no vistos.

- n_estimators (list): Número de árboles en el bosque. Opciones: [100, 200].
- max_depth (list): Profundidad máxima del árbol. Opciones: [10, 20, None].
- min_samples_split (list): Número mínimo de muestras requeridas para dividir un nodo interno.
- min_samples_leaf (list): Número mínimo de muestras requeridas para estar en un nodo hoja.
- class_weight (list): Ponderación asociada a las clases. Opciones: ['balanced'].

Modelo Máquinas de Vectores de Soporte (SVM):

Se aplica un pipeline para entrenar y optimizar un modelo de clasificación basado en la función run_svm . Comienza cargando los conjuntos de datos de entrenamiento, validación y prueba, aplanando las entradas si es necesario y decodificando etiquetas de formato one-hot a índices de clase. Combina los datos de entrenamiento y validación para realizar una búsqueda de hiperparámetros con GridSearchCV, probando distintas combinaciones de C, kernel y gamma mediante validación cruzada de 3 pliegues y usando precisión como métrica. Tras encontrar la mejor configuración, entrena el modelo final, lo guarda con joblib y lo evalúa en el conjunto de prueba independiente, generando métricas como

precisión, recall, F1-score y matriz de confusión para estimar su rendimiento real. Las SVM permiten separar clases mediante hiperplanos óptimos y, con kernels, manejar datos no linealmente separables.

- C: Un valor bajo crea un margen más amplio, pero permite más errores de clasificación (modelo más regularizado). Un valor alto de C crea un margen más estrecho, pero intenta clasificar todos los puntos correctamente (menos regularizado, propenso a overfitting).
- kernel: Define la función de transformación utilizada para mapear los datos a un espacio de mayor dimensión.
- gamma: Define la influencia de un solo ejemplo de entrenamiento. Valores bajos significan una influencia lejana, mientras que valores altos significan una influencia cercana. scale utiliza 1/(n_features* X.var()) como valor, lo que es a menudo una buena opción predeterminada.
- Validación Cruzada (cv=3): Divide el conjunto de entrenamiento (en este caso, combinado x_combined y y_combined) en K pliegues (3 en este caso). El modelo se entrena K veces, usando K-1 pliegues para el entrenamiento y el pliegue restante para la validación. Esto proporciona una estimación más robusta del rendimiento del modelo en datos no vistos que una simple división train/validation.

Modelo Xgboost.

Se utiliza la función run_xgboost describe un flujo de trabajo para entrenar un modelo de clasificación con XGBoost, un algoritmo eficiente de gradient boosting que usa árboles de decisión como predictores bases. Se realiza ajuste de hiperparámetros con RandomizedSearchCV, buscando combinaciones óptimas de parámetros como n_estimators, max_depth y learning_rate, usando validación cruzada de 3 pliegues para evaluar cada configuración. Tras encontrar el mejor modelo, se

entrena con los datos combinados de entrenamiento y validación y se evalúa en un conjunto de prueba independiente, calculando métricas como precisión y pérdida logarítmica multiclase (mlogloss). El flujo incluye pasos de preprocesamiento (carga de datos, decodificación de etiquetas one-hot) y guarda el modelo entrenado con joblib para reutilizarlo posteriormente.

- n_estimators: Este hiperparámetro controla el número de árboles de refuerzo (boosted trees) en el modelo. Más árboles generalmente mejoran el rendimiento, pero también pueden llevar a un sobreajuste y a un mayor tiempo de entrenamiento. En el grid de búsqueda, se exploran los valores [100, 200].
- max_depth: Este hiperparámetro define la profundidad máxima de cada árbol individual en el modelo. Una profundidad mayor permite que los árboles capturen interacciones más complejas entre las características, pero también aumenta el riesgo de sobreajuste. Los valores explorados son [6, 10].
- learning_rate (también conocido como eta en la documentación original de XGBoost): Este hiperparámetro controla el tamaño del paso en cada iteración de refuerzo. Un valor más pequeño requiere más estimadores para alcanzar el mismo nivel de rendimiento, pero puede resultar en un modelo más robusto y menos propenso al sobreajuste. Los valores explorados son [0.1, 0.01].
- use_label_encoder=False: Desactiva el uso del LabelEncoder interno de XGBoost, lo cual es una práctica recomendada y elimina una advertencia de depreciación.
- eval_metric=mlogloss: Especifica la métrica de evaluación utilizada durante el entrenamiento para problemas de clasificación multi-clase. mlogloss (logloss multi-clase) es una métrica común para este tipo de problemas.
- estimator: El modelo XGBClassifier a optimizar.

- param_distributions: El diccionario que define los hiperparámetros y los valores o distribuciones a muestrear.
- n_iter=4: El número de combinaciones de hiperparámetros a muestrear aleatoriamente. En este caso, solo se prueban 4 combinaciones de las posibles.
- scoring=accuracy: La métrica utilizada para evaluar el rendimiento de cada combinación de hiperparámetros. Se busca maximizar la precisión (accuracy).
- cv=3: El número de pliegues (folds) para la validación cruzada durante la búsqueda. El conjunto combinado de entrenamiento y validación se divide en 3 partes, y el modelo se entrena en 2 partes y se evalúa en la parte restante, rotando los pliegues.
- n_jobs=-1: Utiliza todos los núcleos de CPU disponibles para acelerar la búsqueda.
- verbose=1: Muestra el progreso de la búsqueda.
- random_state=0: Fija la semilla aleatoria para la reproducibilidad de la búsqueda aleatoria.

Modelo Red Neuronal Simple

La arquitectura planteada para este modelo consta de una red que tiene las siguientes capas:

- *Capa de entrada (Input)***:** Forma de entrada: (x_train.shape[1], 1); lo que significa que Cada muestra de entrada es un vector 1D con x_train.shape[1] características.
- *Capa de aplanamiento (Flatten)***:** Convierte la entrada 2D en un vector 1D, eliminando la estructura adicional introducida por la dimensión (..., 1).
- *Capa densa***:** (Dense(256, activation='relu')); Capa completamente conectada con 256 neuronas y función de activación ReLU, aprende combinaciones no lineales de las características de entrada.

- *Capa de regularización*: (Dropout(0.5)); desactiva aleatoriamente el 50 % de las neuronas durante cada época de entrenamiento para prevenir el sobreajuste.
- *Capa de salida:* (Dense(6, activation='softmax')); capa con 6 neuronas (una por cada clase de emoción). utiliza softmax para producir una distribución de probabilidad sobre las clases.
- *Los parámetros de compilación*
- *Optimizador: Adam*: Algoritmo de optimización eficiente basado en gradientes, que ajusta los pesos utilizando momentos de primer y segundo orden (Diederik P. Kingma, 2015)
- *Función de pérdida: categorical_crossentropy:* Se utiliza para problemas de clasificación multiclase cuando las etiquetas están en formato one-hot. Calcula la diferencia entre las distribuciones de probabilidad predicha y verdadera.
- *Métrica*: *accuracy*; Indica el porcentaje de predicciones correctas sobre el total de muestras.

Los parámetros de entrenamiento

- *Épocas* = 20: El conjunto de entrenamiento será recorrido completamente 20 veces.
- *Tamaño de lote* = 32: Los datos se dividen en lotes de 32 muestras antes de actualizar los pesos. Esto mejora la eficiencia y estabiliza el gradiente.
- *Conjunto de validación* = (x_test, y_test): Se evalúa el desempeño del modelo en cada época sobre un conjunto de validación independiente.

- *Callback: ReduceLROnPlateau()* Reduce la tasa de aprendizaje si la métrica de validación (p. ej., val_loss) no mejora después de cierto número de épocas, lo que ayuda a afinar el aprendizaje en fases más lentas.

El accuracy obtenido para este modelo corresponde al 65%

Modelo CNN

La arquitectura propuesta para este modelo se compone de varias capas convolucionales y densas con regularización y normalización, que se describen a continuación:

- *Capa de entrada (Input(shape=input_shape))*: La forma de entrada corresponde a una señal 1D multicanal que son las características extraídas de audio.
- *Primera capa convolucional (Conv1D***):** Filtros: 128, Tamaño del kernel: 5, Activación: ReLU; Regularización L2 con λ = 0.001: penaliza pesos grandes para reducir el sobreajuste; padding='same': la salida mantiene la misma longitud que la entrada.
- *Normalización por lotes (BatchNormalization)*: Estabiliza y acelera el entrenamiento al normalizar las activaciones.
- *Reducción de dimensionalidad (MaxPooling1D)*: Reduce la longitud de la secuencia a la mitad (pool_size=2).
- *Regularización (Dropout(0.3)):* Desactiva aleatoriamente el 30% de las neuronas durante el entrenamiento.
- *Segunda y tercera capas convolucionales***:** Segunda: 64 filtros, kernel 5. Tercera: 32 filtros, kernel 3. Ambas siguen el mismo patrón: Conv1D → BatchNorm → MaxPooling (excepto la última, que no tiene pooling) → Dropout.

- *Aplanamiento (Flatten):* Convierte la salida 3D en un vector para conectar con las capas densas.
- *Capa densa (Dense(64, activation='relu')):* 64 neuronas completamente conectadas con regularización L2.
- *Capa de salida (Dense(6, activation='softmax')):*6 neuronas de salida (correspondientes a las clases de emoción), con activación softmax para clasificación multiclase.

Parámetros de compilación

- *Optimizador: Adam:* Combina las ventajas de AdaGrad y RMSProp. La tasa de aprendizaje es 1e-3, un valor estándar.
- *Función de pérdida*: *categorical_crossentropy:* Apropiada para clasificación multiclase con etiquetas one-hot.
- *Métrica: accuracy:* Mide el porcentaje de predicciones correctas sobre el total.
- *Parámetros de entrenamiento*
- *Época***s**: *100*: Máximo de pasadas sobre todo el conjunto de entrenamiento.
- *Tamaño del lote*: *32* Se actualizan los pesos cada 32 muestras.
- *Conjunto de validación: (x_test, y_test)*: Evalúa el rendimiento del modelo en datos no vistos.
- *EarlyStopping***:** Detiene el entrenamiento si val_loss no mejora después de 10 épocas. Evita el sobreajuste.
- *ReduceLROnPlateau***:** Reduce la tasa de aprendizaje a la mitad si val_loss no mejora tras 5 épocas. Ayuda a refinar la convergencia.

El accuracy obtenido para este modelo corresponde al 75,5% de aciertos sobre el conjunto de datos de validación.

Modelo Multilayer Perceptron

La arquitectura propuesta para este modelo tiene la siguiente estructura:

- *Capa de entrada (Input(shape=(x_train.shape[1],))):* La red recibe un vector de características de longitud x_train.shape[1].
- *Primera capa densa (Dense(512, activation='relu'))* 512 neuronas con activación ReLU, útil para modelar relaciones no lineales.
- *Dropout (0.3)* Se apaga el 30 % de las neuronas durante el entrenamiento para prevenir el sobreajuste.
- *Segunda capa densa (Dense(256, activation='relu'))* 256 neuronas con ReLU.
- *Dropout (0.2)*Desactiva el 20 % de las neuronas, complementando la regularización.
- *Tercera capa densa (Dense(128, activation='relu'))*Otra capa intermedia de reducción de dimensionalidad progresiva.
- *Capa de salida (Dense(n_classes, activation='softmax'))* n_classes = 6 neuronas para clasificación multiclase que corresponde a cada una de las emociones evaluadas. La función softmax asigna probabilidades a cada clase.

Parámetros de compilación

- *Optimizador: Adam* Basado en momentos adaptativos; eficaz para entrenamientos rápidos y estables. Se especifica una tasa de aprendizaje baja (0.0005), que favorece la convergencia fina.

- *Función de pérdida: CategoricalFocalLoss(gamma=2.0)* Variante de la entropía cruzada, propuesta por Lin et al. (2017), que penaliza más los errores en clases difíciles y menos los ejemplos bien clasificados, ayudando en problemas de clases desbalanceadas. *γ=2.0* amplifica esta penalización.
- *Métrica: accuracy* Evalúa la proporción de predicciones correctas.

Parámetros de entrenamiento

- *Épocas: 100 (máximo)*
- *Tamaño del lote: 32*
- *Validación:* Se usa un conjunto separado (x_val, y_val)
- *Verbose=2:* Proporciona un resumen por época.
- *EarlyStopping:* Detiene el entrenamiento si la función de pérdida de validación (val_loss) no mejora después de 10 épocas, restaurando los mejores pesos.

El accuracy obtenido para este modelo corresponde al 76.1% de aciertos sobre el conjunto de datos de validación.

Modelo CNN 1D

La arquitectura propuesta para el modelo tiene la siguiente estructura:

- *Épocas: 50,* el número de épocas indica cuántas veces el modelo pasará por el conjunto de datos de entrenamiento. En este caso, se han configurado 50 épocas, lo que permite al modelo aprender de los datos sin riesgo significativo de sobreajuste.
- *Capa de Entrada:* (x_train.shape[1], 1). Esta capa obtiene las características del audio en una dimensión.

- *Capas Convolucionales:* Hay tres capas definicdas Conv1D con 32, 64 y 128 filtros con un tamaño de kernel 3, activación ReLU.
- *Capas de MaxPooling:* Estas tres capas tiene la función de reducir la dimensionalidad
- *Capa de Aplanamiento:* Convierte la salida de las capas convolucionales en un vector.
- *Capa Densa:* Con 128 neuronas y activación ReLU.
- *Capa de Dropout:* Con una tasa de 0.6 para prevenir el sobreajuste.
- *Capa de Salida:* Densa con un número de neuronas igual al número de clases de emociones y activación softmax.
- *Tamaño de Lote (Batch Size):* El tamaño de lote de 32, es decir, que el modelo actualiza sus pesos después de ejecutar 32 muestras. Este valor se escogió para equilibrar la velocidad de entrenamiento del modelo.

Modelo CNN 1D L2

- *Épocas: 50.* Igual que el modelo anterior se han configurado 50 épocas, lo que permite al modelo aprender de los datos sin riesgo significativo de sobreajuste.
- *Capa de Entrada:* (x_train.shape[1], 1)**.** Esta capa obtiene las características del audio en una dimensión.
- *Capas Convolucionales:* Hay tres capas definicdas Conv1D con 32, 64 y 128 filtros con un tamaño de kernel 3, activación ReLU y regularizacion de L2 con coeficiente 0.0001.
- *Capas de MaxPooling:* Estas tres capas tiene la función de reducir la dimensionalidad tienen un tamano de 2.
- *Capa de Aplanamiento:* Convierte la salida de las capas convolucionales en un vector.

- *Capa Densa:* Con 128 neuronas, activación ReLU y regularización L2 con coeficiente 0.001.
- *Capa de Dropout:* Con una tasa de 0.6 para prevenir el sobreajuste.
- *Capa de Salida:* Con 6 neuronas una para cada clase y activación softmax que convierte las salidas en probabilidades.
- *Tamaño de Lote* (Batch Size): El tamaño de lote de 32 quiere decir que el modelo actualiza sus pesos después de ejecutar 32 muestras. Este valor se escogió para equilibrar la velocidad de entrenamiento del modelo.

Para la Compilación del Modelo se trabajó con los siguientes hiperparámetros

- *Optimizador:* Adam
- *Tasa de Aprendizaje:* 0.0005. El valor de 0.0005 es bajo, es decir, que el modelo realizará pequeños cambios en los pesos durante cada iteración. Esto sirve para evitar grandes variaciones en la convergencia.
- *Función de Pérdida:* Categorical Crossentropy. Ayuda a tener la mejor precisión en la distribución de probabilidades.
- Callbacks Definidos: *Early Stopping:* Monitorea la precisión de validación, con paciencia de 5 épocas y recuperación de los mejores pesos.
- *Reducción de Tasa de Aprendizaje:* Supervisa la pérdida de validación, reduce la tasa en un factor de 0.2, con paciencia de 3 épocas y un mínimo de 0.00001.

Modelo CNN-LSTM

- *Épocas:* 50

- *Capa de Entrada:* (x_train.shape[1], 1)**.** Reshape se utiliza para ajustar la entrada a lo solicitado por la capa convolucional.
- *Capas Convolucionales:* Hay dos capas definicdas Conv1D con 512, y 256 filtros con un tamaño de kernel 5, activación ReLU, normalización BatchNormalization().
- *Capas de MaxPooling:* Estas capas tiene la función de reducir la dimensionalidad
- *Capa Densa:* Con 128 neuronas y activación ReLU.
- *Capa de Dropout:* Con una tasa de 0.3 para prevenir el sobreajuste.
- *Capas LTSM:* Hay tres capas de 200 unidades, dos de ellas con return_sequences=True y una con return_sequences=False.
- *Capa de Dropout:* Con una tasa de 0.5
- *Capa de Salida:* Densa con un número de neuronas igual al número de clases de emociones y activación softmax.
- *Tamaño de Lote* (Batch Size): 32

Compilación del Modelo:

- *Optimizador:* Adam
- *Tasa de Aprendizaje:* 0.0001
- *Función de Pérdida:* Categorical Crossentropy
- Callbacks. *ModelCheckpoint:* Guarda el mejor modelo apoyado en val_accuracy.
- *EarlyStopping:* Susoende el entrenamiento si no hay mejora en val_loss durante 10 épocas.
- *ReduceLROnPlateau:* Disminuye la tasa de aprendizaje si val_loss no mejora durante 5 épocas.

Modelo CNN-LSTM L2

- Épocas: 50
- *Capa de Entrada:* (x_train.shape[1], 1)**.** Reshape se utiliza para ajustar la entrada a lo solicitado por la capa convolucional.
- *Capas Convolucionales:* Hay dos capas definicdas Conv1D con 512, y 256 filtros con un tamaño de kernel 5, activación ReLU, normalización BatchNormalization() y regularización L2 con un factor de 0.00001.
- *Capas de MaxPooling:* Estas capas tiene la función de reducir la dimensionalidad
- *Capa Densa:* Con 128 neuronas y activación ReLU.
- *Capa de Dropout:* Con una tasa de 0.3 para prevenir el sobreajuste.
- *Capas LTSM:* Hay tres capas de 200 unidades, dos de ellas con return_sequences=True y una con return_sequences=False y regularización L2 con un factor de 0.00001.
- *Capa de Dropout:* Con una tasa de 0.5
- *Capa de Salida:* Densa con un número de neuronas igual al número de clases de y_traina, ctivación softmaxy y regularización L2 con un factor de 0.0001.
- *Tamaño de Lote* (Batch Size): 32

Compilación del Modelo:

- *Optimizador:* Adam
- *Tasa de Aprendizaje:* 0.001
- *Función de Pérdida***:** Categorical Crossentropy
- Callbacks. *ModelCheckpoint:* Guarda el mejor modelo apoyado en val_accuracy.

- *EarlyStopping:* Susoende el entrenamiento si no hay mejora en val_loss durante 10 épocas.
- *ReduceLROnPlateau:* Disminuye la tasa de aprendizaje si val_loss no mejora durante 5 épocas.

Interpretabilidad de Modelos - Características Principales

Con el objetivo de analizar e identificar las principales características de audio que influyen en la predicción de emociones, es decir; saber qué características obtenidas con MFCC, ZCR, RMS, etc., fueron más importantes para que el modelo clasificara una muestra como cierta emoción (ira, tristeza, etc.). se utilizó el método SHAP.

Se seleccionaron aleatoriamente 100 muestras del conjunto x_data del modelo identificado como mejor, para usarlas como conjunto de referencia background = shap.sample(x_data, 100, random_state=0) x_sample = x_data[:100]

Se utiliza adicionalmente KernelExplainer, que es una clase del paquete SHAP (SHapley Additive exPlanations) diseñada para explicar cualquier modelo de machine learning, incluso si es una "caja negra" como una red neuronal, SVM, o ensemble.

explainer = shap.KernelExplainer(model.predict_proba, background)

Finalmente, se calculan los valores SHAP para cada muestra en x_sample, usando el explainer en este caso, un KernelExplainer

shap_values = explainer.shap_values(x_sample)

CAPITULO 4

ANÁLISIS DE RESULTADOS

Pruebas de Concepto

Modelo Random Forest

Como prueba de concepto se valida el modelo generado random_forest_model con datos no vistos por el modelo y se presentan sus resultados.

Figura 5

Métricas modelo

```
📊 Métricas de evaluación:
✔ Accuracy: 0.6498
✔ Precision (weighted): 0.6627
✔ Recall (weighted): 0.6498
✔ F1-score (weighted): 0.6463
✔ Balanced Accuracy: 0.6501
```

Nota: Se observa que el valor del Accuracy es del 65,01% de aciertos con respecto al conjunto de datos de validación; datos que no han sido vistos por el modelo.

Figura 6

Métricas por clase

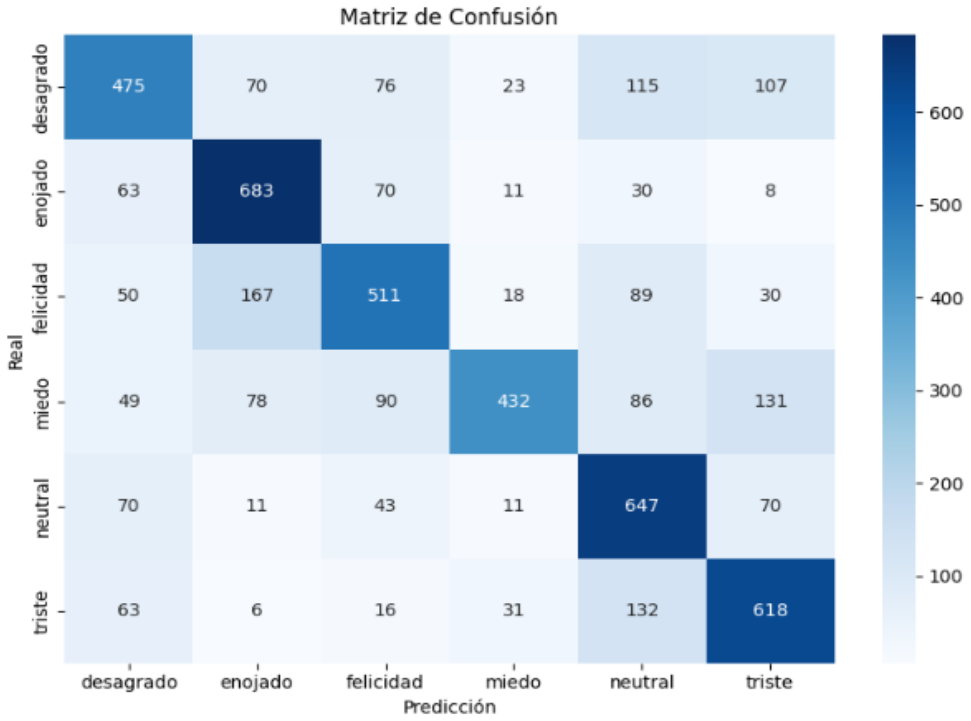
📋 Reporte de Clasificación por clase:

	precision	recall	f1-score	support
desagrado	0.62	0.55	0.58	866
enojado	0.67	0.79	0.73	865
felicidad	0.63	0.59	0.61	865
miedo	0.82	0.50	0.62	866
neutral	0.59	0.76	0.66	852
triste	0.64	0.71	0.68	866
accuracy			0.65	5180
macro avg	0.66	0.65	0.65	5180
weighted avg	0.66	0.65	0.65	5180

De manera similar a lo observado durante el entrenamiento y validación del modelo con el conjunto de prueba, en el nuevo conjunto de datos de validación la clase mejor detectada corresponde a enojado con un F1 del 73% lo que indica que detecta bien los casos reales, mientras que la clase con el menor rendimiento relativo es desagrado. Todas las clases están representadas en el conjunto de datos.

Figura 7

Matriz de confusión



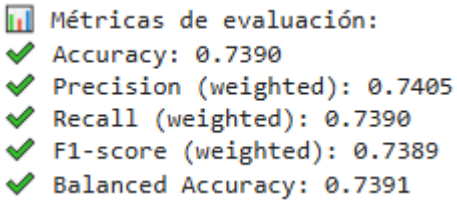
Nota: En la matriz de confusión se puede observar que la clase miedo se confunde con triste y la clase triste se confunde con neutral.

Modelo SVM (Support Vector Machine)

Como prueba de concepto se valida el modelo generado svm_best_model con datos no vistos por el modelo y se presentan sus resultados.

Figura 8

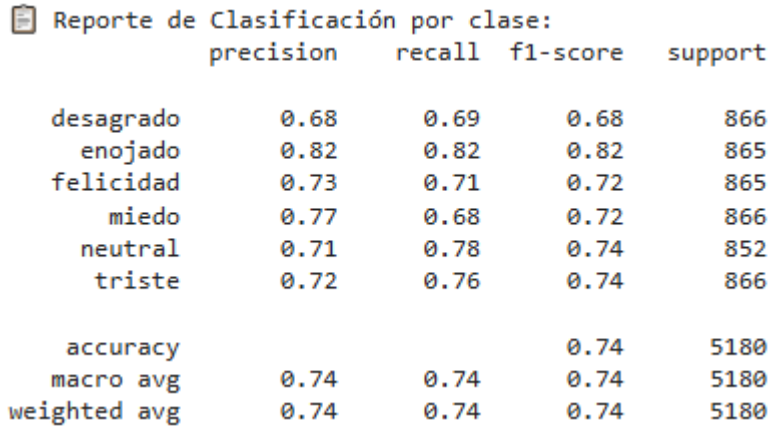
Métricas modelo svm_best_model.pkl



Nota: Se observa que el valor del Accuracy es del 73,91% de aciertos con respecto al conjunto de datos de validación; datos que no han sido vistos por el modelo.

Figura 9

Métricas por clase

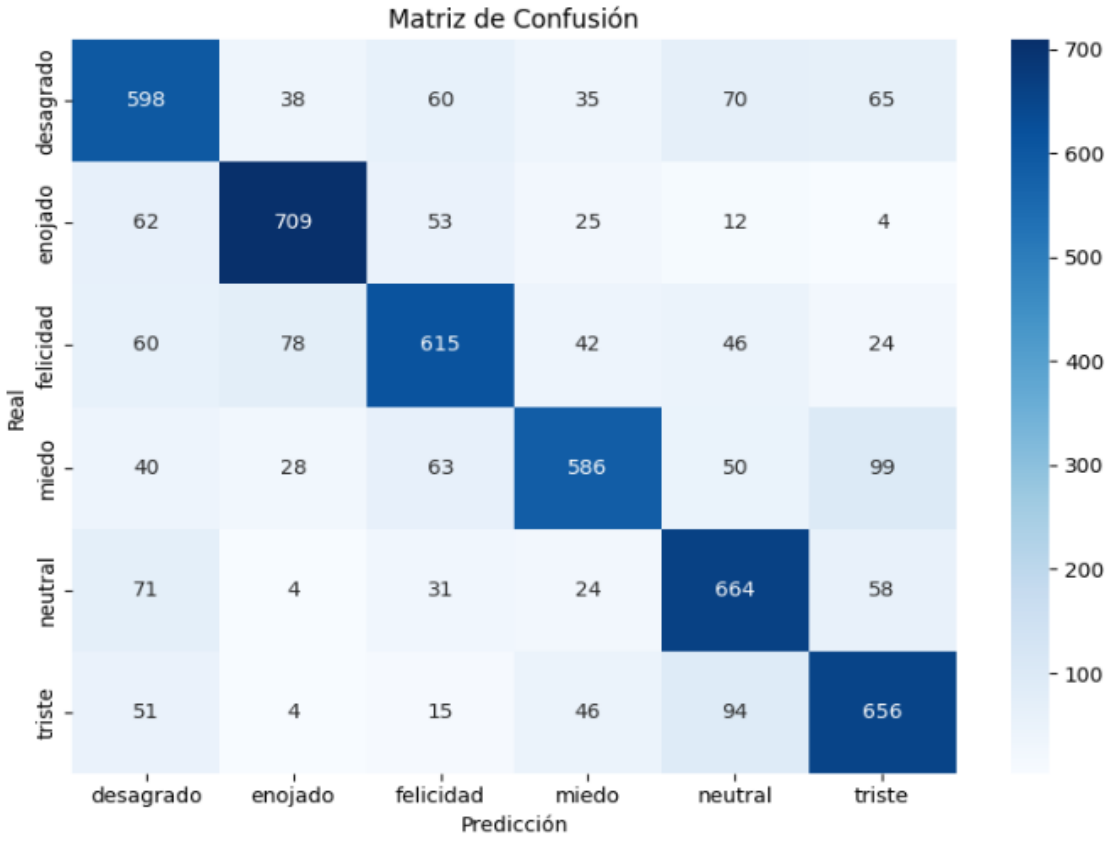


	precision	recall	f1-score	support
desagrado	0.68	0.69	0.68	866
enojado	0.82	0.82	0.82	865
felicidad	0.73	0.71	0.72	865
miedo	0.77	0.68	0.72	866
neutral	0.71	0.78	0.74	852
triste	0.72	0.76	0.74	866
accuracy			0.74	5180
macro avg	0.74	0.74	0.74	5180
weighted avg	0.74	0.74	0.74	5180

De manera similar a lo observado durante el entrenamiento y validación del modelo con el conjunto de prueba, en el nuevo conjunto de datos de validación la clase mejor detectada corresponde a enojado con un F1 del 82% lo que indica que detecta bien los casos reales, mientras que la clase con el menor rendimiento relativo es desagrado. Todas las clases están representadas en el conjunto de datos.

Figura 10

Matriz de confusión



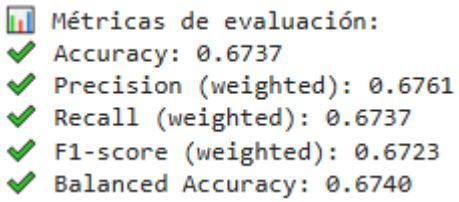
Nota: En la matriz de confusión se puede observar que la clase miedo se confunde con triste.

Modelo XGboost

Como prueba de concepto se valida el modelo generado xgboost_best con datos no vistos por el modelo y se presentan sus resultados.

Figura 11

Métricas modelo xgboost_best.pkl



Nota: Se observa que el valor del Accuracy es del 67,4% de aciertos con respecto al conjunto de datos de validación; datos que no han sido vistos por el modelo.

Figura 12

Métricas por clase

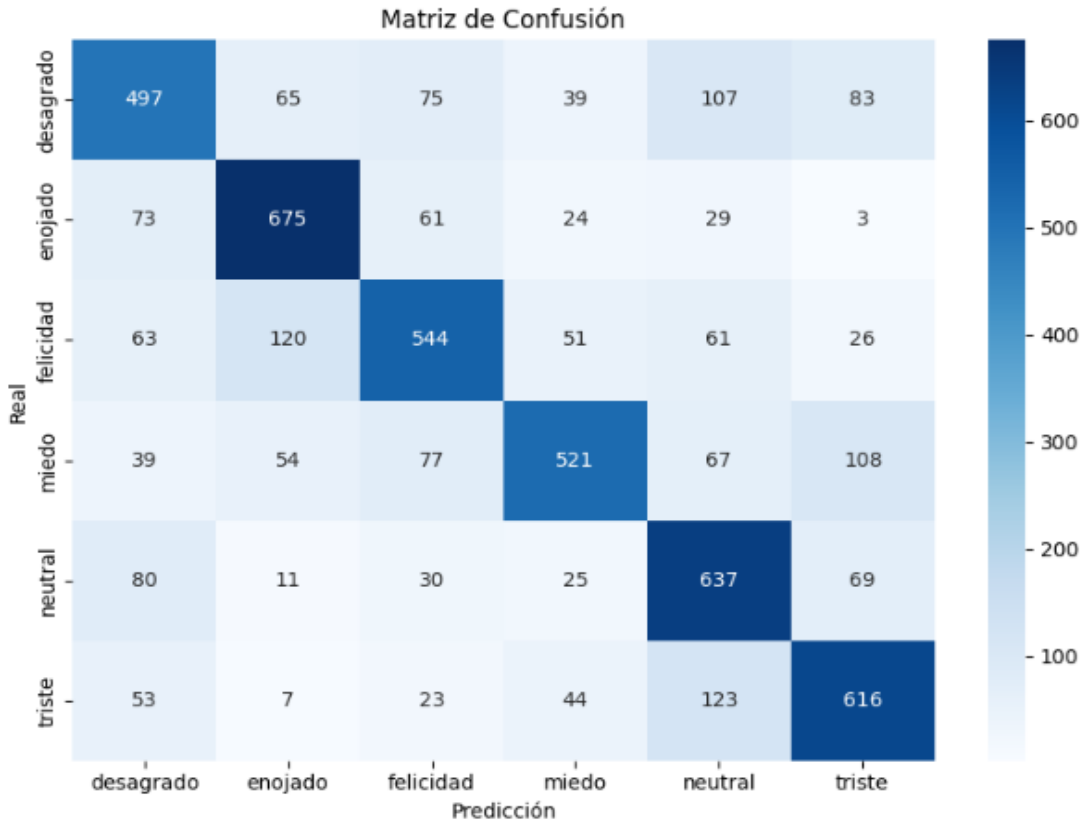
Reporte de Clasificación por clase:

	precision	recall	f1-score	support
desagrado	0.62	0.57	0.59	866
enojado	0.72	0.78	0.75	865
felicidad	0.67	0.63	0.65	865
miedo	0.74	0.60	0.66	866
neutral	0.62	0.75	0.68	852
triste	0.68	0.71	0.70	866
accuracy			0.67	5180
macro avg	0.68	0.67	0.67	5180
weighted avg	0.68	0.67	0.67	5180

De manera similar a lo observado durante el entrenamiento y validación del modelo con el conjunto de prueba, en el nuevo conjunto de datos de validación la clase mejor detectada corresponde a enojado con un F1 del 75% lo que indica que detecta bien los casos reales, mientras que la clase con el menor rendimiento relativo es desagrado. Todas las clases están representadas en el conjunto de datos.

Figura 13

Matriz de confusión



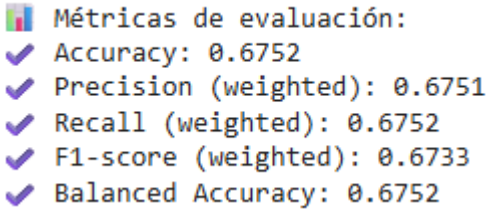
Nota: En la matriz de confusión se puede observar que la clase triste se confunde con neutral.

Modelo Red Neuronal Simple

Como prueba de concepto se valida el modelo generado rn_model con datos no vistos por el modelo y se presentan sus resultados.

Figura 14

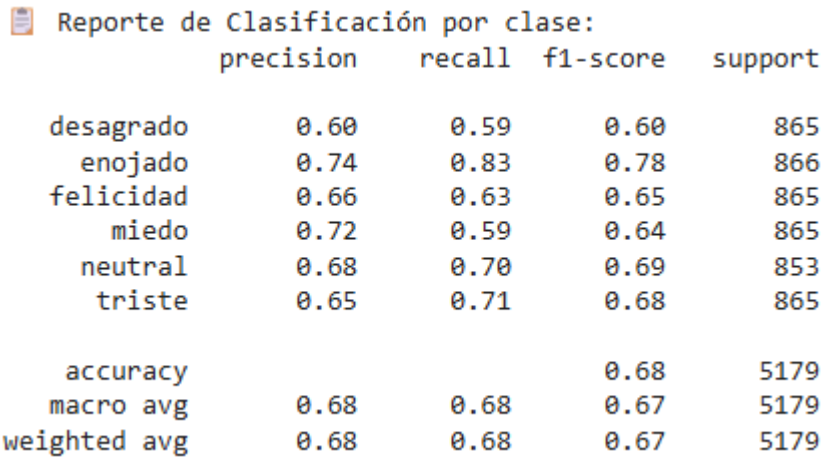
Métricas modelo rn_model.pkl



Nota: Se observa que el valor del Accuracy es del 67,5% de aciertos con respecto al conjunto de datos de validación; datos que no han sido vistos por el modelo.

Figura 15

Métricas por clase



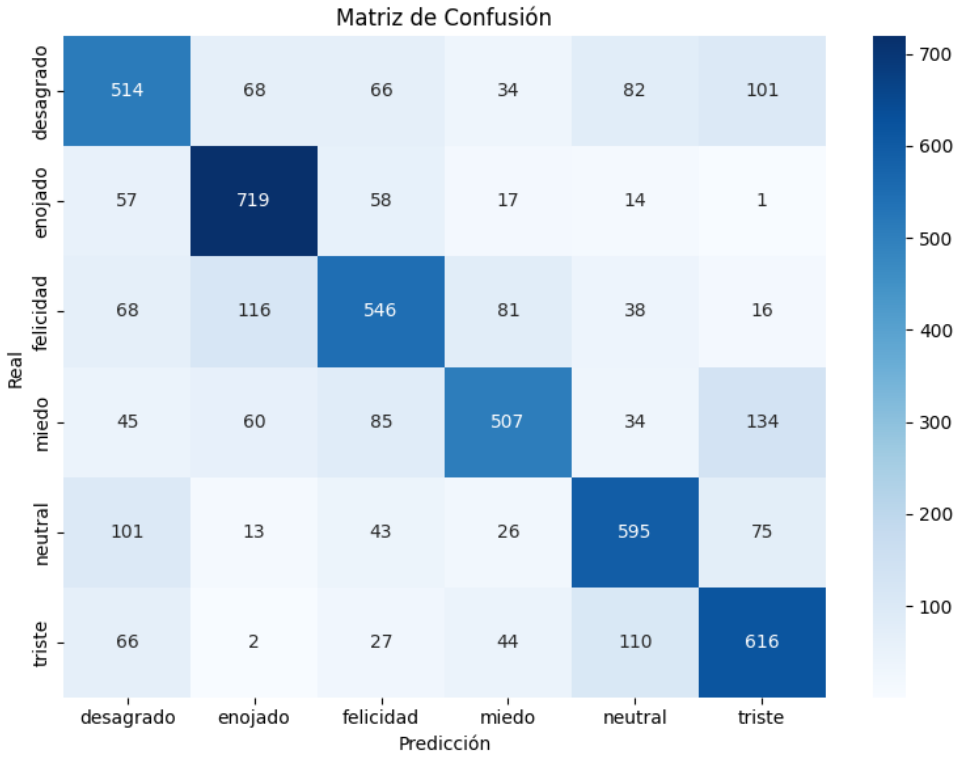
	precision	recall	f1-score	support
desagrado	0.60	0.59	0.60	865
enojado	0.74	0.83	0.78	866
felicidad	0.66	0.63	0.65	865
miedo	0.72	0.59	0.64	865
neutral	0.68	0.70	0.69	853
triste	0.65	0.71	0.68	865
accuracy			0.68	5179
macro avg	0.68	0.68	0.67	5179
weighted avg	0.68	0.68	0.67	5179

De manera similar a lo observado durante el entrenamiento y validación del modelo con el conjunto de prueba, en el nuevo conjunto de datos de validación la clase mejor detectada corresponde a

enojado con un F1 del 78% lo que indica que detecta bien los casos reales, mientras que la clase con el menor rendimiento relativo es desagrado. Todas las clases están representadas en el conjunto de datos.

Figura 16

Matriz de confusión



Nota: En la matriz de confusión se puede observar que la clase de miedo se confunde con tristeza.

Modelo CNN

Como prueba de concepto se valida el modelo generado cnn con datos no vistos por el modelo y se presentan sus resultados.

Figura 17

Métricas de evaluación

```
📊 Métricas de evaluación:
✔ Accuracy: 0.7741
✔ Precision (weighted): 0.7738
✔ Recall (weighted): 0.7741
✔ F1-score (weighted): 0.7737
✔ Balanced Accuracy: 0.7741
```

Nota: Se observa que el valor del Accuracy es del 77,4% de aciertos con respecto al conjunto de datos de validación; datos que no han sido vistos por el modelo.

Figura 18

Métricas por clase

📋 Reporte de Clasificación por clase:

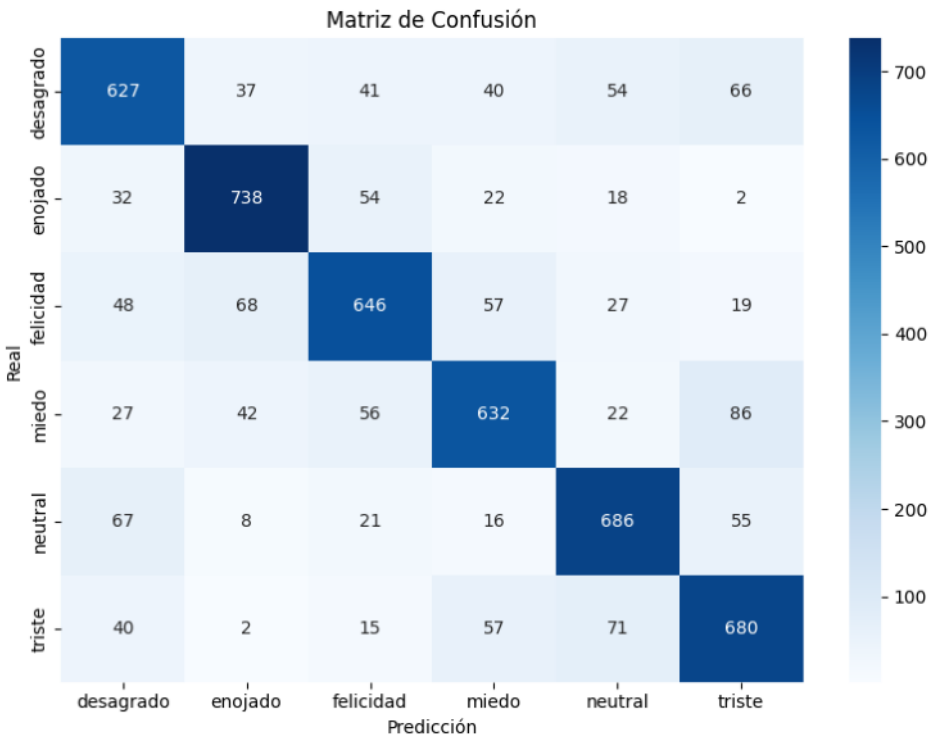
	precision	recall	f1-score	support
desagrado	0.75	0.72	0.74	865
enojado	0.82	0.85	0.84	866
felicidad	0.78	0.75	0.76	865
miedo	0.77	0.73	0.75	865
neutral	0.78	0.80	0.79	853
triste	0.75	0.79	0.77	865
accuracy			0.77	5179
macro avg	0.77	0.77	0.77	5179
weighted avg	0.77	0.77	0.77	5179

De manera similar a lo observado durante el entrenamiento y validación del modelo con el conjunto de prueba, en el nuevo conjunto de datos de validación la clase mejor detectada corresponde a

enojado con un F1 del 84% lo que indica que detecta bien los casos reales, mientras que la clase con el menor rendimiento relativo es desagrado. Todas las clases están representadas uniformemente en el conjunto de datos.

Figura 19

Matriz de confusión



Nota: Se observa en la matriz de confusión que la clase de *miedo* se confunde con las clases de tristeza y neutral.

Modelo Multilayer Perceptron

Para este modelo ya se realiza el entrenamiento con los datos de entrenamiento y validación y se realiza la predicción con el conjunto de datos de prueba, los resultados obtenidos son:

Figura 20

Métricas de evaluación modelo MLP

```
📊 Métricas de evaluación:
✔ Accuracy: 0.7614
✔ Precision (weighted): 0.7635
✔ Recall (weighted): 0.7614
✔ F1-score (weighted): 0.7613
✔ Balanced Accuracy: 0.7614
```

Nota: Se observa que el valor del Accuracy es del 76,1% de aciertos con respecto al conjunto de datos de prueba.

Figura 21

Métricas por clase

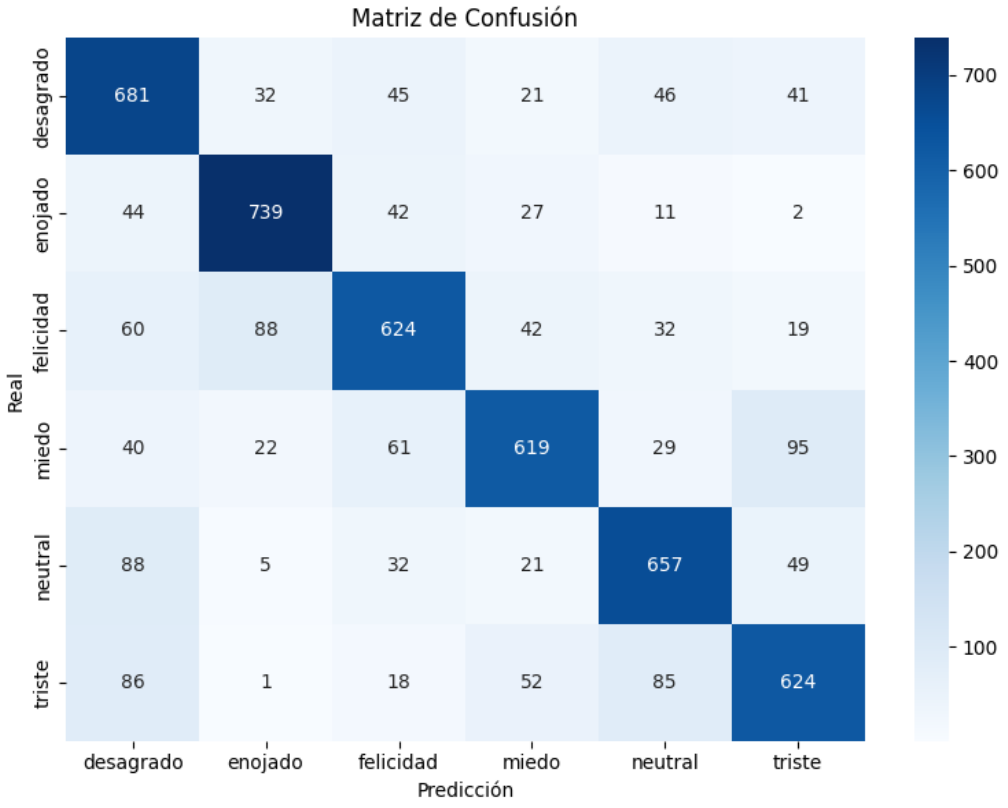
📋 Reporte de Clasificación por clase:

	precision	recall	f1-score	support
desagrado	0.68	0.79	0.73	866
enojado	0.83	0.85	0.84	865
felicidad	0.76	0.72	0.74	865
miedo	0.79	0.71	0.75	866
neutral	0.76	0.77	0.77	852
triste	0.75	0.72	0.74	866
accuracy			0.76	5180
macro avg	0.76	0.76	0.76	5180
weighted avg	0.76	0.76	0.76	5180

Se observa que la clase mejor detectada corresponde a *enojado* con un F1 del 84% lo que indica que detecta bien los casos reales, mientras que la clase con el menor rendimiento relativo es desagrado. Todas las clases están representadas uniformemente en el conjunto de datos.

Figura 22

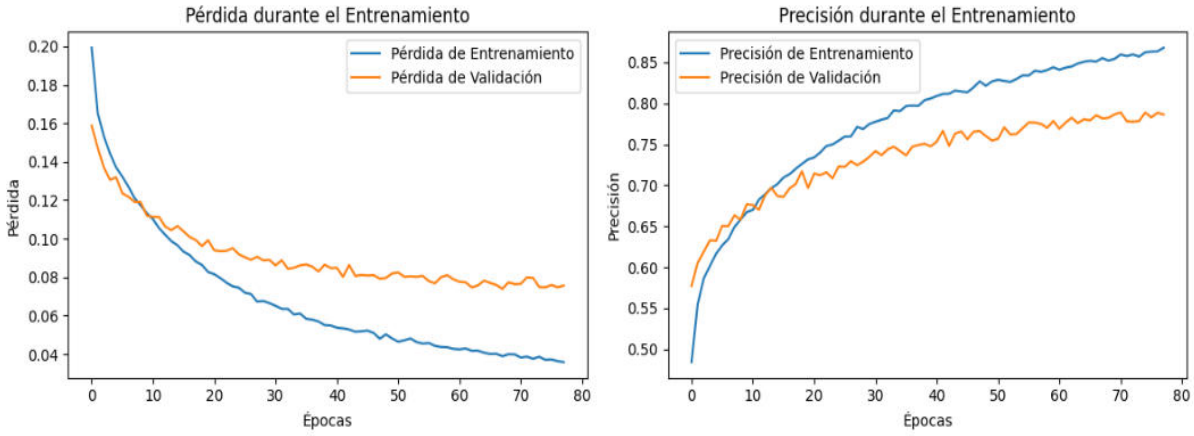
Matriz de confusión



Nota: En la matriz de confusión se observa que la clase de *miedo* en este modelo se confunde con tristeza, otra clase notoria la confusión es *miedo* con *tristeza*.

Figura 23

Gráfico de pérdida



Nota: En la gráfica se puede observar que no se tiene sobreajuste de datos.

Modelo CNN 1D

Figura 24

Métricas Modelo CNN 1D

```
📊 Métricas de evaluación:
✅ Accuracy: 0.7031
✅ Precision (weighted): 0.7030
✅ Recall (weighted): 0.7031
✅ F1-score (weighted): 0.7026
✅ Balanced Accuracy: 0.7032
```

Nota: Se observa que el valor del Accuracy es del 70,3% de aciertos con respecto al conjunto de datos de validación.

Figura 25

Métricas por clase del modelo CNN 1D

📋 Reporte de Clasificación por clase:

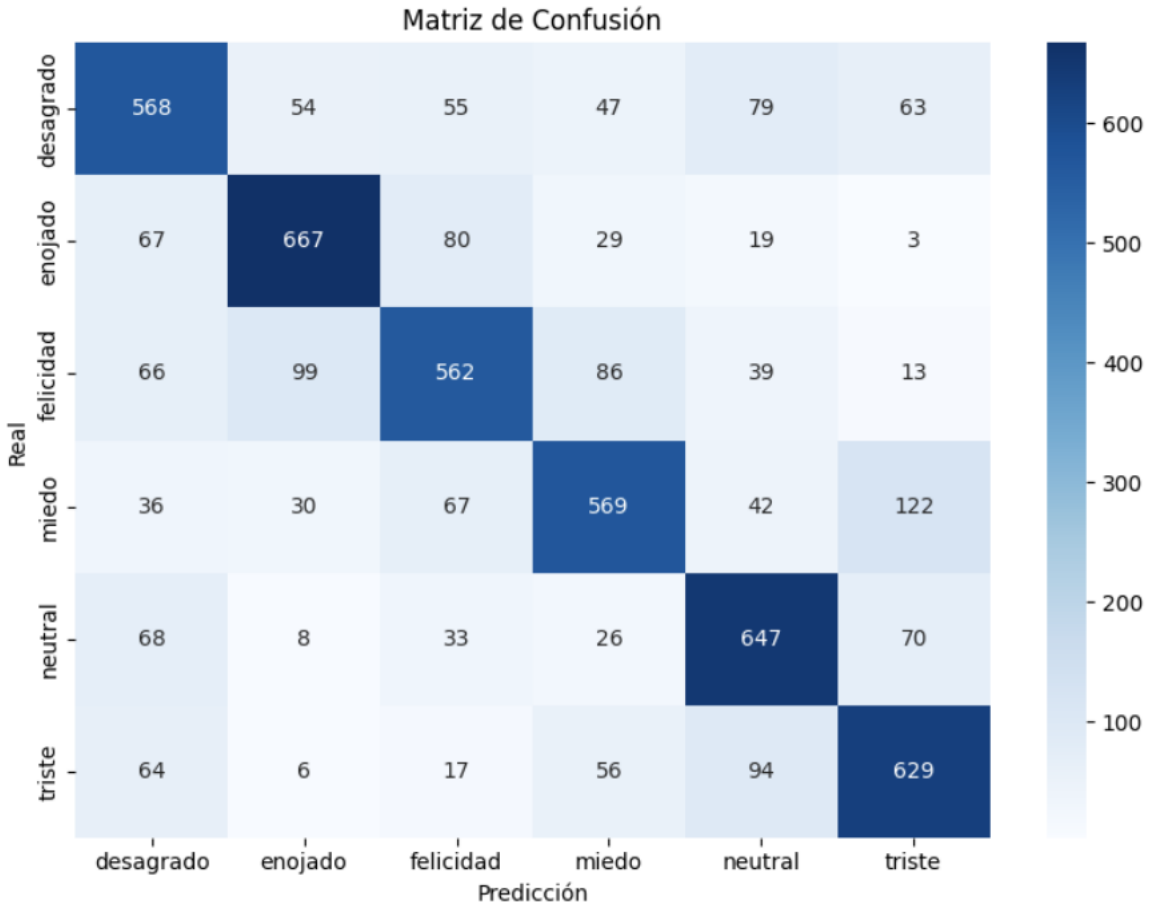
	precision	recall	f1-score	support
desagrado	0.65	0.66	0.65	866
enojado	0.77	0.77	0.77	865
felicidad	0.69	0.65	0.67	865
miedo	0.70	0.66	0.68	866
neutral	0.70	0.76	0.73	852
triste	0.70	0.73	0.71	866
accuracy			0.70	5180
macro avg	0.70	0.70	0.70	5180
weighted avg	0.70	0.70	0.70	5180

Estos resultados muestran que el modelo tiene un mejor desempeño en la clase "enojado", con datos altos de precisión y puntaje F1. La clase "desagrado" presenta los valores más bajos en estas métricas.

Adicional, la clase "neutral" tiene el support más bajo, lo que sugiere que el modelo tiene dificultades para clasificar correctamente esta clase.

Figura 26

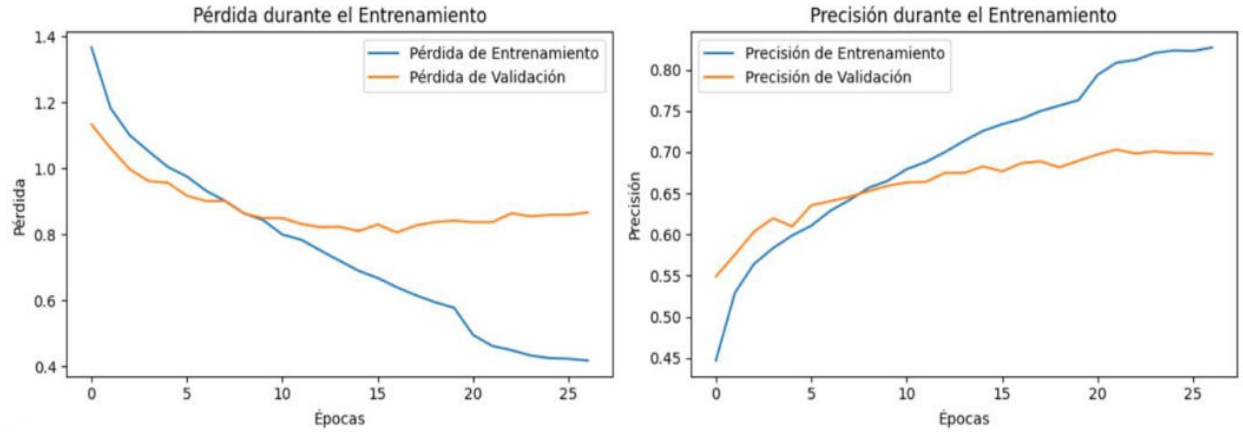
Matriz de Confusión del modelo CNN 1D



Nota: En la matriz de confusión se observa que la clase de “enojado”, “neutral” y “triste” tiene valores altos en la diagonal principal, se presenta confusiones entre las clases ""miedo" y "triste".

Figura 27

Gráfico de precisión y Pérdida durante el entrenamiento Modelo CNN 1D

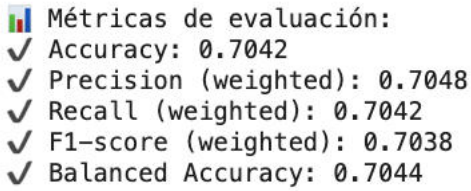


Nota: Este gráfico muestra que el modelo se está entrenando de manera apropiada, con una buena capacidad de generalización y sin signos de sobreajuste.

Modelo CNN 1D L2

Figura 28

Métricas Modelo CNN 1D L2



Nota: Se observa que el valor del Accuracy es del 70,4% de aciertos con respecto al conjunto de datos de validación.

Figura 29

Métricas por clase del modelo CNN 1D L2

📋 Reporte de Clasificación por clase:

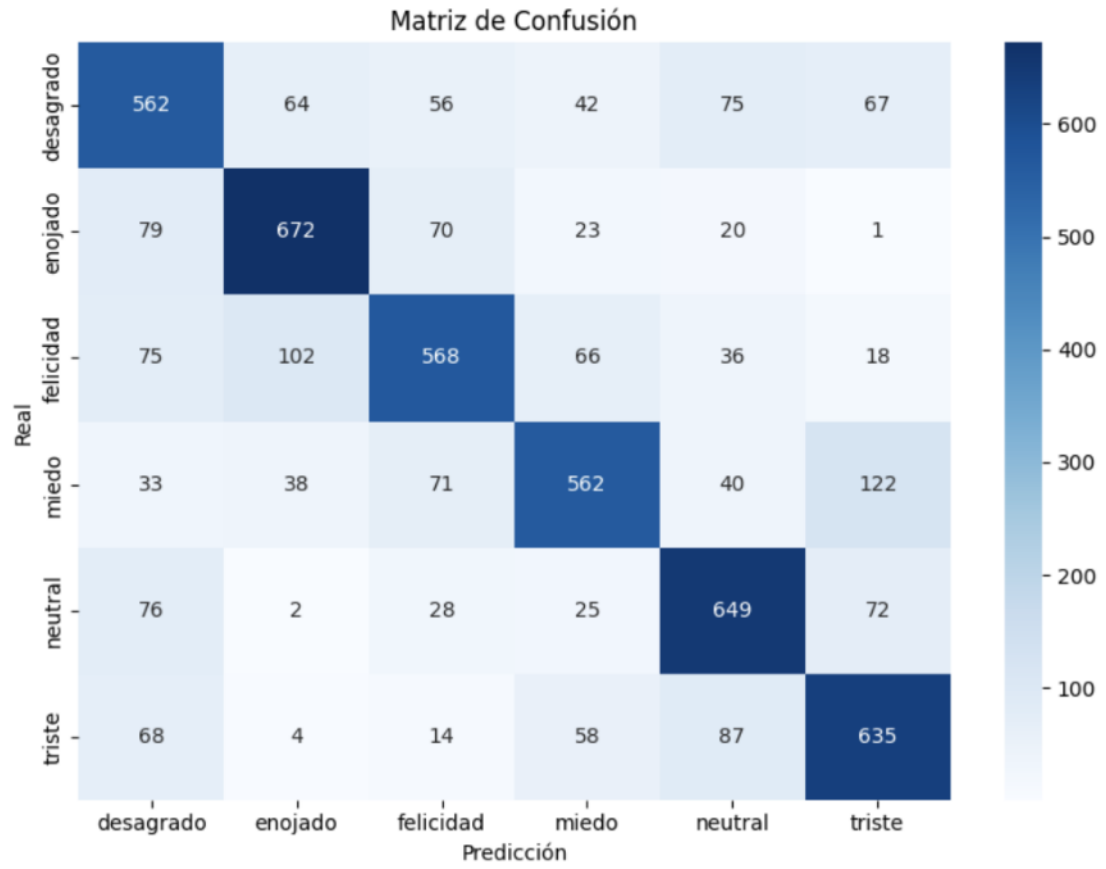
	precision	recall	f1-score	support
desagrado	0.63	0.65	0.64	866
enojado	0.76	0.78	0.77	865
felicidad	0.70	0.66	0.68	865
miedo	0.72	0.65	0.68	866
neutral	0.72	0.76	0.74	852
triste	0.69	0.73	0.71	866
accuracy			0.70	5180
macro avg	0.70	0.70	0.70	5180
weighted avg	0.70	0.70	0.70	5180

Estos resultados muestran que el modelo tiene un mejor desempeño en la clase "enojado", con datos altos de precisión y puntaje F1. La clase "desagrado" presenta los valores más bajos en estas métricas.

Adicional, la clase "neutral" tiene el support más bajo, lo que sugiere que el modelo tiene dificultades para clasificar correctamente esta clase.

Figura 30

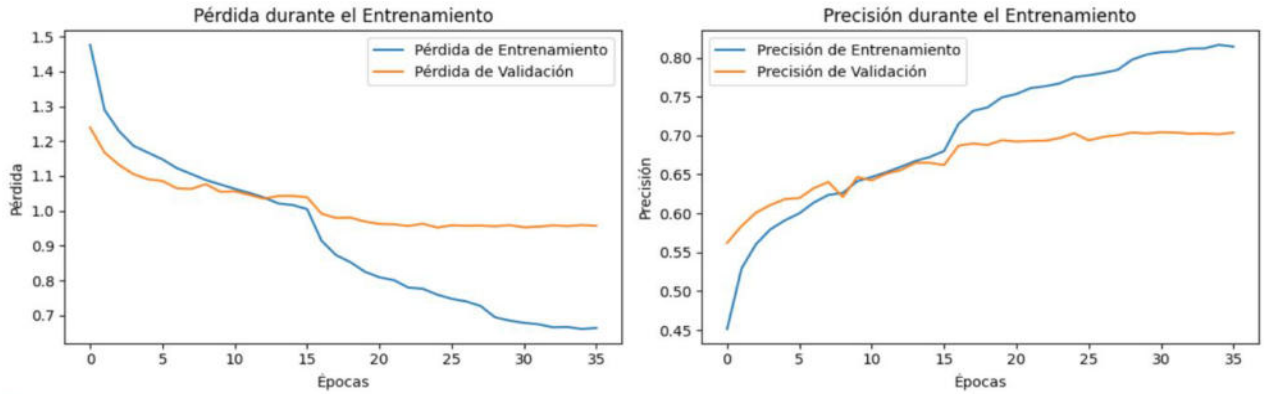
Matriz de Confusión modelo CNN 1D L2



Nota: En la matriz de confusión se observa que la clase de “enojado”, “neutral” y “triste” tiene valores altos en la diagonal principal, se presenta confusiones entre las clases ""miedo" y "triste" y entre “felicidad” y “enojado”.

Figura 31

Gráfico de precisión y Pérdida durante el entrenamiento Modelo CNN 1D L2



Nota: Este gráfico muestra que el modelo se está entrenando de manera apropiada, con una buena capacidad de generalización y sin signos de sobreajuste

Modelo CNN-LSTM

Figura 32

Métricas Modelo CNN - LSTM

```
📊 Métricas de evaluación:
✓ Accuracy: 0.6498
✓ Precision (weighted): 0.6531
✓ Recall (weighted): 0.6498
✓ F1-score (weighted): 0.6496
✓ Balanced Accuracy: 0.6500
```

Nota: Muestra que el valor del Accuracy es del 64,9% de aciertos con respecto al conjunto de datos de validación.

Figura 33

Matriz por clase del modelo CNN - LSTM

📋 Reporte de Clasificación por clase:

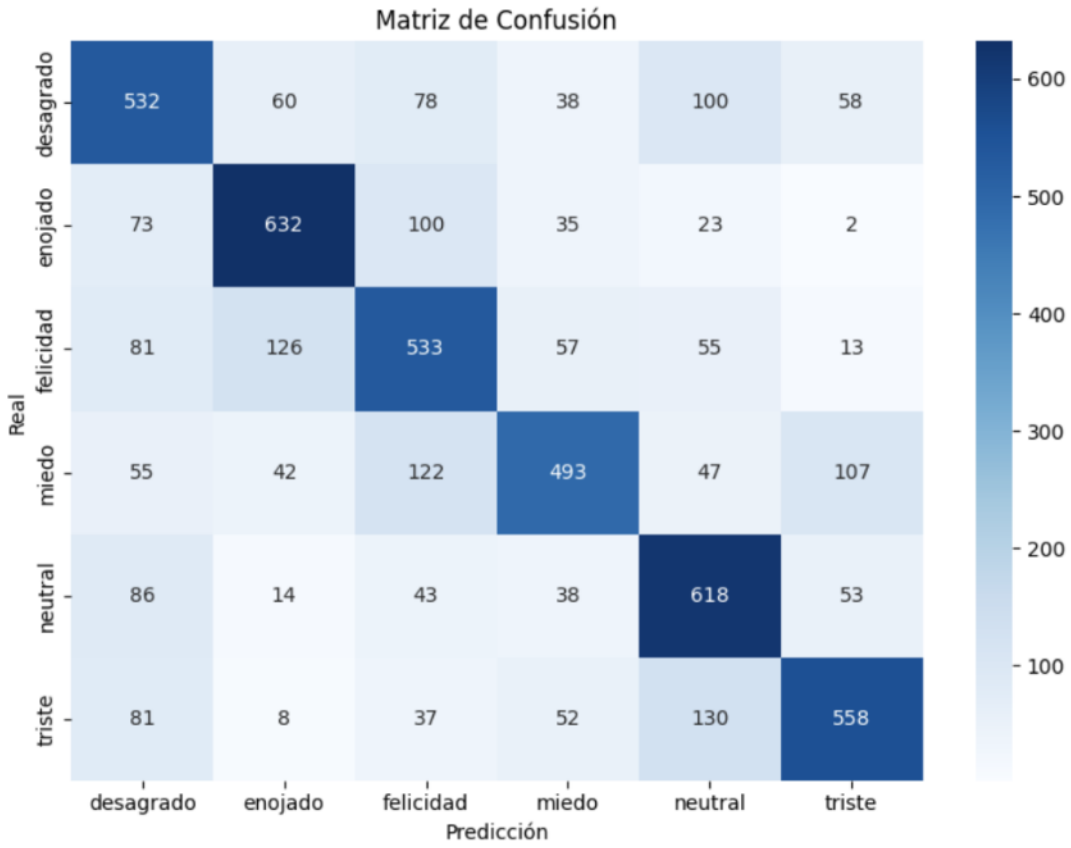
	precision	recall	f1-score	support
desagrado	0.59	0.61	0.60	866
enojado	0.72	0.73	0.72	865
felicidad	0.58	0.62	0.60	865
miedo	0.69	0.57	0.62	866
neutral	0.64	0.73	0.68	852
triste	0.71	0.64	0.67	866
accuracy			0.65	5180
macro avg	0.65	0.65	0.65	5180
weighted avg	0.65	0.65	0.65	5180

Estos resultados muestran que el modelo tiene un mejor desempeño en la clase "enojado", con información alta de precisión y puntaje F1. La clase "desagrado" presenta los valores más bajos en estas métricas.

Adicional, la clase "neutral" tiene el support más bajo, lo que sugiere que el modelo tiene dificultades para clasificar correctamente esta clase.

Figura 34

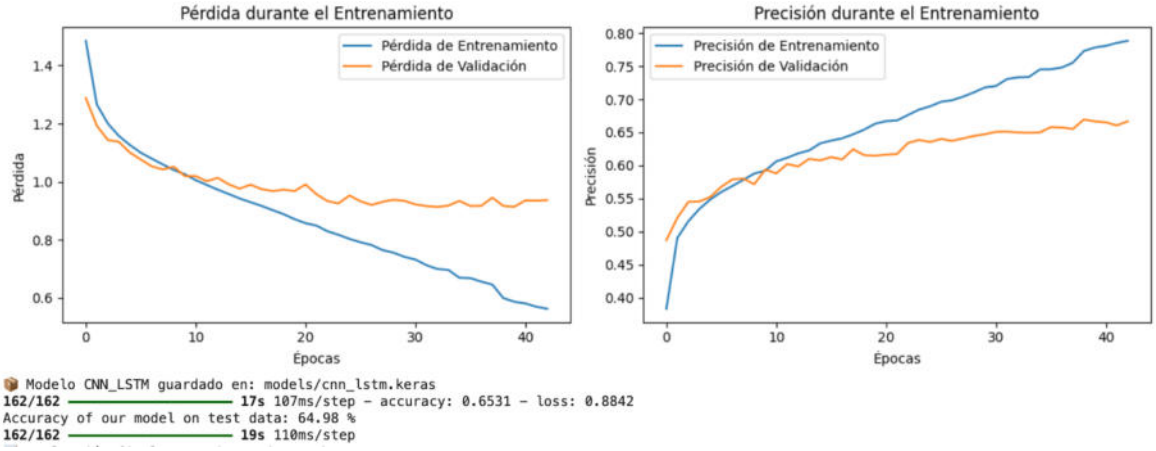
Matriz de Confusión modelo CNN - LSTM



Nota: En la matriz de confusión se observa que la clase de “enojado”, “neutral” tiene valores altos en la diagonal principal, se presenta confusiones entre las clases "miedo" y "triste" , entre “felicidad” y “enojado” y entre "miedo" y “felicidad”.

Figura 35

Gráfico de precisión y Pérdida durante el entrenamiento Modelo CNN - LSTM

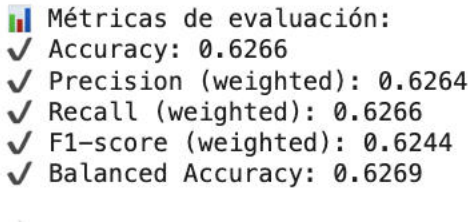


Nota: Se observa en este gráfico que el modelo se está entrenando de manera apropiada, con una buena capacidad de generalización y sin sobreajuste

Modelo CNN-LSTM L2

Figura 36

Métricas Modelo CNN - LSTM L2



Nota: Muestra que el valor del Accuracy es del 62,6% de aciertos con respecto al conjunto de datos de validación.

Figura 37

Matriz por clase del modelo CNN - LSTM L2

Reporte de Clasificación por clase:

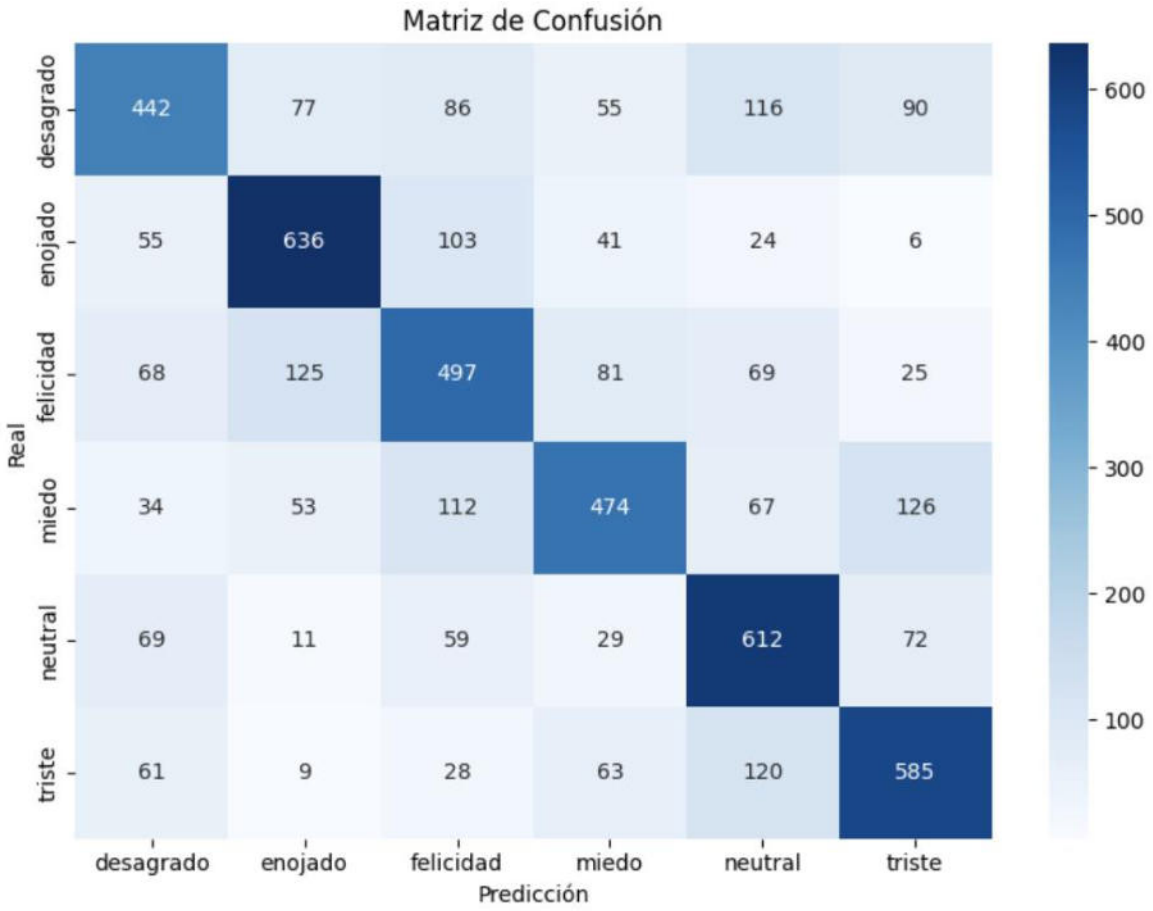
	precision	recall	f1-score	support
desagrado	0.61	0.51	0.55	866
enojado	0.70	0.74	0.72	865
felicidad	0.56	0.57	0.57	865
miedo	0.64	0.55	0.59	866
neutral	0.61	0.72	0.66	852
triste	0.65	0.68	0.66	866
accuracy			0.63	5180
macro avg	0.63	0.63	0.62	5180
weighted avg	0.63	0.63	0.62	5180

Estos resultados muestran que el modelo tiene un mejor desempeño en la clase "enojado", con información alta de precisión y puntaje F1. La clase "desagrado" presenta los valores más bajos en estas métricas.

Adicional, la clase "neutral" tiene el support más bajo, lo que sugiere que el modelo tiene dificultades para clasificar correctamente esta clase.

Figura 38

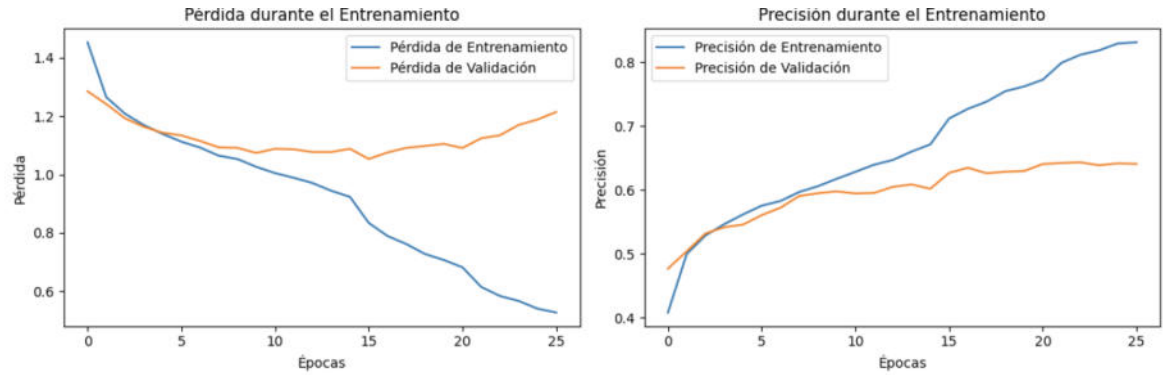
Matriz de Confusión modelo CNN - LSTM L2



Nota: En la matriz de confusión se observa que la clase de “enojado”, “neutral” tiene valores altos en la diagonal principal, se presenta confusiones entre las clases "miedo" y "triste" , entre “felicidad” y “enojado” y entre "miedo" y “felicidad”.

Figura 39

Gráfico de precisión y Pérdida durante el entrenamiento Modelo CNN - LSTM L2




```
📦 Modelo CNN_LSTM guardado en: models/cnn_lstml2.keras
162/162 ━━━━━━━━━━━━━━━━━━━━ 16s 99ms/step - accuracy: 0.6332 - loss: 1.0505
Accuracy of our model on test data: 62.66 %
162/162 ━━━━━━━━━━━━━━━━━━━━ 17s 103ms/step
```

Nota: Se observa en este gráfico que el modelo se está entrenando de manera apropiada, con una buena capacidad de generalización y sin sobreajuste.

Análisis de Resultados

Análisis De Resultados - Comparación De Modelos

En esta sección se presentan los resultados obtenidos para distintos modelos de clasificación aplicados al reconocimiento de emociones a partir de señales de voz. Se evaluaron las métricas, con énfasis en el F1-score por clase, así como las conclusiones derivadas de las matrices de confusión, las cuales permiten identificar patrones de aciertos y errores comunes entre las emociones. El objetivo de este análisis es determinar qué modelos presentan mejor desempeño general, cuáles emociones son más fácilmente identificables por los modelos y en qué casos se presentan mayores confusiones, especialmente entre emociones con características acústicas similares.

Tabla 6

Métricas de los modelos entrenados

Modelo	Accuracy	Precisión	Recall	F1	Balance Accuracy
Random Forest	64.98%	66.27%	64.98%	64.63%	65.01%
SVM	73.90%	74.05%	73.90%	73.89%	73.91%
Xgboost	67.37%	67.61%	67.37%	67.23%	67.40%
Red Neuronal Simple	67.525	67.515	67.52%	67.33%	67.52%
CNN	75.48%	75.49%	75.48%	75.45%	75.49%
MLP	76.14%	76.35%	76.14%	76.13%	76.14%
CNN1D	70.31%	70.30%	70.31%	70.26%	70.32%
CNN1DL2	70.42%	70.48%	70.42%	70.38%	70.44%
CNNLSTM	64.98%	65.31%	64.98%	64.96%	65%
CNNLSTML2	62.66%	62.64%	62.66%	62.44%	62.69%

Tabla 7

Análisis de F1 y matriz de confusión

Modelo	Clases que mejor predice	Clases que peor predice	Matriz de confusión
Random Forest	Enojado 73%	Desagrado 58%	Enojado es la emoción que mejor predice Existe confusión entre triste y neutral y miedo - triste
SVM	Enoojado 82%	Desagrado 68%	Enojado es la emoción que mejor predice Existe confusión miedo - triste

Xgboost	Enojado 75%	Desagrado 59%	Enojado es la emoción que mejor predice Existe confusión entre triste y neutral
Red Neuronal Simple	Enojado 78%	Desagrado 60%	Enojado es la emoción que mejor predice Existe confusión entre miedo - triste
CNN	Enojado 84%	Desagrado 74%	Enojado es la emoción que mejor predice Existe confusión entre triste y neutral y miedo - triste
MLP	Enojado 85%	Desagrado 73%	Enojado es la emoción que mejor predice Existe confusión entre triste y neutral y miedo - triste
CNN1D	Enojado 77%	Desagrado 65%	Enojado es la emoción que mejor predice Existe confusión entre miedo - triste
CNN1DL2	Enojado 77%	Desagrado 65%	Enojado es la emoción que mejor predice Existe confusión entre miedo - triste
CNNLSTM	Enojado 72%	Desagrado y Felicidad 60%	Enojado es la emoción que mejor predice

			Existe confusión entre triste y neutral, miedo – triste, felicidad - enojado
CNNLSTML2	Enojado 72%	Desagrado y Felicidad 60%	Enojado es la emoción que mejor predice Existe confusión entre triste y neutral

De los modelos entrenados en el presente proyecto, el MLP es el que mejor métricas obtuvo, con un accuracy de 76.14%, lo que quiere decir que aproximadamente el 76% de las predicciones de las emociones son correctas y el 24% generan confusión entre las distintas emociones. Las métricas obtenidas como recall, precision, F1-score y balanced accuracy— con valores cercanos al 76%, lo que signifca que el modelo de MLP para reconocimiento de emociones en voz tiene un rendimiento aceptable y equilibrado en la clasificación de las seis emociones. Estas métricas permiten identificar correctamente las emociones (recall), evitar falsos positivos (precision) y mantener un desempeño uniforme entre clases. Estos resultados muestran que el modelo ha aprendido patrones relevantes y generaliza bien en el reconocimiento emocional.

Con base en los resultados de F1-score y las observaciones de las matrices de confusión presentados en la tabla precedente, se evidencia que la emoción “enojado” es claramente la mejor identificada en todos los modelos, con un máximo desempeño en MLP (85%) y CNN (84%), lo que representa que las características acústicas de esta emoción son más representativas. Por el contrario, "desagrado" y "felicidad" son las emociones que presentan mayores dificultades, con F1-scores bajos (por ejemplo, 58% en Random Forest y 60% en CNNLSTM para ambas clases). Las confusiones más frecuentes en las matrices de confusión ocurren entre "triste" y "neutral", y entre "miedo" y "triste", lo

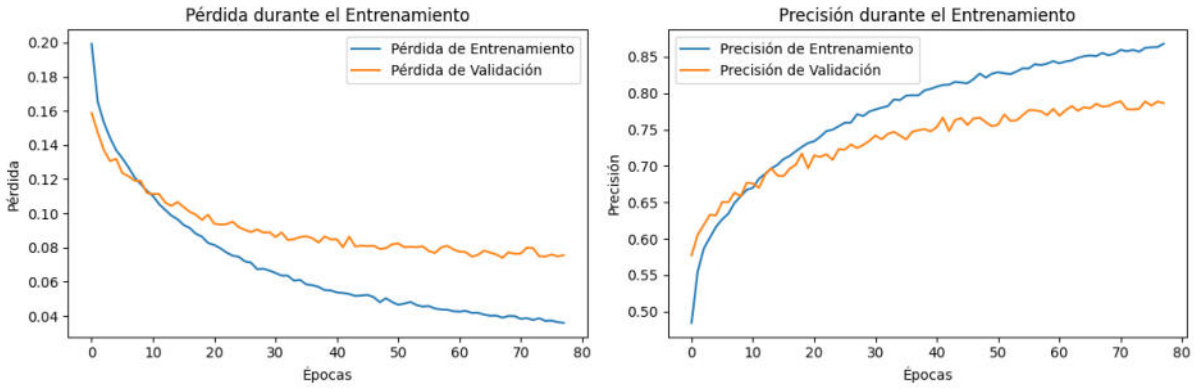
cual indica una superposición acústica significativa entre estas emociones, probablemente debido a características vocales similares. Algunos modelos, como CNNLSTM, también muestran confusión adicional entre "felicidad" y "enojado", lo que sugiere que en ciertos casos las emociones de alta activación pueden ser difíciles de diferenciar.

Análisis Del Rendimiento Del Modelo

En la Figura 40 muestra la evolución de la pérdida (loss) y la precisión (accuracy) tanto para el conjunto de entrenamiento como para el de validación durante 80 épocas de entrenamiento del modelo.

Figura 40

Gráfica de Pérdida durante el entrenamiento



Perdida durante el entrenamiento: En la gráfica se observa una disminución progresiva y sostenida de la pérdida en el conjunto de entrenamiento (línea azul), lo cual indica que el modelo aprende correctamente a minimizar el error durante el ajuste de sus parámetros. En las primeras 15 a 20 épocas, la pérdida de validación (línea tomate) también decrece significativamente, lo que sugiere un buen aprendizaje inicial y capacidad de generalización. Posteriormente, esta curva se estabiliza y presenta ligeras fluctuaciones sin incrementos relevantes, lo cual indica que no existe un sobreajuste severo.

A pesar de que la brecha entre la pérdida de entrenamiento y la de validación aumenta a partir de la época 20, esta diferencia se mantiene controlada. Esto sugiere una ligera tendencia al sobreajuste, aunque sin comprometer gravemente la capacidad del modelo para generalizar sobre datos no vistos.

Precisión durante el entrenamiento: En la gráfica, se observa un comportamiento coherente con el análisis anterior. La precisión del entrenamiento aumenta de manera continua y alcanza valores cercanos al 87 %, mientras que la precisión en validación se incrementa rápidamente hasta estabilizarse alrededor del 77 %. Estos valores indican que el modelo logra un buen rendimiento predictivo, aunque existe una brecha de aproximadamente 10 puntos porcentuales entre la precisión de entrenamiento y validación, lo cual refuerza la hipótesis de un leve sobreajuste.

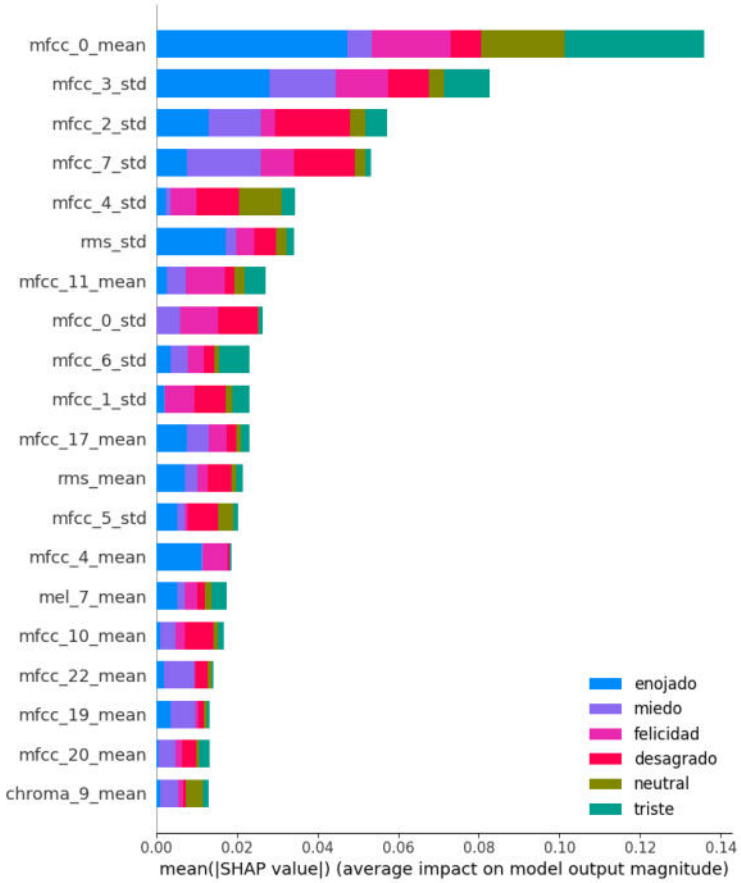
Convergencia y generalización: En la gráfica se puede observar que las curvas indican que el modelo alcanza una convergencia estable a partir de la época 40, tanto en pérdida como en precisión. Aunque existe una brecha entre la pérdida de entrenamiento y la de validación, esta no es excesiva ni creciente, por lo que se puede indicar que el modelo generaliza de manera aceptable a los datos no vistos.

Principales Características – Interpretabilidad

Para identificar las características que tienen mayor importancia se utiliza SHAP, a continuación, se realiza el detalle de las principales características que afectan a cada una de las clases.

Figura 41

Gráfica de Pérdida durante el entrenamiento



Principales Características En Orden De Importancia

Las siguientes características tienen mayor peso en la decisión del modelo:

mfcc_0_mean: Tiene el mayor impacto general, influye fuertemente en la predicción de las clases triste, neutral y felicidad, mfcc_0: está fuertemente relacionado con la energía y tono general del habla.

mfcc_3_std, mfcc_2_std, mfcc_7_std: La variabilidad (desviación estándar) de estos coeficientes MFCC influye notablemente en varias emociones. Implican cambios acústicos importantes probablemente diferencias prosódicas entre enojo, miedo y disgusto.

mfcc_4_std: Puede capturar matices del timbre

rms_std: variabilidad de la energía de la señal, emociones intensas como enojo o miedo.

mfcc_11_mean, mfcc_0_std: Menos influyentes, pero con impacto distribuido en varias emociones.

Tabla 8

Interpretación por clase

Clase	**Característica**
Enojado (azul)	mfcc_0_mean, mfcc_3_std, rms_std tienen alta influencia en emociones como enojo se relacionan con energía alta y variabilidad.
Miedo (morado)	Características similares al enojo con menos intensidad
Felicidad (rosa)	influenciada por mfcc_0_mean, mfcc_2_std, mfcc_7_std, habla más armónica y con menor variabilidad en el tono.
Triste (verde)	fuertemente influenciado por mfcc_0_mean, con menor energía y tono.
Mostaza (neutral)	más homogénea, sin rasgos extremos, aparece en múltiples características de forma moderada.

CAPITULO 5

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

El conjunto de datos utilizado en el presente estudio consiste en una serie de audios en idioma inglés, grabados de forma profesional bajo ambientes controlados y con la vocalización de actores o actrices. Sin embargo; en los ambientes reales, no se tienen estas condiciones ideales, por lo que bajo circunstancias normales el comportamiento, predicción y resultados pueden ser diferentes. Como se mencionó durante el proyecto, incluso el ámbito cultural podría influir en condiciones reales de aplicabilidad, lo que lo vuelve un problema complejo.

El proyecto planteó el ejercicio de efectuar el reconocimiento de emociones en la voz a partir de la extracción y análisis de diversas características acústicas, tales como MFCC, espectrograma de Mel, RMS, Zero-Crossing Rate (ZCR) y Chroma STFT, características que representan diferentes aspectos del comportamiento vocal, desde la energía (RMS), la textura espectral (Mel, STFT), hasta patrones fonéticos (MFCC), todos considerados como relevantes al momento de identificar la expresión emocional en el habla.

Para interpretar cómo estas características influían en las decisiones del mejor modelo MLP, se utilizó SHAP y el método KernelExplainer, el cual permitió obtener explicaciones del comportamiento del modelo. Esta herramienta reveló, por ejemplo, que, entre las características acústicas analizadas, especialmente los coeficientes MFCC, desempeñan un papel importante en la predicción de emociones. El promedio del coeficiente mfcc_0 destaca por su influencia significativa en emociones como tristeza, neutralidad y felicidad, al reflejar aspectos fundamentales del tono y la energía del habla. Sin embargo, su influencia también podría explicar parte de la confusión entre clases como tristeza y neutral, ya que ambas comparten patrones acústicos similares en este parámetro. Asimismo, la desviación estándar de

ciertos MFCC (mfcc_2, mfcc_3, mfcc_7) y del valor rms muestra cómo la variabilidad prosódica y energética está estrechamente relacionada con emociones de mayor intensidad como el enojo y el miedo.

En general, los modelos basados en redes neuronales profundas, especialmente MLP y CNN, muestran mejor capacidad para discriminar emociones, aunque persisten desafíos en la separación de emociones similares o de baja intensidad.

Si bien el tratamiento del problema, abordado en el proyecto a través de las características acústicas y modelos clásicos, generan los resultados esperados; se podrían plantean otras formas de formular este tipo de análisis. Los modelos basados en deep learning, como convolutional neural networks (CNNs) podrían ser aplicados a espectrogramas, o transformers entrenados directamente sobre las señales crudas. Sin embargo, se debe considerar que estos modelos también implican una mayor complejidad y posiblemente una menor interpretabilidad perdiendo el sentido de que el modelo sea analíticamente comprensible.

La implementación de sistemas de detección de emociones puede beneficiar significativamente sectores como la atención médica, el marketing y la seguridad, permitiendo una mejor comprensión de las necesidades humanas y optimizando las estrategias de respuesta.

Recomendaciones

Si bien el objetivo del presente proyecto era el análisis de reconocimiento de emociones en voz, el principal desafío que tiene actualmente la Inteligencia Artificial es volver las aplicaciones más personalizadas, que capten emociones y necesidades de los usuarios. Para ello se plantea un análisis combinado de datos acústicos con información visual que permita analizar las expresiones faciales y complementar con el análisis de texto o transcripciones del habla, lo que permitiría crear un enfoque de análisis emocional más integral.

Implementar protocolos de seguridad y privacidad robustos es esencial en los sectores de atención médica, marketing y seguridad, principalmente cuando se manejan datos sensibles relacionados con la salud mental de los usuarios. Estas medidas no solo protegen a los individuos, sino que también fortalecen la confianza del público en el uso de tecnologías que analizan emociones, promoviendo un desarrollo ético y responsable en todos los ámbitos.

Para mejorar la capacidad de generalización y la robustez del sistema de reconocimiento de emociones, se recomienda emplear datasets más extensos y diversos que contemplen una amplia gama de características demográficas y culturales. Esto incluye voces de hablantes de distintas edades (infantes, adultos jóvenes, personas mayores), diferentes géneros, acentos y variedades lingüísticas. Además, resulta esencial priorizar la recopilación o el uso de grabaciones que reflejen emociones expresadas de manera espontánea en contextos reales, en lugar de emociones actuadas o inducidas en laboratorio. Este tipo de datos naturales permitirá que el modelo aprenda patrones emocionales más representativos y aplicables a situaciones del mundo real, reduciendo el riesgo de sobreajuste a un conjunto de datos específico y aumentando su utilidad práctica en aplicaciones como asistentes virtuales, atención al cliente o sistemas de monitoreo emocional.

Referencias

Cabrelles Sagredo, M.S. La influencia de las emociones en el sonido de la voz. Biblioteca Virtual Miguel de Cervantes. https://www.cervantesvirtual.com/obra-visor/la-influencia-de-las-emociones-en-el-sonido-de-la-voz/html/

PHR ROBOTICS. (21 abril). Terapia emocional asistida por robots: una nueva frontera en la salud mental. https://www.phr-robotics.com/robots/terapia-emocional-robots

Fahmi, A. (Aug 30, 2024). The Future of AI-Powered Personal Assistants: Beyond Siri and Alexa. Medium, https://medium.com/@fahmiadam/the-future-of-ai-powered-personal-assistants-beyond-siri-and-alexa-85777dc5b3cb

Livingstone SR, RAVDESS Emotional speech audio, Kaggle. https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

Eu Jin Lok, CREMA - D, Kaggle. https://www.kaggle.com/datasets/ejlok1/cremad

Eu Jin Lok, Toronto emotional speech set (TESS), Kaggle. https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess

Eu Jin Lok, Surrey Audio-Visual Expressed Emotion (SAVEE), Kaggle. https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee

López, A. C. (2025). Machine Learning - Aprendizaje Automático.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Springer, Ed.).

MacDonald y Roland Langrock, W. Z. I. L. (2016). Hidden Markov Models for Time Series: An Introduction Using R (C. A. Hall/CRC, Ed.).

Reynolds, D. A. (2009). *Gaussian Mixture Models*. En *Encyclopedia of Biometrics* (Springer, Ed.).

Librosa (2025), https://librosa.org/doc/latest/index.html

Influencias culturales en la expresión emocional: Navegando por el mosaico de los sentimientos humanos, Medium, https://medium.com/introducci%C3%B3n-a-la-psicolog%C3%ADa-ugr/influencias-culturales-en-la-expresi%C3%B3n-emocional-navegando-por-el-mosaico-de-los-sentimientos-29bf025150fb

Ecuador: tres iniciativas diferentes de regulación de la IA, Centro Competencia, https://centrocompetencia.com/ecuador-tres-iniciativas-diferentes-de-regulacion-de-la-ia/

Kingma, D. P., & Ba, J. L. (2015). Adam: A Method for Stochastic Optimization. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6980

OpenAI. (2025, julio 4). *ChatGPT* (versión 4o). https://chat.openai.com/chat

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Leanpub. https://christophm.github.io/interpretable-ml-book/

Schuller, B., Rigoll, G., & Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine–belief network architecture

Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks

Correa, L. (2019). Reconocimiento automático de emociones en español mediante señales de voz

López, F. (2019). Reconocimiento de emociones en la voz mediante aprendizaje profundo

Khalil R., Jones E., Babar M. et al. (2017). Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access

Livingstone S. & Russo F. (2018) "*The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)"*

Ian Goodfellow, Y. B. y. A. C. (2016). *Deep Learning*. MIT Press

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Tan, M., & Le, Q. V. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*". Proceedings of the 36th International Conference on Machine Learning (ICML), 97, 6105-6114.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.

Lin, T:Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929–1958.