Universidad Católica de Santa María

Facultad de Ciencias e Ingenierías Físicas y Formales

Escuela Profesional de Ingeniería de Sistemas

Desarrollo e implementación de un sistema para detectar y explicar patrones emocionales en la voz en escenas de películas usando redes neuronales y clasificadores interpretativos

Tesis presentada por el Bachiller:

Monje Bolivar, Ronaldo Alejandro

ORCID: 0009-0003-6245-7671

para optar el Título Profesional de Ingeniero de Sistemas

Asesor:

Dr. Esquicha Tejada, José David

ORCID: 0000-0002-0191-7174

Arequipa – Perú

2026

UCSM-ERP

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

INGENIERIA DE SISTEMAS

TITULACIÓN CON TESIS

DICTAMEN APROBACIÓN DE BORRADOR

Arequipa, 17 de Noviembre del 2025

Dictamen: 014608-C-EPIS-2025

Visto el borrador del expediente 014608, presentado por:

2016250451 - MONJE BOLIVAR RONALDO ALEJANDRO

Titulado:

DESARROLLO E IMPLEMENTACIÓN DE UN SISTEMA PARA DETECTAR Y EXPLICAR PATRONES EMOCIONALES EN LA VOZ EN ESCENAS DE PELÍCULAS USANDO REDES NEURONALES Y CLASIFICADORES INTERPRETATIVOS

Nuestro dictamen es:

APROBADO

Titulo Profesional/Titulo de Segunda Especialidad/Grado Académico a optar:

INGENIERO DE SISTEMAS

29643112 - GUEVARA PUENTE DE LA VEGA KARIM
DICTAMINADOR

29601217 - ROSAS PAREDES KARINA
DICTAMINADOR

71132586 - ANGULO OSORIO JAVIER FERNANDO
DICTAMINADOR

Desarrollo e implementación de un sistema para detectar y explicar patrones emocionales en la voz en escenas de películas usando redes neuronales y clasificadores interpretativos

INFORME DE ORIGINALIDAD

16%	15%	11%	11%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	Submitted to Universidad Católica de Santa María Trabajo del estudiante	1%
2	www.mdpi.com Fuente de Internet	1%
3	theses.hal.science Fuente de Internet	<1%
4	publikationen.bibliothek.kit.edu Fuente de Internet	<1%
5	iieta.org Fuente de Internet	<1%
6	repositorio.ucsm.edu.pe Fuente de Internet	<1%
7	Submitted to Asia Pacific International College Trabajo del estudiante	<1%

DEDICATORIA

A mis familiares, en especial a mi madre Marielena, por permitirme alcanzar muchos sueños,

incluso ante las mayores dificultades.

A las comunidades de videojuegos, slackline, así como a las personas y espacios que conectan y valoran intentar algo nuevo, donde el hecho de no compartir un idioma enriquece la experiencia: cada quien recorre su propio camino, un proceso de descubrimiento para entendernos, compartir y avanzar juntos, llegando a un resultado común. Porque la emoción

está dentro de todos, un lenguaje que todos hablamos.

A las series, juegos y películas que me enseñaron cosas que no entendía con palabras: el deseo de tener emociones con Violet Evergarden, la diferencia entre usar una palabra, como

un demonio, y entenderla por un humano en Frieren.

El riesgo y la confusión de querer, por encima del miedo, la espontaneidad necesaria que implica amar, que me hizo reflexionar From Me to You.

La valentía de volver a cuidar después de fallar en The Last of Us, especialmente en ti, aunque sea un paso a la vez, como en Bunny Girl Senpai. 'Nadie nace en este mundo para estar solo',

según One Piece.

AGRADECIMIENTOS

Agradezco a los docentes y directivos que me ofrecieron una guía inigualable, ayudando desde

lo más personal y humano.

Al Mag. Diego Iquira, quien me acompañó en gran parte del proceso, organizando las ideas

para plasmarlas en elementos factibles.

Al Dr. José Esquicha, con una gran facilidad para identificar áreas de mejora y brindar claridad ante un proyecto muy ambicioso.

RESUMEN

Las personas desean comunicarse y ser comprendidas, siendo esencial la expresión emocional, pero su uso implica un desafío de precisión, rapidez, interpretación y eficacia, que puede ser simplificado con Inteligencia Artificial (AI). Analizar las emociones a menudo se vincula con gran experiencia en el área o sistemas complejos multidisciplinarios, influenciables por factores contextuales, difíciles de comprender, en servicios poco adaptativos o muy limitados. El sistema busca detectar y explicar patrones emocionales, entre características numéricas de la voz y representaciones visuales Mel, usando modelos interpretativos que vinculan razonamiento intermedio (embeddings) en una arquitectura combinada de Redes Neuronales Convolucionales (Convolutional Neural Networks, CNN) y Transformadores (Transformers). El sistema propuesto ofrece precisión, explicabilidad y una solución viable, reproducible y de bajo costo computacional para el análisis emocional en voz, con potencial en educación, salud y tecnologías interactivas. Este estudio tecnológico y aplicado, con un enfoque exploratorio y explicativo, usa las decisiones simplificadas de árboles de decisión (decision trees, DT) junto a LassoCV (Least Absolute Shrinkage and Selection Operator cross-validation) para detectar patrones entre las bases de datos de entrenamiento y muestra. Se encontró que hasta el 87% de representaciones embeddings en la base de datos de CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) (41/42) y el 62% en EMO-STIM (Emotional Film Clips with Discrete and Componential Assessment) (59/68) podrían ser explicadas por características vocales. Los modelos interpretativos (surrogate models) lograron un coeficiente de determinación (R^2) de 0.76, un error cuadrático medio (MSE) de 0.00165 y su raíz (RMSE) de 0.041, correspondientes a una representación con 10 características, con fidelidad superior al 98%, y etiquetas emocionales con precisión del 99.7% usando arquitectura CNN básica, con soporte complementario de transformadores. El estudio demuestra que muchas decisiones del modelo se basan en una amplia gama de características vocales. Esto sugiere que tanto la expresión como la respuesta de emociones básicas no solo se apoya en las más convencionales o evidentes.

Palabras claves: Inteligencia artificial emocional, redes neuronales convolucionales, interpretabilidad

ABSTRACT

People seek to communicate and be understood, and emotional expression is essential to achieve this. However, using emotions in communication involves challenges of precision, speed, interpretation, and effectiveness that can be simplified through AI. Emotional analysis is often associated with expert knowledge or complex multidisciplinary systems influenced by contextual factors difficult to understand, in services that are poorly adaptable or highly limited. This system aims to detect and explain emotional patterns between numerical voice features and Mel visual representations, using interpretive models that link embeddings from a combined architecture of CNN and transformers. The proposed system offers accuracy, explainability, and a computationally efficient, viable, and reproducible solution for emotional voice analysis, with potential applications in education, healthcare, and interactive technologies. This applied and technological study, with an exploratory and explanatory approach, uses simplified decisions from DT and LassoCV to detect patterns across training clips and movie databases. Results show that up to 87% of the embeddings in the CREMA-D dataset (41/42) and 62% in EMO-STIM (59/68) could be explained through vocal features. Surrogate interpretive models achieved a theoretical R^2 above 0.76, MSE of 0.00165, and RMSE of 0.041, corresponding to a representation with 10 features, with over 98% fidelity and emotional labels accuracy of 99.7% using a basic CNN architecture, with complementary support from transformers. The study demonstrates that many of the model's decisions are based on a wide range of vocal features. This suggests that both the expression and response of basic emotions are supported not only by the most conventional or obvious cues.

Keywords: Emotional artificial intelligence, convolutional Neural Networks, interpretability

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

INTRODUCCIÓN

La voz humana es uno de los medios más poderosos para expresar emociones. Hoy en día, gracias a la AI, ya es posible detectar emociones a partir del sonido de la voz con modelos muy precisos, como las redes neuronales y los Transformers, logrando resultados de precisión superiores al 90 %. Sin embargo, operan mayormente como sistemas de caja negra, sin explicar claramente cómo llegan a sus conclusiones. Esta falta de interpretabilidad representa una barrera importante para su aplicación científica, ética y social.

Esta investigación nace del interés por detectar los patrones relacionados a nuestro estado emocional, más allá de la parte técnica. La motivación personal nace de una comprensión profunda, aunque tardía, del valor de la comunicación emocional. Expresar lo que sentimos con claridad no solo mejora nuestras relaciones, sino que también fortalece nuestra salud mental, reduce el aislamiento y nos permite conectar de forma más auténtica con los demás.

El problema que se aborda tiene impacto en múltiples contextos donde la comunicación emocional es clave para el desempeño: desde situaciones cotidianas hasta roles profesionales, cuyo desempeño se mide por su capacidad para comunicar eficazmente sus emociones y comprender las de los demás. La voz no solo transmite información, también refleja intención, empatía y credibilidad. Por eso, existe una necesidad técnica de hacer explicables los modelos de reconocimiento emocional, herramientas que permitan mejorar la comunicación y conocer si estas relaciones son estáticas o dinámicas.

A través de un enfoque que integra CNN, Transformer y técnicas de selección de características explicables como DT y métodos como LassoCV, se explora si es posible detectar patrones emocionales interpretables.

El objetivo no es solo la detección emocional, sino permitir que los resultados sean verificables y medibles, reduciendo su complejidad, en un lenguaje más comprensible, sin perder el poder analítico. Conociendo rutas más libres entre generaciones, profesiones y contextos, abriendo la posibilidad de intervenciones más justas, éticas y humanas.

En resumen, este trabajo busca tender un puente entre el avance tecnológico y la necesidad humana de ser entendidos. Puede convertirse en una herramienta útil para quienes tienen dificultades para expresarse emocionalmente, ya sea por inseguridad, discapacidad o falta de herramientas. Comprender las emociones a través de la voz es un paso necesario hacia una comunicación más empática, clara y real.

Para su desarrollo, el trabajo se organiza principalmente en cuatro capítulos interrelacionados y varias secciones no numeradas. El Capítulo I establece la base conceptual de la investigación: presenta el problema, los objetivos, el enfoque y la solución general propuesta. También define la línea de investigación y el camino lógico que se seguirá, permitiendo entender desde el inicio qué se hará, por qué y cómo se conecta cada parte del estudio. El Capítulo II desarrolla los antecedentes y fundamentos teóricos, incluyendo el estado del arte, y analiza enfoques previos relacionados con la AI emocional, para situar esta propuesta dentro del panorama actual. En el

Capítulo III se define el marco metodológico. Aunque algunos aspectos ya han sido introducidos, aquí se abordan desde una perspectiva más operativa y lógica, permitiendo traducir las ideas generales en decisiones concretas. Se precisa el tipo y nivel de investigación, los alcances, limitaciones, técnicas de recopilación de datos y el plan de trabajo, aportando una estructura sólida para la ejecución del estudio. El Capítulo IV presenta el desarrollo de la solución propuesta, siguiendo un flujo estructurado que inicia con la comprensión, preparación y representación de los datos, y culmina con la arquitectura híbrida del sistema, los modelos de reconocimiento emocional y los modelos explicativos surrogate que permiten interpretar los patrones detectados. En las secciones no numeradas posteriores, se exponen los resultados de forma directa, sin interpretaciones; la discusión crítica de dichos resultados, comparada con elementos críticos de otras investigaciones; se abordan hallazgos inesperados, limitaciones, implicancias éticas y su relación con estudios previos; se presentan las conclusiones, el valor tanto del sistema de reconocimiento como de los modelos interpretativos; se plantean las recomendaciones y trabajos futuros, orientados a ampliar el trabajo, considerando nuevos conjuntos de datos, distintos contextos de aplicación y mejoras metodológicas para una mayor generalización y comprensión de los patrones emocionales. Finalmente, se incluyen las referencias bibliográficas y los anexos que complementan la información presentada en los capítulos principales.

CAPÍTULO I

1 Planteamiento de la Investigación

1.1 Planteamiento del Problema

Hoy más que nunca, debemos enfocarnos en comunicarnos efectivamente. Esta habilidad nos da la capacidad de transmitir ideas claras (UNESCO, 2015), mientras que comunicarse emocionalmente permite que las interacciones sociales sean exitosas, productivas y con un ambiente saludable (Turković et al., 2022). Según la Organización Mundial de la Salud (OMS), con datos de la Organización Internacional del Trabajo (OIT), el 60 % de las personas están trabajando, afectadas de forma prevalente por una comunicación poco efectiva que contribuye al estrés y la ansiedad. De hecho, el 50 % del costo asociado a estas condiciones mentales se debe a la reducción de productividad, generando pérdidas de 1 billón de dólares a la economía mundial, vinculadas a la mala comunicación (World Health Organization, 2022). Por ello, la comunicación no solo depende de las palabras, sino de cómo lo decimos; estas características son claves para entender las emociones que estamos comunicando oralmente (Guyer et al., 2021). Esto implica analizar los desafíos de la comunicación, como la posibilidad de interferencias provocadas por factores externos que pueden afectar la precisión e interpretación del mensaje (Alhussein et al., 2025). Asimismo, es necesario considerar los parámetros acústicos más allá de los elementos dominantes que contribuyen a la expresión de las emociones (Liscombe, 2007).

A nivel mundial, la mala comunicación tiene repercusiones significativas. El sector empresarial tiene un 56 % de proyectos fallidos, principalmente causados por problemas de comunicación (Project Management Institute, 2013). En el ámbito de la salud, se ha reportado que, en 2013, los errores médicos causaron la muerte de entre 44,000 y 98,000 pacientes cada año, siendo imperativo explorar nuevas oportunidades mediante medidas preventivas factibles, que contribuyan tanto a pacientes como al bienestar de los trabajadores (Riehle et al., 2013). El 80 % de eventos adversos hospitalarios se asocian a traspasos inadecuados, y el 53 % de los eventos adversos críticos se deben a fallas en la comunicación, según el informe de Joint Commission Sentinel Event Database de 2004 a 2009 (Joint Commission International, 2021).

A nivel latinoamericano, estudios como el realizado en Colombia resaltan que el 40 % de los trabajadores que dependen de su voz como herramienta principal, como docentes y operadores telefónicos, enfrentan alteraciones en las cualidades vocales debido a factores como el estrés laboral, la falta de formación en técnicas y la sobrecarga vocal. Estas alteraciones pueden incluir cambios en el tono, la resonancia, la intensidad y la calidad de la voz, afectando tanto la comunicación efectiva como la salud a largo plazo. La investigación subraya la necesidad urgente de intervenciones focalizadas que no solo prevengan los trastornos vocales, sino que también promuevan un monitoreo adecuado de las características de la voz para detectar y

abordar posibles anomalías antes de que se conviertan en problemas mayores (Figueredo J.N. & Castillo J.A., 2016).

A nivel nacional, el estudio realizado en la Universidad Nacional Autónoma de Chota reveló que el 51,1 % de los estudiantes con un nivel alto de inteligencia emocional mostró una correlación positiva con el rendimiento académico (p = 0,021) (Idrogo Zamora & Asenjo-Alarcón, 2021). Estos hallazgos subrayan la importancia de desarrollar habilidades emocionales en el contexto educativo para manejar los desafíos académicos, contribuyendo a un mejor rendimiento. En Arequipa, estudios como el realizado en el Hospital Regional Julio Pinto Manrique revelaron que el 64,29 % de las enfermeras con satisfacción laboral moderada presentaron un adecuado control emocional, habilidad crucial en profesiones de alto contacto humano, que favorece no solo su bienestar personal, sino que mejora la calidad de la atención y la solución de situaciones (Vargas Mamani & Villanueva Villena, 2022).

1.2 Objetivos de la Investigación

1.2.1 Objetivo General

Desarrollar e implementar un sistema basado en arquitecturas combinadas de redes neuronales y clasificadores interpretativos para detectar y explicar patrones emocionales en la voz con escenas de películas.

1.2.2 Objetivos Específicos

1. Recolectar y preparar bases de datos de audio de escenas de películas para el entrenamiento y validación del sistema.
2. Extraer y seleccionar características vocales relevantes que permitan representar las emociones expresadas en el audio.
3. Diseñar e implementar una arquitectura híbrida de redes neuronales para el modelado de los patrones emocionales en la voz.
4. Aplicar y evaluar modelos interpretativos surrogate para explicar las decisiones del sistema sobre las emociones detectadas.
5. Analizar y validar el desempeño del sistema en términos de precisión, fidelidad y capacidad explicativa, identificando áreas de mejora.

1.3 Preguntas de Investigación

1. ¿Qué tan adecuadas son las bases de datos de escenas de películas para representar emociones auditivas de forma fiable en modelos de aprendizaje automático?

2. ¿Cuáles son las características vocales más relevantes para la detección de emociones y cómo se comportan en diferentes conjuntos de datos?
3. ¿Con qué precisión se generarán las etiquetas emocionales de la muestra, usando un modelo basado en redes neuronales y Transformers entrenado con espectrogramas Mel?
4. ¿Qué tan fuerte es la relación entre las características vocales y los embeddings emocionales, según el coeficiente de determinación R^2?
5. ¿Qué tan bien pueden modelos surrogate, como DT o LassoCV, imitar las decisiones del modelo emocional en términos de fidelity?

1.4 Línea y Sub-línea de Investigación

1.4.1 Línea

Inteligencia Artificial

1.4.2 Sublínea

Computación neuronal

1.5 Enfoque

Cuantitativo, se medirá la precisión del sistema de IA en el dataset de entrenamiento y patrones en la muestra. Además, en el árbol de decisión se obtendrá el rendimiento según la profundidad y los tramos temporales seleccionados, en modelos representativos o completos según el dataset.

1.6 Solución Propuesta

El sistema propuesto tiene como función principal generar un archivo de etiquetado emocional, a partir de archivos de audio, usando dataset comprobado para el análisis emocional basado en redes neuronales convolucionales con transformadores y la interpretación de estos patrones con DT.

1.6.1 Justificación e Importancia

La comunicación oral nos ayuda a exponer ideas claras y buscar soluciones en sociedad, es clave para mejorar nuestras relaciones sociales y profesionales, pero puede convertirse en un desafío cuando no captamos adecuadamente las emociones o características vocales de los demás, especialmente con diferencias generacionales, culturales o de idioma. Estas barreras pueden llevarnos a evitar conversaciones, generar tensiones o afectar la productividad laboral, la calidad del servicio al cliente y hasta la atención médica.

Los avances tecnológicos en AI han revolucionado el análisis de la comunicación; por ejemplo, para identificar patrones emocionales en la voz se han implementado técnicas como el espectrograma log-mel que, en combinación con algoritmos de IA no supervisados, han mostrado ser especialmente eficaces en identificar emociones (Alhussein et al., 2025).

Por otro lado, el análisis de la comunicación efectiva se basa en la identificación de características de la voz, integrándolas con herramientas precisas y adecuadas (Liscombe, 2007), como es el caso de OpenSmile, que utiliza diferentes conjuntos de elementos relevantes (Mustafa et al., 2024) para identificar la voz o las emociones.

Ante esta realidad, se propone una solución innovadora que busca la sencillez, con una mínima cantidad de elementos en un ambiente claro emocional y vocal, por la selección de variables, incorporando IA por sus capacidades.

La elección de CNN, Transformers, Lasso y árboles de decisión no fue arbitraria, sino el resultado de una selección metodológica orientada al equilibrio entre rendimiento, interpretabilidad y escalabilidad. Las redes neuronales convolucionales (CNN) destacan en tareas de reconocimiento por su capacidad para extraer características locales de los datos, ¡especialmente en dominios visuales y auditivos; sin embargo, presentan limitaciones en interpretabilidad, al funcionar como modelos de tipo caja negra.

Los Transformers, en cambio, permiten modelar relaciones globales mediante mecanismos de atención, ofreciendo una visión contextual del input y complementando las capacidades locales de las CNN.

El componente Lasso se incorporó como un método de selección automática de variables relevantes, que penaliza la complejidad y reduce la dimensionalidad del modelo, previniendo el sobreajuste. Esta propiedad es especialmente valiosa en contextos con muestras pequeñas o alta correlación entre variables.

Finalmente, los árboles de decisión (DT) se integraron no solo por su bajo costo computacional, sino por su capacidad de representar decisiones como reglas comprensibles, lo que aumenta la transparencia y la trazabilidad del sistema.

Esta combinación metodológica se diseñó para aprovechar las fortalezas complementarias de cada técnica, descartando de forma fundamentada otros modelos con resultados poco prometedores o teóricamente incompatibles. En conjunto, la propuesta constituye una arquitectura híbrida capaz de equilibrar precisión y explicabilidad, ofreciendo una solución robusta a los desafíos actuales del reconocimiento emocional, donde comprender el modelo es tan importante como su exactitud.

1.6.2 Descripción de la Solución

El sistema propuesto tiene como función principal generar un archivo de etiquetado emocional, a partir de archivos de audio, usando un dataset comprobado para el análisis emocional basado en redes neuronales convolucionales con transformadores.

Los módulos tienen como función principal reconocer, en tramos de tiempo de generación, los sets de características vocales, etiqueta, las dimensiones de transformación de audios a imágenes espectrales, su almacenamiento y los tramos de procesamiento por lote(batch).
Es fundamental verificar que se cumpla con la precisión en el modelo de IA, tener el formato correcto de los archivos y ajustar las métricas de generación de prueba para evitar un margen de error manual en la precisión, el tiempo de entrenamiento y los recursos computacionales.
Finalmente, estos serán agrupados para la interpretación de los patrones encontrados; como técnica principal se usarán los DT de clasificación, y según los resultados encontrados se seleccionarán uno o varios modelos.
El proceso tiene 2 etapas.

Etapa de Generación y Extracción de elementos básicos
Se selecciona un conjunto de audios etiquetados con emociones, con un set de características de sonido que identifiquen la voz, escogiendo las más relevantes según los perfiles o personalidades que se desean etiquetar, como se muestra en la Tabla 1, con el proceso de entrenamiento, prueba y ajuste del modelo con datos etiquetados.

Tabla 1
Flujo del procesamiento de audios

Inicio
Importar dataset de audios etiquetados con emociones
Establecer categorías emocionales y set de características
Definición de métricas, parámetros y configuraciones
Entrenamiento del modelo de reconocimiento emocional oral
Extracción de embeddings
Prueba del modelo con resultado de precisión
Fin

Fuente: Elaboración propia.

Se ingresa el audio experimental, o su representación visual o numérica, en el formato adecuado; el sistema lo procesará identificando la emoción. Como se ve en la Tabla 2, sobre el uso del sistema para el procesamiento de audios experimentales.
Tabla 2 *Flujo de preselección de subconjuntos*

Inicio
Cargadeclips
Selecciónde candidatospormodelospreliminares
Gráficodeprofundidadymuestreopormodelosinterpretativos
Registroderesultados,relacionesymétricas
Fin

Fuente: Elaboración propia.

Tras el proceso iterativo de selección, generación y análisis de las representaciones o audios, se incorpora un componente innovador basado en Lasso, concebido como un módulo de conexión orientado a mejorar la interpretabilidad del sistema. Esta integración surge a partir de los distintos resultados obtenidos en Orange y KNIME, utilizando modelos auxiliares PCA, FR, DBSCAN y DT con diferentes subgrupos de datos, y de la revisión de metodologías afines según su compatibilidad con los tipos de datos y clasificadores estándar.

La propuesta enfrenta retos característicos de los enfoques Post hoc, ya que CNN son modelos de tipo caja negra que, por su naturaleza, deben implementarse para poder ser estudiadas. Existen múltiples formas de analizar o interpretar sus capas internas, por lo que resulta casi indispensable apoyarse en modelos auxiliares complementarios que permitan seleccionar las métricas más adecuadas y definir rutas interpretativas coherentes.

El componente Lasso actúa como un puente metodológico que, de forma metafórica, replica el comportamiento de los transformadores, al prevenir el sobreajuste y penalizar la multidimensionalidad. Estas propiedades explican el incremento de precisión observado en la CNN, diferenciándola de otros modelos más convencionales.

Finalmente, tanto MultiTaskLasso como LassoCV permiten regular los parámetros de manera casi automática y sencilla, siendo el segundo más adecuado para datos básicos, categóricos o etiquetados. No obstante, es importante considerar el alto costo computacional asociado a estos procesos, donde la cinematografía adquiere relevancia: aunque su naturaleza es compleja, este tipo de datos se distingue por su eficiencia con muestras pequeñas pero ricas en contenido emocional, lo que, en conjunto con un modelo Ante hoc DT surrogate, permite alcanzar una interpretabilidad inherente del sistema.

1.6.3 Aporte

El principal aporte de esta investigación radica en el desarrollo de una arquitectura híbrida e interpretable para el reconocimiento de emociones en la voz, que integra la potencia de las CNN y los mecanismos de atención de los Transformers, con el componente de regulación y reducción de dimensionalidad de Lasso y la interpretabilidad inherente de los DT. Esta combinación permite no solo alcanzar una alta precisión en la detección de emociones vocales,

sino también superar las limitaciones de opacidad propias de los modelos de tipo caja negra, ofreciendo una solución equilibrada entre rendimiento, explicabilidad y eficiencia.

Recientes recopilaciones y revisiones en el campo de la interpretabilidad, explicabilidad, IA y XAI destacan la creciente necesidad en auge de metodologías explicativas, especialmente en dominios monomodales como el reconocimiento emocional en audio. Estas investigaciones evidencian la escasez de estudios que logren combinar alta precisión con interpretabilidad, y recomiendan desarrollar enfoques más transparentes, trazables y adaptados al tipo de datos para cerrar esta brecha.

En este contexto, el sistema de IA propuesto identifica patrones vocales asociados a distintas emociones, aportando a una mejor comprensión de la expresión emocional y la comunicación humana, y reduciendo la brecha contextual oral. Además, abre nuevas posibilidades para investigaciones futuras en análisis afectivo, detección de patrones e interpretabilidad basada en enfoques XAI, consolidándose como una contribución metodológica y práctica en el campo del reconocimiento emocional.

CAPÍTULO II

2 Fundamentos teóricos

2.1 Estado del Arte

El estudio de la comunicación con IA toma un papel cada vez más importante en las interacciones, al cambiar la forma en que nos relacionamos en sectores clave, influyentes en nuestro trabajo, salud y relaciones personales. Entre sus avances está la comodidad de transmitir mensajes en menor tiempo, que pueden ajustarse y permiten el acceso entre dificultades. Sin embargo, la tecnología requiere comprender el contexto, ser cuidadosa y responsable al manejar grandes cantidades de datos para evitar ser poco empática, injusta o invasiva.

Para esta revisión, se seleccionaron estudios de los últimos cinco años relacionados con la comunicación efectiva o emocional, en los cuales se utiliza IA, con énfasis en CNN y transformers. Estos estudios se centran en aspectos clave del habla, como la prosodia, que se refiere a las características relacionadas con el volumen, tono y ritmo de la voz, así como su análisis a través del espectrograma.

2.1.1 IA Aplicada en la comunicación

La IA entiende el contexto en la comunicación gracias a su capacidad de adaptación y precisión emocional.

El objetivo de este artículo es optimizar el servicio de recepción de llamadas de emergencias en Rumanía sin comprometer significativamente su sistema de precisión emocional, que se logró al identificar y agregar el estado 'irritado', con un 91.82% de precisión ponderada (Marghescu et al., 2023a). Pocos estudios sobre los servicios de emergencia han considerado preguntarse cuál es el componente crucial para dar respuestas rápidas y efectivas, como mostró este estudio, que destacó el reconocimiento emocional en la voz (Marghescu et al., 2023b).

Sin embargo, pocos estudios han considerado determinar en vivo las intenciones o emociones con subtítulos, como lo hace un enfoque con tecnologías de realidad virtual, características acústicas y reconocimiento del habla, en una clasificación multicapa (Ubur & Gracanin, 2025). El estudio destinado a comprender las emociones humanas en el discurso constituye un nicho crítico que aún necesita ser abordado (Bhanbhro et al., 2025). La importancia de este estudio radica en su potencial de adaptabilidad en diferentes idiomas ricos en fonética, como el Punjabi, logrando precisión emocional usando redes neuronales convolucionales, un avance significativo en la comunicación efectiva dentro de comunidades específicas (Sharma et al., 2023).

2.1.2 AI comunicativa y emocional

Las transformaciones en la interacción humano-computador por IA están presentes en muchas áreas, predominando la salud. Estas aplicaciones son producto del análisis de señales del habla, características fáciles o señales emocionales, abriendo puertas por ser más objetivos o menos invasivos que los métodos tradicionales.

El estudio de la precisión complementaria de la comunicación puede ser bastante complejo, como saber implementar armoniosamente los elementos significativos del habla o los aspectos faciales (Pan et al., 2024). La importancia de este tema radica en ofrecer experiencias más naturales, como las ofrecidas en redes neuronales profundas mediante la integración de señales emocionales (Vardhan et al., 2024).

Estudios recientes sobre tecnologías de reconocimiento emocional han mostrado un gran potencial para su aplicabilidad en diversas áreas, particularmente en interacciones digitales y contextos sociales, donde la detección de emociones se apoya en señales objetivas menos invasivas y más precisas que los métodos tradicionales (Guo et al., 2024).

En los últimos años, ha surgido un creciente interés por métodos más objetivos y automatizados para el reconocimiento de emociones, utilizar características acústicas, lingüísticas y temporales de los audios, así como técnicas de aprendizaje automático y profundo, que pueden identificar patrones en la detección temprana del Trastorno de Estrés Postraumático(TEPT), con un enfoque que no consume tanto tiempo como el diagnostico tradicional (Islam & ElSayed, 2024). Sin embargo, pocos estudios han considerado, para diagnósticos tempranos, apoyarse en modelos entrenados con datos emocionales y musicales (Josephine Mary Juliana et al., 2023).

2.1.3 Análisis de características emocionales y vocales

El análisis de las características del habla influye en la comunicación; ya sea analizando las palabras o la forma como se dicen, tienden a usar técnicas de IA en redes neuronales junto con métodos avanzados para identificar emociones complejas. Además, se encuentran mayormente en servicios de asistentes virtuales y atención al cliente.

Este estudio propone priorizar técnicas según su uso, como ser a largo plazo, con LSTM, o si es suficiente con un reconocimiento inicial, en CNN, aunque esta tenga menor precisión. Incluye restricciones de tamaño, diversidad y matices emocionales complejos por la variabilidad individual (Wani et al., 2020).

El enfoque adoptado en estos estudios busca superar las limitaciones de los enfoques tradicionales de la comunicación oral, que se centran en el contenido verbal, para poner énfasis en características de la voz que afectan la interpretación y evaluación del mensaje (Guyer et al., 2021), así como la percepción de confianza y credibilidad del emisor (Guyer et al., 2021). Este trabajo propone una solución que integra dos análisis de sonido de espectrograma Mel, el clásico e inverso, con una precisión emocional de 94.79%, pero limitada por la poca variedad de datos entrenados y la complejidad computacional (J. Li et al., 2022). Este artículo presenta un enfoque

en las redes de cápsulas frente al tradicional CNN, siendo prometedor para aplicaciones avanzadas en emociones complejas y mixtas del habla (Trinh Van et al., 2022).
Pocos estudios han explorado la aplicabilidad de métodos como Multi-Layer Perceptron (MLP) y Random Forest (RF) para el aprendizaje automático en la comunicación, que explora su enfoque técnico para superar limitaciones en enfoques prácticos como diálogos y servicio al cliente (Arun et al., 2021).

2.1.4 Uso de espectrograma en AI

El uso de espectrogramas como representación visual del contenido frecuencial del audio ha cobrado gran relevancia en el campo de la AI, particularmente en aplicaciones de procesamiento del habla y reconocimiento emocional.
Sin embargo, persisten limitaciones debido al alto consumo de recursos computacionales por parte de arquitecturas transformers que modelen matices emocionales complejas y duraderas (H. Li et al., 2024).
También existen dificultades para capturar con precisión la señal y las emociones en ambientes dinámicos y en diferentes tramos de tiempo (Kawade et al., 2022).
No obstante, investigaciones innovadoras con un framework de Deep Learning (DL) para la detección emocional pueden reducir el trabajo manual y de diseño, ofreciendo adaptabilidad y eficiencia en la representación de información temporal y frecuencias de audio (Khalil et al., 2019).
A pesar de los avances en sistemas no invasivos, sigue siendo un reto generalizar la efectividad del uso de escalas cuasi-logarítmicas en espectrogramas de Mel, donde el rendimiento varía según la selección de características vocales y el contexto de aplicación (Harar et al., 2018). Se subraya el potencial de la música para superar estas brechas, apoyándose en algoritmos de aprendizaje profundo para capturar dimensiones emocionales más ricas (Lu & Hao, 2023).

2.1.5 Redes neuronales convolucionales junto con clasificadores

La AI está impulsada por datos, lo que requiere la colaboración entre humanos y máquinas para la creación de algoritmos y la evaluación de resultados. Este enfoque, que integra conceptos estadísticos y la intervención humana, es fundamental para mejorar la interpretabilidad y la reproducibilidad de los modelos de IA (Yu & Kumbier, 2018).
Se argumenta que las redes neuronales convolucionales son eficaces en la clasificación de imágenes, pero a menudo carecen de explicaciones claras sobre cómo toman sus decisiones. Este desafío se aborda mediante el uso de DT, que descomponen las representaciones de características de las CNN y proporcionan explicaciones semánticas sobre qué partes de un objeto activan los filtros y cómo contribuyen a la predicción (Zhang et al., 2019).
La capacidad de las arquitecturas CNN, una forma especializada de redes neuronales artificiales (Artificial Neural Network, ANN), reside en su diseño para procesar imágenes y extraer

automáticamente sus características, lo que las hace ideales para tareas de visión por computadora (Elngar et al., 2021). Estas redes se utilizan, por ejemplo, en la identificación de pacientes con autismo, donde resultan especialmente útiles en el diagnóstico médico; mediante la clasificación de patrones complejos y el análisis de la conectividad entre diferentes regiones cerebrales, se ha alcanzado hasta un 80% de precisión (Kashef, 2022).

2.2 Bases Teóricas de la Investigación

La creciente variedad de parámetros acústicos ha llevado a la utilización inconsistente de conjuntos de datos, dificultando la comparación entre estudios. Este artículo propone un conjunto de características acústicas para el análisis de voz, utilizando openSMILE, que permite generalizar mejor la información, integrar sistemas y lograr un alto rendimiento, incluso con un tamaño reducido (Eyben et al., 2016).

En el estudio de emociones aplicadas a AI, las bases de datos son fundamentales para entrenar y evaluar modelos capaces de reconocer emociones en la voz. Las más útiles son aquellas con variedad de participantes, emociones claramente definidas y evaluaciones realizadas por personas.

El dataset CREMA-D es uno de los más usados para analizar la expresión emocional vocal. Contiene 7,442 clips de voz y video, grabados por 91 actores de distintas etnias que interpretan frases neutras con seis emociones básicas (alegría, tristeza, miedo, ira, asco y neutralidad). Más de 2,400 personas evaluaron cada clip, indicando la emoción percibida y su intensidad. Aunque incluye video, destaca por la información contenida solo en la voz, alcanzando un 40.9% de reconocimiento emocional sin señales visuales (Cao et al., 2014).

Por su parte, EmoStim ofrece un enfoque diferente basado en la inducción emocional. Incluye 99 fragmentos de películas seleccionados por su capacidad de evocar emociones específicas. Fueron evaluados por 638 personas a través de la plataforma CrowdFlower. Su valor principal es su alta validez ecológica: las emociones reflejan reacciones más naturales y contextuales, como las que se experimentan en la vida diaria (Somarathna et al., 2024).

Ambos datasets han sido validados por múltiples participantes y cubren emociones básicas, enfocándose en la expresión e inducción emocional, respectivamente.

Una revisión sistemática sobre la efectividad de los DT en tareas de clasificación destaca su rol como herramienta de interpretabilidad en sistemas complejos como las CNN. Dado que las CNN son altamente efectivas en tareas de clasificación, pero carecen de transparencia, la incorporación de DT como CART para una interpretabilidad mejorada es crucial en aplicaciones en campos críticos. Las técnicas explicativas como CART pueden ayudar a mejorar la comprensión de los modelos de clasificación, promoviendo el desarrollo de AI explicable (XAI) (Zaman & Hassan, 2021).

En el campo del diagnóstico médico, los DT presentan técnicas para representar y clasificar datos con simplicidad y efectividad, revisando diferentes algoritmos de DT (Charbuty &

Abdulazeez, 2021), proporcionando una manera comprensible de desglosar las predicciones hechas por modelos complejos como las CNN, lo que es clave en la búsqueda de métodos que mejoren la interpretabilidad en una variedad de contextos.

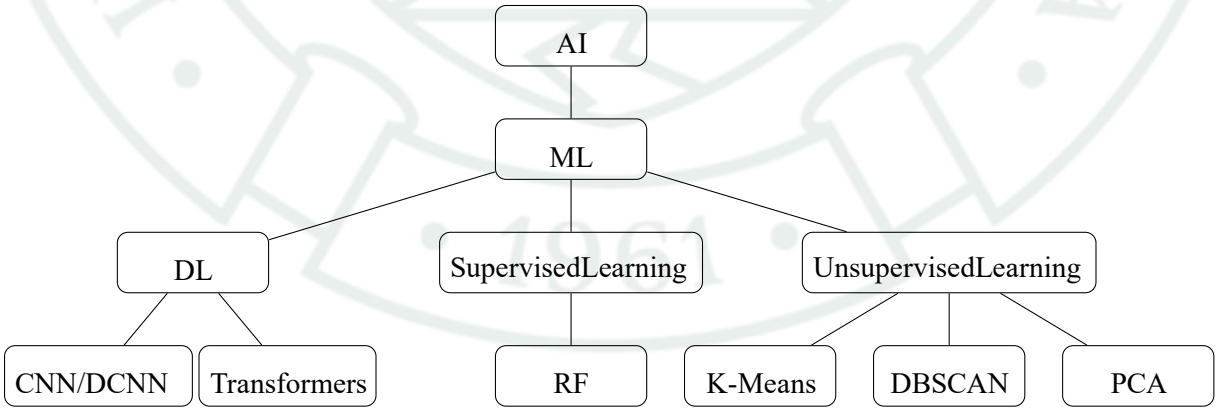
Fundamentos de la AI

La AI o máquina pensante es un sistema que da una solución práctica a problemas que en este momento necesitan de inteligencia humana para resolverse, donde la categoría seleccionada permite prescindir de elementos como etiquetado, entrenamiento o supervisión humana. Una de las ramas con mayor atención e inversión es el aprendizaje de máquina (Machine Learning, ML), debido a su factibilidad para manejar grandes volúmenes de datos en sectores digitales y su costo computacional asequible, lo que la lleva a rivalizar con especialistas a un nivel de producción alto. ML se introduce en nuestra vida cotidiana transformando entradas x en salidas esperadas, de modo que $y = f(x)$ (Landgraf, 2021).

La inteligencia y emociones funcionan efectivamente cuando se complementan, en una conexion clara, un proceso de empatizar, mas alla de construir pensamientos logicos necesitamos los sentimentales en su mision de evaluar la realidad (Oritsegbemi, 2023).

La Figura 1 muestra cómo se organizan los principales tipos de algoritmos en inteligencia artificial. El DL se ubica aparte de los métodos supervisados y no supervisados porque puede aplicarse en ambos tipos de aprendizaje. Dentro del aprendizaje no supervisado, aparecen herramientas como Density-Based Spatial Clustering of Applications with Noise (DBSCAN), para agrupar datos sin etiquetas, y Principal Component Analysis (PCA), que ayuda a simplificar datos complejos. Aunque no son el foco de este trabajo, estos modelos sirven como referencia para guiar la investigación hacia los algoritmos más adecuados y efectivos según la tarea planteada.

Figura 1 *Jerarquía de algoritmos de AI, ML y DL utilizados*



Fuente: Elaboración propia.

En la Figura 1, los DL pueden ser hibridos, combinando enfoques supervisados y no supervisados.

Fundamentos de las emociones y prosodia

La emoción está en la base de todo comportamiento y de todo proceso cognitivo. Las transiciones emocionales nos permiten reconocer los cambios en nuestro "dominio de acción", ya sea en la amistad, la familia o en los distintos grupos sociales. De este modo, las emociones compartidas en la interacción social configuran nuestra identidad colectiva. Finalmente, nos identificamos con grupos ontológicos, a través de las emociones con las que nos reconocemos, y con marcos formales, a través de la razón que los establece (Totaro, 2021).

En la comunicación diaria, las emociones surgen de pequeños cambios o rupturas en lo esperado, lo que muestra cómo las personas adaptan su conducta, ajustan su expresión y comprensión, y necesitan métodos que preserven la naturaleza de la situación (Käsermann et al., 2000).

La expresión emocional con la voz no cambia únicamente por un solo factor; es una combinación que puede incluir el nivel de activación (arousal), el hecho de que sea positiva o negativa, o que implique mayor o menor control (Goudbeek & Scherer, 2010).

La prosodia, a través de características del habla, revela información sobre el estado interno del hablante y modula cómo se percibe el mensaje, cambios prosódicos que pueden indicar estados paralingüísticos, asociados a emociones; pragmáticos, que señalan intenciones comunicativas como preguntas; y de competencia lingüística, reflejando dominio del idioma y acentos, rasgos suprasegmentales que permiten discriminar emociones, interpretar intenciones y evaluar la fluidez del hablante, mostrando cómo la prosodia codifica información clave para la comunicación (Liscombe, 2007).

CAPÍTULO III

3 Marco metodológico

3.1 Alcances y limitaciones

3.1.1 Alcances

El presente estudio inicia de forma exploratoria, abarca el desarrollo e implementación de un sistema de AI orientada a la detección y explicación de emociones básicas a partir de señales acústicas de la voz. El alcance se delimita al análisis de características prosódicas, como tono, intensidad y ritmo, sin considerar el contenido semántico ni el idioma de las expresiones. Al manejar bases de datos validadas y comprobadas masivamente, se amplía la extensión y aplicabilidad a contextos culturales más amplios.

El sistema se construye y evalúa a partir de dos bases de datos con enfoques complementarios: CREMA-D, centrada en la emoción expresada por el hablante, y EMO-STIM, orientada a la percepción de emociones en escenas de películas por parte del espectador. De esta manera, el proyecto integra tanto la dimensión expresiva como la perceptiva, permitiendo identificar patrones significativos en la voz y validar su consistencia en contextos más complejos. El modelo diseñado se caracteriza por su simplicidad, bajo costo computacional y flexibilidad para ser replicado en escenarios particulares comprobados, como educación, salud, monitoreo de voz o sistemas de interacción humano-máquina.

La generalización del patron principal encontrado entre los modelos con mejor equilibrio en desempeño y fidelidad, parecen compartir alta compatibilidad para explicarse con caracteristicas vocales.

3.1.2 Limitaciones

Una de las principales limitaciones de este estudio es la imposibilidad de controlar el sistema en tiempo real, debido a consideraciones éticas sobre el almacenamiento temporal de datos, aunque se utilizan datos más “contextualizados” que datasets de reconocimiento emocional, todavía persiste cierta limitación de naturalidad, ya que la metodología aún no se encuentra comprobada en contextos sin validación previa de análisis emocional.

Los resultados carecen de una validación perceptual directa que confirme su fiabilidad más allá del análisis automatizado, lo que restringe su aplicabilidad inmediata en entornos prácticos. El uso de ciertas escenas de películas clásicas presenta inconvenientes, ¡ya que algunos clips incluyen sonidos comunicativos no verbales como quejidos, llantos, risas; ruidos de animales o elementos fantásticos, además de la repetición de fragmentos dentro de una misma película, lo que podría introducir cierto ruido.

El empleo de redes neuronales profundas como CNN y Transformers dificulta la comprensión directa de las decisiones internas del modelo, mientras que la segmentación manual de clips sería demasiado compleja y poco reproducible, por lo que se optó por duraciones fijas. Sin embargo, no se consideraron los espacios en blanco entre fragmentos, los cuales podrían tener un efecto distinto, especialmente en tramos largos, dado que la base de entrenamiento estaba diseñada para resultados en intervalos cortos.

Dado que el estudio propone un enfoque novedoso y simplificado para la evaluación de emociones, algunos factores relacionados con la duración de los fragmentos, la posible mezcla de emociones o las variaciones culturales y generacionales podrían no estar completamente considerados, lo que limita la generalización de los resultados a otros contextos y escenarios. Para superar los desafíos de rendimiento y generalización en muestras o bases de datos que usan ML, es necesario saber que es un tema en auge. Algunas tecnologías, como la realidad mixta (XR), carecen de investigaciones previas, por lo que los estudios sobre computación afectiva y su interacción e interpretación necesitan nuevas perspectivas para poder confiar y adaptar estas ’cajas negras’ a nuestros objetivos (Afzal et al., 2024).

Aunque LassoCV se evalúa con fidelidad, en investigaciones exploratorias no hay suficientes recopilaciones que identifiquen el desempeño de este modelo en el área, si bien no suele relacionarse a indicadores de sobreajuste, la complejidad computacional de trabajar y analizar cajas negras para evaluar qué tan adecuada es para el tipo de dato para modelos posteriores o previos es incierta. Futuras direcciones de la IA explicable (Explainable AI; XAI) consideran esencial identificar nuevos métodos en computación afectiva, junto con la importancia de aplicarse en modelos basados en audio. Si bien interpretar y explicar se usan de forma intercambiable en este ámbito, el primero se asocia a objetivos más amplios. Es fundamental reconocer, según las características o modelos, su nivel post hoc, relacionado con la generación de cajas negras, o ante hoc, con los inherentemente explicables como los DT, resaltando que estos últimos, aunque transparentes, sufren de capacidades limitadas y no siempre logran capturar la complejidad (Johnson et al., 2025).

3.1.3 Tipo y Nivel de investigación

De acuerdo con las características de este estudio, no se formula una hipótesis específica, ni se definen de manera conceptual u operacional las variables, ya que los patrones entre las variables acústicas y las emociones no se consideran hechos concluyentes. El enfoque principal de la investigación es exploratorio, orientado a detectar fenómenos y generar hallazgos iniciales que sirvan como base para estudios posteriores y secundario el explicativo.

En cuanto a la conveniencia de formular hipótesis, la literatura señala que no todas las investigaciones cuantitativas requieren plantearla. En los estudios exploratorios, no se propone hipótesis, y en las explicativas, se puede si su propósito es pronosticar cifras o hechos; Asimismo, se debe tener en cuenta que las variables deben estar siempre vinculadas

directamente con la hipótesis; y que, en ausencia de una hipótesis causal, indicar variables independientes y dependientes se considera un error metodológico (González Mares, 2019).

Según las características de la investigación no se planteará una hipótesis, ni precisar o definir conceptual u operacionalmente sus variables. Los patrones en la relación de variables acústicas y emociones no se consideran hechos concluyentes, ya que el enfoque exploratorio del estudio se orienta en detectar fenómenos y generar hallazgos iniciales.

El análisis de la conveniencia para formular hipótesis menciona que no todas las investigaciones cuantitativas plantean una; precisamente en el caso del alcance exploratorio no se formula; con el explicativo, este se le podría considerar si en los alcances iniciales del estudio intentaran pronosticar cifras o hechos. Por otro lado, las variables de forma indispensable deben aclarar su relación con la hipótesis, además exceptuando tener una hipótesis causal, indicar las variables independientes y dependientes es considerado un error (González Mares, 2019).

Tipo de investigación

La investigación aplicada se orienta a adelantos y productos tecnológicos. Además, los planteamientos cuantitativos con esencia exploratoria se dirigen a propósitos como la exploración de fenómenos o variables (González Mares, 2019).

Este estudio se clasifica como aplicado y tecnológico, ya que desarrolla un sistema funcional para detectar y explicar emociones en la voz, aplicando conocimiento técnico a un problema concreto. Se utilizan conjuntos de datos seleccionados bajo criterios de calidad y diversidad, en un entorno controlado y evaluado mediante métricas como precisión, fidelidad y coeficiente de determinación. Estas características responden a finalidades señaladas por Sampieri y Lester, como evaluar, interpretar, comparar y generar resultados transferibles a futuras aplicaciones.

Nivel de investigación

Una investigación cuantitativa puede tener diferentes alcances, entre ellos el exploratorio y el explicativo. El alcance exploratorio se aplica cuando el fenómeno aún no ha sido suficientemente estudiado y se requiere recolectar información inicial para establecer patrones preliminares. Por su parte, el alcance explicativo busca determinar las causas y efectos entre variables, explicando por qué ocurren ciertos fenómenos (González Mares, 2019).

Este estudio adopta un enfoque cuantitativo con un alcance tanto exploratorio como explicativo. Principalmente exploratorio porque los referentes y marcos teóricos aún no permiten formular con claridad hipótesis; por consiguiente, aborda una combinación poco estudiada: el uso de inteligencia artificial y modelos interpretativos para el reconocimiento emocional en la voz. A través del análisis de clips de películas, se busca observar y validar patrones emocionales representativos. Al mismo tiempo, el estudio es explicativo porque emplea una base de datos de entrenamiento estructurada, con etiquetas emocionales controladas, que permite analizar cómo las decisiones emocionales son tomadas por el sistema. Esto posibilita examinar relaciones entre

variables acústicas, influyen en categorías emocionales y respuestas emocionales, con métricas como la fidelidad y precisión del modelo.

3.2 Población

La población de este estudio se centra en escenas de películas, un medio ideal para el análisis de la voz por su diversidad y riqueza emocional. Sin embargo, podría ser aplicable a cualquier contexto en el que la voz humana sea fundamental para la comunicación oral, como en atención al cliente, discursos, entrevistas y otros momentos en los que las emociones juegan un papel clave.

3.3 Muestra

La muestra de esta investigación se seleccionó mediante un muestreo no probabilístico e intencional, basado en criterios técnicos y de viabilidad, tales como la compatibilidad en categorías emocionales básicas de respuesta frente a la expresada, el uso validado para el análisis emocional en recursos cinematográficos, tramos de división en clips principalmente en inglés y accesibilidad a los recursos.

El objetivo principal es detectar, en películas y bajo estos criterios de compatibilidad, patrones consistentes con la base de entrenamiento, utilizando características acústicas y embeddings relacionados. Se aprovecharán las etiquetas validadas, ya sean las expresadas en la base de entrenamiento o las percibidas en las películas, mientras que la base de datos restante servirá para corroborar la semejanza de los comportamientos.

Es importante precisar que no se aplicó una fórmula estadística de muestra finita ni un muestreo probabilístico clásico, debido a que la población de referencia, escenas de películas con predominancia de voz y carga emocional, no constituye un universo homogéneo exacto en la clasificación emocional, requiriendo filtros orientados a los objetivos investigativos. El número de clips y películas a considerar se estableció a partir de antecedentes empíricos.

Algunos estudios tienden a obtener mejores resultados con muestras pequeñas y controladas bajo criterios de investigación. Por ejemplo, comenzando con 437 escenas recomendadas por expertos, solo 70 cumplían con los criterios principales y únicamente 28 presentaban la diversidad emocional esperada (Michelini et al., 2019). Otro trabajo validó una batería de 57 películas para la inducción de emociones básicas, por diferenciación, analizando su capacidad de activación emocional (Fernández Megías et al., 2011). Esta misma batería fue utilizada posteriormente para analizar la valencia emocional positiva y negativa, evaluando la reactividad emocional en función del tiempo y el sexo, aunque con grupos de tan solo 10 escenas (Michelini et al., 2015). El análisis se centrará en una selección cercana a 20 escenas, extraídas de un mínimo de 8 películas. Bajo estas condiciones, EMO-STIM se redujo de 99 a 42 clips, que se organizaron de forma alfabética en dos grupos iguales A y B, como se muestra en la Tabla 3, trabajando el análisis final con el grupo A.

Tabla 3 *Distribución de escenas en los subconjuntos A y B*

Muestra A	Muestra B
12YearsASlave-clip-1.wav	LifeIsBeautiful-clip-2.wav
28DaysLater-clip-3.wav	LoveActually-clip-1.wav
28DaysLater-clip-5.wav	LoveActually-clip-2.wav
28DaysLater-clip-6.wav	LoveActually-clip-3.wav
28DaysLater-clip-7.wav	LoveActually-clip-4.wav
AmericanHistoryX-clip-1.wav	LoveActually-clip-6.wav
APerfectWorld-clip-1.wav	LoveActually-clip-7.wav
Bambi-clip-2.wav	Misery-clip-1.wav
BatmanReturns(1992)-clip-2.wav	MyGirl-clip-1.wav
CryFreedom-clip-2.wav	Philadelphia.wav
DangerousMinds.wav	RememberTheTitans-clip-1.wav
DeadManWalking-clip-1.wav	Scream1-clip-1.wav
ET-clip1.wav	Seven-clip-1.wav
ForrestGump-clip-1.wav	ShawshankRedemption-clip-1.wav
HotelRwanda-clip-1.wav	TheChamp-clip-2.wav
HotelRwanda-clip-3.wav	TheChamp-clip-3.wav
HotelRwanda-clip-4.wav	TheDeparted-clip-1.wav
HotelRwanda-clip-7.wav	ThePianist-clip-6.wav
HotelRwanda-clip-8.wav	TheShining-clip-3.wav
InTheNameOfTheFather-clip-1.wav	Trainspotting-clip-1.wav
KillBill1-clip-2.wav	Whenamanlovesawoman-clip-1.wav

Fuente: Elaboración propia.

3.4 Técnicas y Procedimientos para la Recolección de Datos

3.4.1 Técnica

El análisis de audio en películas ha demostrado ser una herramienta eficaz para estudiar la expresión emocional sin requerir la intervención directa de participantes. Las películas ofrecen datos de alta calidad para analizar patrones emocionales en el habla, permitiendo su evaluación y entrenamiento en distintos modelos. En investigaciones previas, se ha utilizado un enfoque de estudio de caso basado en escenas de películas seleccionadas para abarcar la aplicabilidad de diversas emociones, eventos y contextos (Zlatintsi et al., 2017).

Se adopta un enfoque de estudio de caso basado en el análisis de audio de escenas de películas validadas, utilizando herramientas de procesamiento de audio para detectar patrones en las características de la voz. Para este proyecto se emplearon herramientas en Python para procesar audio y entrenar modelos de AI. Se usaron librosa y openSMILE para extraer características de la voz, como espectrogramas y parámetros acústicos. El modelo fue desarrollado con PyTorch, utilizando redes neuronales convolucionales y transformers. Para evaluar su desempeño se usó scikit-learn, y los resultados se graficaron con matplotlib y seaborn.

3.4.2 Procedimiento

Dentro de los estudios sobre respuestas emocionales, la selección de películas no asume que las emociones complejas necesariamente estén presentes junto a las emociones básicas esperadas, ya que en un conjunto de 16 películas que provocaron exitosamente emociones, no se encontraron en todos los casos las emociones esperadas, recomendándose según el estudio analizarlas por separado (Gross & Levenson, 1995).

Mientras que algunas investigaciones utilizan un gran número de películas para conocer las emociones experimentadas, en este caso solo se usaron tres para representar la tristeza, la neutralidad y la diversión. Bajo la condición de expresar o suprimir emociones, los efectos fueron evidentes en estas películas emocionales (Gross & Levenson, 1997).

El procedimiento sigue un enfoque de optimización que busca minimizar la necesidad de grandes volúmenes de datos de entrenamiento y reducir la carga computacional mediante la transformación de datos y el manejo selectivo del almacenamiento de los elementos temporales, logrando un balance adecuado entre rendimiento y eficiencia, y facilitando la selección de variables relevantes.

Se empleará un enfoque progresivo. Inicialmente, se entrenará un modelo CNN+Transformer para la clasificación emocional de espectrogramas, aprovechando la extracción automática de etiquetas para las muestras. Posteriormente, se utilizarán DT para analizar los embeddings generados, y se aplicará regresión LassoCV para vincularlos con características acústicas extraídas del audio, con el fin de detectar patrones relevantes.

3.4.3 Instrumento

La investigación sobre la extracción de características acústicas ha dado lugar a diversas soluciones basadas en algoritmos. OpenSMILE destaca como una herramienta unificadora, capaz de reconocer y filtrar ruido durante el procesamiento, garantizando un rendimiento confiable y una adaptabilidad a las necesidades de diferentes investigaciones, gracias a su modularidad (Eyben et al., 2010).

En estudios previos, se observó que una clasificación para la identificación automática del habla en euskera alcanzó una precisión del 92.3% utilizando solo los 6 mejores elementos, frente a un total de 86 componentes o 512 subcomponentes, logrando una precisión máxima del 98.4% (Luengo et al., 2005).

Se empleará esta herramienta OpenSMILE, que ha demostrado ser altamente efectiva en diversas investigaciones debido a su robustez y flexibilidad para la extracción de características acústicas en la voz, utilizando modelos que identifican la voz y se relacionan con emociones, como eGeMAPS.

3.4.4 Herramienta

Se utilizarán los cuadernos de Google Colab para la ejecución del código. Este servicio en la nube, basado en máquinas virtuales, ofrece opciones gratuitas y de pago con acceso a recursos computacionales, como GPU y TPU, aunque con ciertas restricciones. Es ampliamente utilizado en educación e investigación para el desarrollo de modelos de AI, principalmente mediante el entorno Jupyter y el lenguaje Python, aprovechando recursos compartidos como el almacenamiento en Google Cloud. Resulta especialmente útil para almacenar y recuperar archivos en segmentos, con gran compatibilidad con librerías especializadas.

Como apoyo al procesamiento programado, se utilizaron herramientas visuales como Orange y KNIME. Estas plataformas permiten analizar datos conectando bloques (nodos) sin necesidad de programar. Orange destaca por su simplicidad y rapidez para probar modelos y ver resultados al instante, mientras que KNIME permite construir flujos más complejos y combinar múltiples fuentes de datos, son entornos efectivos para integrar múltiples técnicas de procesamiento de datos en etapas exploratorias o comparativas de forma visual e intuitiva.

3.5 Análisis de los datos

El análisis de los datos se realizará bajo un enfoque cuantitativo, permitiendo comparar los resultados obtenidos en el entrenamiento con las características vocales extraídas de los clips de películas.

CAPÍTULO IV

4 Desarrollo de la Propuesta de Investigación

4.1 Introducción al Flujo del Sistema

La arquitectura del sistema propuesto se organiza en una secuencia modular de componentes interdependientes, diseñados para implementar el enfoque progresivo descrito en el procedimiento metodológico. Cada módulo desempeña un rol específico en la transformación, clasificación e interpretación de señales de audio con contenido emocional.

Gracias a esta estructura modular, es posible incorporar mecanismos de selección adaptativa que permiten identificar, a lo largo del flujo, configuraciones óptimas para el procesamiento de datos. Por ejemplo, se evaluaron diferentes combinaciones de clips, ventanas temporales y subconjuntos de escenas (agrupadas según contenido vocal), generando múltiples versiones del dataset base. Esto habilitó decisiones informadas sobre cuál configuración ofrecía mayor estabilidad, riqueza emocional o claridad interpretativa.

El flujo inicia con la preparación de los datos, donde las señales de voz extraídas de escenas fílmicas son normalizadas y convertidas en representaciones visuales (espectrogramas de Mel) y vectores acústicos. Estas representaciones son procesadas por un modelo compuesto por una red CNN seguida de un Transformer, cuya salida corresponde a una clasificación emocional automática y embeddings latentes.

A partir de estos embeddings, se habilita una segunda etapa interpretativa. Modelos de árbol de decisión se utilizan para explorar las reglas implícitas que guían la clasificación, mientras que la regresión LassoCV permite vincular variables acústicas específicas con componentes de los embeddings, generando así un puente entre el modelo de caja negra y su explicación transparente.

Durante el desarrollo, el flujo permitió tomar decisiones iterativas fundamentadas en métricas como la desviación estándar emocional por grupo, la fidelidad del etiquetado y la estabilidad de los embeddings. Esto facilitó la refinación del conjunto de datos efectivo, reduciendo la complejidad sin sacrificar riqueza emocional, y favoreciendo una mayor reproducibilidad del modelo. Esta arquitectura modular permite separar claramente las etapas de etiquetado automático, interpretación simbólica y selección de variables relevantes, manteniendo la eficiencia del sistema y facilitando su análisis posterior. En las siguientes secciones se detallan cada uno de estos módulos, sus estructuras internas y los criterios utilizados para su implementación.

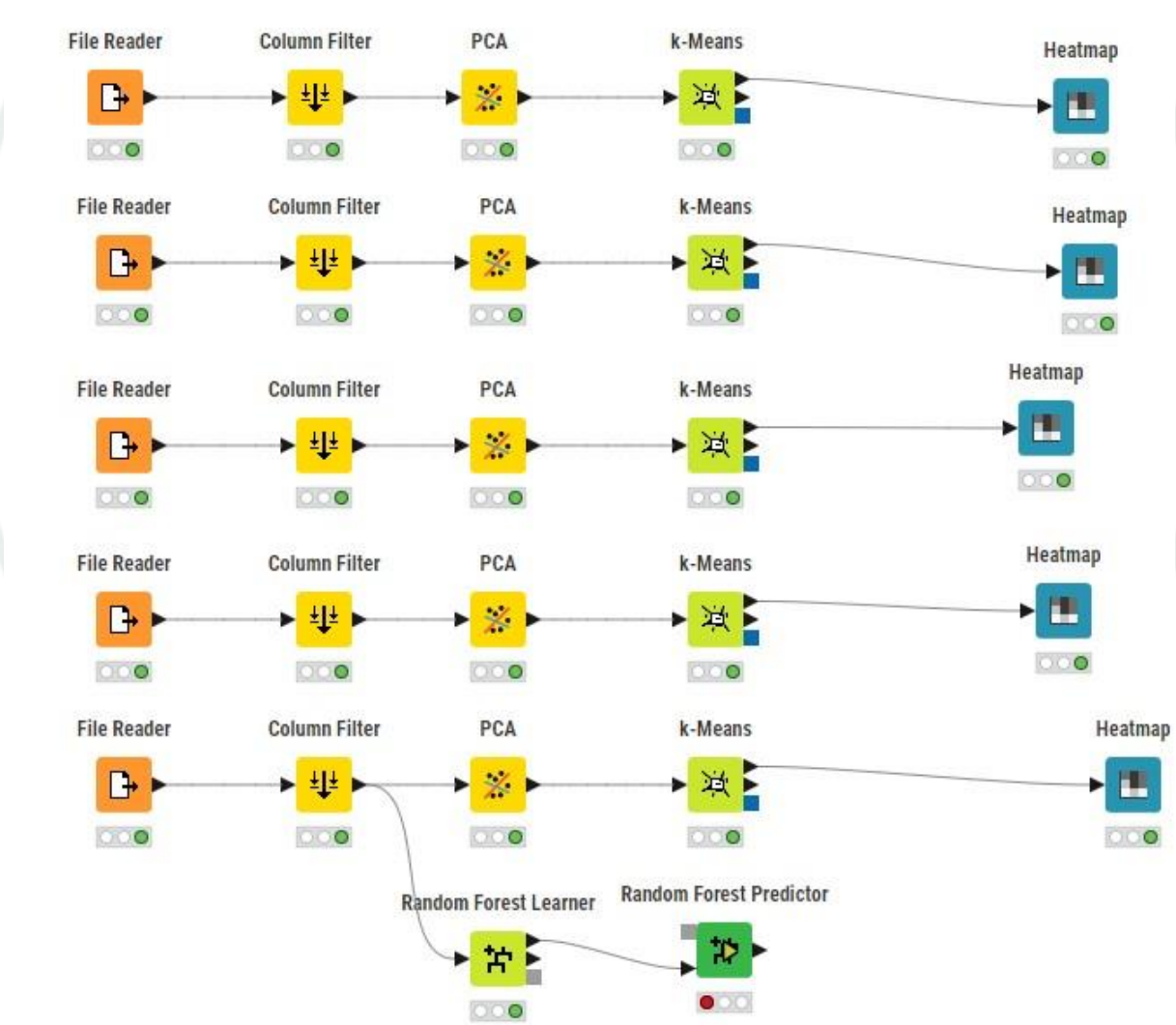
Exploración preliminar de datos para IA

Al construir un sistema que aproveche capacidades únicas de diferentes categorías de AI en distintas etapas, es posible combinar procesos automáticos que reduzcan los datos con

modelos que tomen decisiones. En la práctica, los primeros pasos se centraron en simplificar y filtrar la información disponible, para luego evaluar si resulta útil un modelo más complejo.
Con un flujo de trabajo creado en Orange, donde dos fuentes de datos diferentes son conectadas y comparadas. Este tipo de análisis permite verificar qué información es relevante antes de pasar a etapas más avanzadas.
En contraste, la Figura 2 presenta un flujo más sencillo en Knime, que aplica pasos de reducción y organización de datos hasta mostrar una visualización clara de los resultados. De esta forma, se obtiene una idea general del valor de los archivos procesados sin necesidad de interpretar cada variable en detalle.

Figura 2

Flujo de trabajo en Knime: simplificación e interpretación



Fuente: Elaboración propia.

Según el ámbito de la investigación, las redes neuronales convolucionales profundas (DCNN) suelen ser las más adecuadas para clasificación, ya que aprovechan la CNN junto a un comportamiento simplificado o reductor de dimensionalidad en los pool y en las capas densas, que constituyen las decisiones finales para alcanzar los objetivos. Sin embargo, las corrientes actuales sugieren tomar diferentes rumbos según la meta del estudio, como es el caso de las arquitecturas híbridas o el uso combinado de varias de ellas en búsqueda de nuevas conexiones intermedias.

4.2 Comprensión de los Datos

Comparación del dataset de entrenamiento y la muestra

Esfundamentalpresentarelementosclavesdelasbasesdedatosutilizadaspararespaldarsu selección y permitir una comprensión clara de sus diferencias. Entre estos aspectos se consideran el contexto de obtención, las características principales, los objetivos y, de manera descriptiva, el idioma predominante (en este caso, inglés), facilitando así su comparación.

La Tabla 4 resume las principales características de los datasets empleados. Ambos se centran en la expresión emocional a través de la voz, pero difieren en su origen y naturaleza. Esta diferencia implica etapas distintas de análisis: mientras que las señales de voz de actores en entornos controlados tienden a ser más homogéneas, las escenas cinematográficas aportan un contexto más natural y diverso. No obstante, en ambos casos las señales acústicas capturan adecuadamente la variabilidad emocional, coherente con los avances en estrategias de modelamiento para reconocimiento vocal, cuya efectividad supera el 90%.

En este sentido, CREMA-D fue diseñado con actores profesionales en un entorno controlado, mientras que la muestra de este estudio (EmoStim) proviene de escenas cinematográficas, lo que ofrece un marco más próximo a situaciones reales, aunque con menor uniformidad en las condiciones de grabación.

Tabla 4 *Descripción general de los datasets utilizados*

Característica	**CREMA-D**	**EmoStim**
Propósito principal	Expresión vocal de emociones actuadas	Inducción emocional mediante escenas audiovisuales
Uso en esta investigación	Análisis vocal de emoc expresadas	Análisis vocal en clips seleccionados con presencia de diálogo
Fuente de datos	Kaggle	Google Drive (acceso a CSV y audios seleccionados)
Formato de audio	WAV (original)	WAV (tras estandarización)
Duración promedio	2 segundos	unos segundos a varios minutos (selección de partes habladas)

Cantidad de clips	7,442	99 (de 139 originales)
Actores / Participantes	91 actores diversos	Múltiples actores en escenas de películas
Etiquetas emocionales	Felicidad, tristeza, enojo, miedo, disgusto, neutral	Emociones básicas según modelo discreto

Fuente: Elaboración propia.

La Tabla 5 profundiza en estas distinciones, mostrando que CREMA-D ofrece mayor consistencia y control experimental, ideal para análisis técnicos, mientras que EmoStim aporta una riqueza contextual valiosa pero con menor homogeneidad. También es importante notar que, mientras CREMA-D ya está balanceado por emoción y actor, EmoStim requirió filtrado manual para garantizar cierta equidad entre clases.

Estas diferencias justifican el uso combinado de ambos conjuntos: uno aporta datos limpios y controlados, el otro, expresiones más espontáneas y realistas, permitiendo evaluar los modelos propuestos en escenarios tanto controlados como naturales.

Tabla 5 *Diferencias clave entre los datasets CREMA-D y EmoStim*

Aspecto	**CREMA-D**	**EmoStim**
Enfoque conceptual	Expresión vocal intencionada (actuada)	Expresión vocal contextual (en escenas cinematográficas)
Balance de clases emocionales	Altamente balanceado por clase y actor	Se realizó un filtrado manual para obtener representatividad emocional y vocal
Validación emocional	Evaluación por crowdsourcing en modalidad auditiva	Validación emocional mediante CrowdFlower
Duración y consistencia	Fragmentos cortos y homogéneos (1–3 s)	Fragmentos variables, centrados en presencia de voz emo-cional
Ventaja principal	Precisión y control experimental	Variedad emocional en con-texto realista

Fuente: Elaboración propia.

En CREMA-D (Lok & Cooper, 2025), los audios están segmentados con una duración promedio de 1 a 3 segundos. Las emociones representadas están distribuidas de forma balanceada entre clases, lo que facilita su uso como base controlada.

En cambio, EmoStim (Mohammadi et al., 2023) presenta audios con duraciones que varían desde unos pocos segundos hasta varios minutos. Las emociones inducidas abarcan una gama más amplia. Sin embargo, muchos clips carecen de diálogos prominentes, lo que genera un

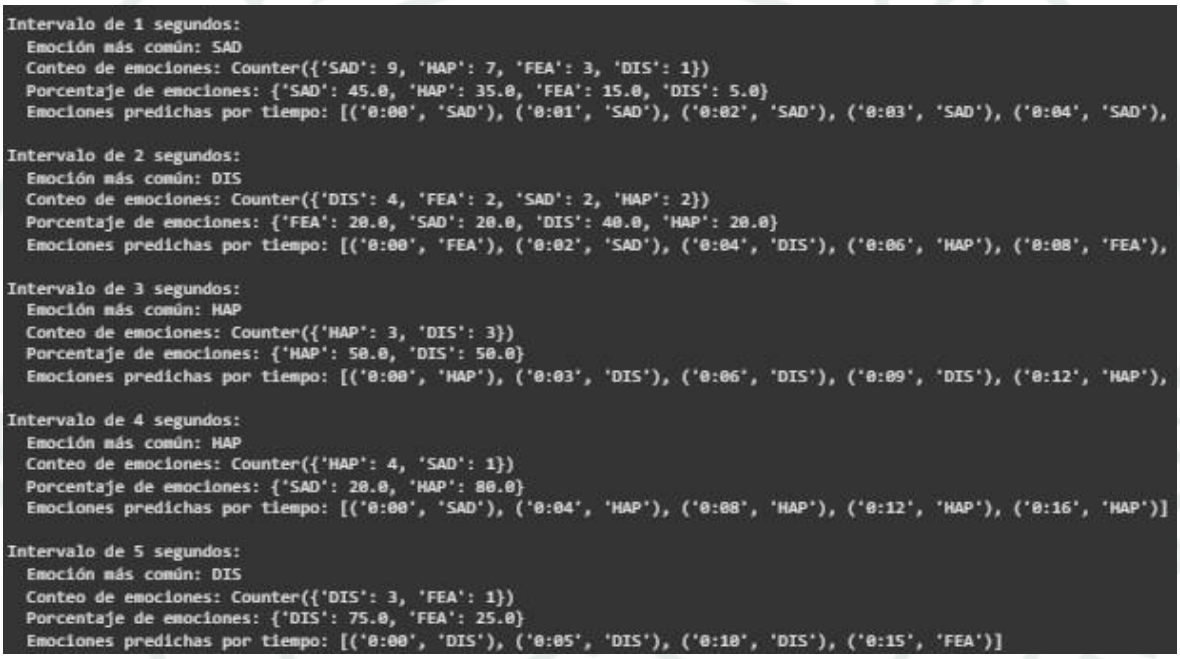
desbalance emocional tanto en la representación por clase como en la presencia de voz, dificultando su uso directo sin preprocesamiento adicional.

Elementos balanceados y desbalanceados

A continuación se presentan diversos análisis visuales sobre la distribución, anomalías y estructura de los audios seleccionados o personales, junto con la metadata y la forma de los datasets utilizados.

Figura 3

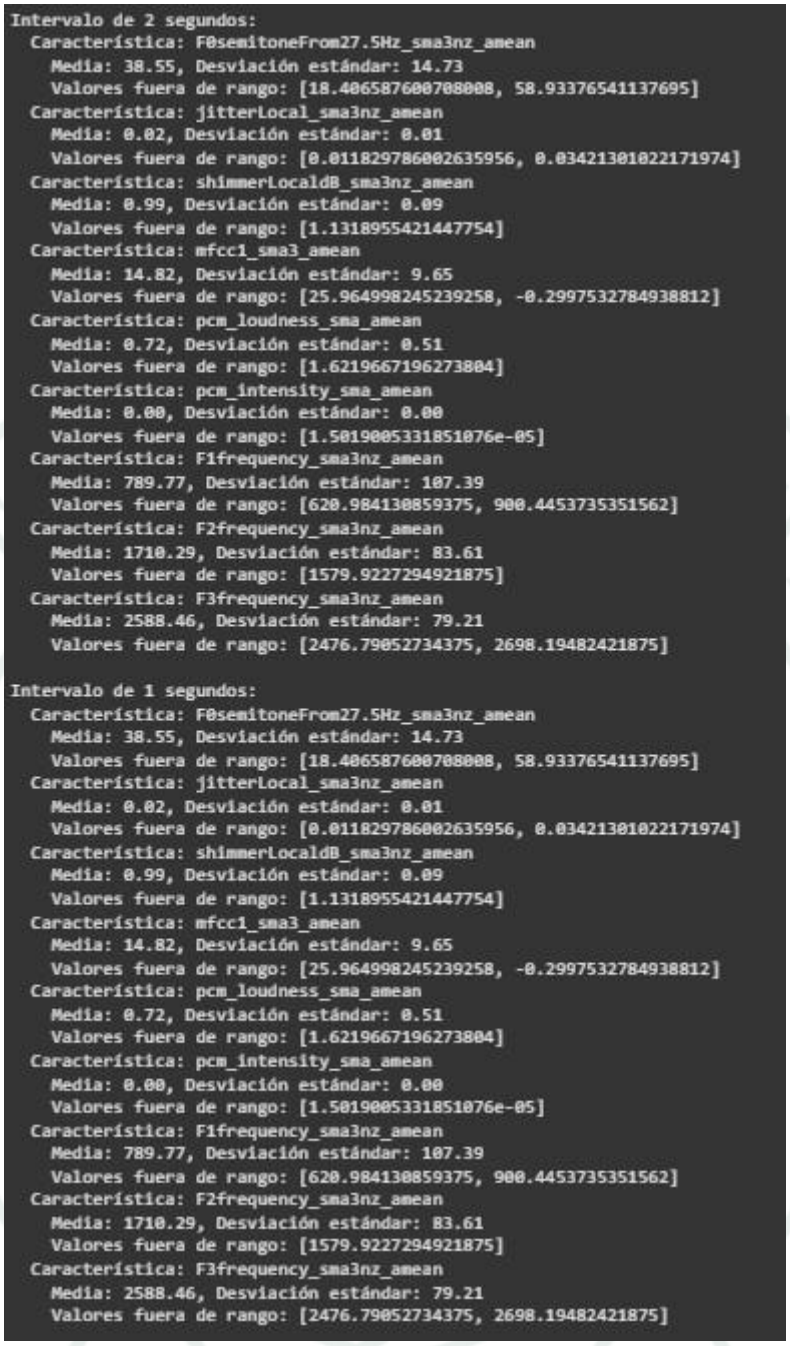
Audios clasificados en tramos por segundo



Fuente: Elaboración propia.

En la Figura 3 se muestra el comportamiento de las emociones en distintos tramos de tiempo. El análisis se presenta mediante el conteo de emociones detectadas, el porcentaje de audio asociado a cada una y el segmento temporal correspondiente, lo que permite identificar variaciones emocionales en intervalos de tiempo específicos.

Figura 4 *Audios: anomalías por desviación estándar*



Fuente: Elaboración propia.

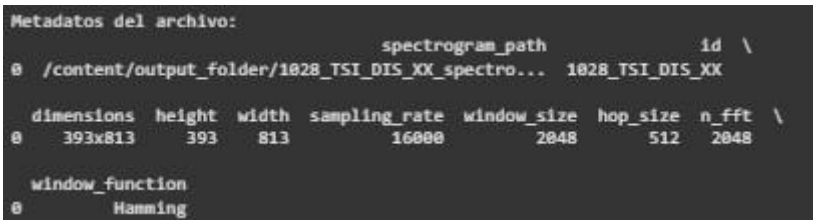
Con la Figura 4 se analizan diferentes tramos de tiempo a partir de las características vocales que presentan anomalías estadísticas según la desviación estándar. Este análisis permite evaluar si los valores de media y rango aportan información relevante al estudio o si se mantienen cercanos a cero, así como identificar segmentos con comportamientos atípicos dentro de cada audio.

Figura 5 *Audios: anomalías en emociones por desviación estándar*

Fuente: Elaboración propia.

Finalmente, la Figura 5 amplía este análisis al agrupar las anomalías por emoción. De esta manera, se identifica si el comportamiento de cada característica se mantiene dentro del rango esperado o si presenta una variabilidad más compleja, aportando una comprensión más detallada del patrón emocional del audio en el tiempo.

Figura 6 *Metadata de archivos de audio analizados*



Fuente: Elaboración propia.

Figura 7 *Distribución (shape) del dataset EmoStim*

```
X_emb = np.load("/content/embeddings/emo_stim/X_emb.npy")
y_pred = np.load("/content/embeddings/emo_stim/y_pred.npy")
val_df = pd.read_csv("/content/embeddings/emo_stim/val_df.csv")
val_df = val_df.reset_index(drop=True)
print(f"X_emb shape: {X_emb.shape}")   # Esperado: (n_samples, n_features)
print(f"y_pred shape: {y_pred.shape}") # Esperado: (n_samples,)
print(f"val_df shape: {val_df.shape}") # Esperado: (n_samples, n_columns)
```

```
X_emb shape: (609, 159744)
y_pred shape: (609,)
val_df shape: (609, 101)
```

Fuente: Elaboración propia.

Figura 8 *Distribución (shape) del dataset CREMA-D*

```
import numpy as np
import pandas as pd

X_emb = np.load("/content/embeddings/crema_d/X_emb.npy")
y_pred = np.load("/content/embeddings/crema_d/y_pred.npy")
val_df = pd.read_csv("/content/embeddings/crema_d/val_df.csv")
val_df = val_df.reset_index(drop=True)
print(f"X_emb shape: {X_emb.shape}")   # Esperado: (n_samples, n_features)
print(f"y_pred shape: {y_pred.shape}") # Esperado: (n_samples,)
print(f"val_df shape: {val_df.shape}") # Esperado: (n_samples, n_columns)
```

```
X_emb shape: (1870, 159744)
y_pred shape: (1870,)
val_df shape: (1870, 101)
```

Fuente: Elaboración propia.

Las Figura 6,7, 8 complementan la descripción de los datasets, mostrando su metadata y estructura interna.

En conjunto, estas figuras permiten comprender las diferencias entre datasets balanceados y desbalanceados, además de facilitar la detección de anomalías y el análisis de la estructura de los audios para un procesamiento posterior adecuado.

4.3 Preparación de Datos

Limpieza y Segmentación de Audio

En el caso de CREMA-D, se unificaron los clips en segmentos temporales consistentes, seleccionando una sola duración entre las tres disponibles. Se conservaron las seis clases emocionales originales.

Para EmoStim, se repitió el procedimiento, agregando una preselección manual debido a la carencia de diálogos o clips poco representativos, pasando de 99 películas a 42. Posteriormente, debido al número propuesto en la muestra y al desbalance emocional del conjunto, y tomando en cuenta el costo computacional de la selección de clips más eficiente, se optó por simplificar la división en dos grupos de 21 clips en orden alfabético, donde se aplica la desviación estándar en la etiqueta emocional.

Normalización y Estandarización de Señales

Los clips de audio fueron segmentados, normalizados o estandarizados utilizando parámetros acústicos multidimensionales, aplicadas en el preprocesamiento del audio se detallan en la Tabla 6.

Tabla 6 *Parámetros espectrales y de preprocesamiento de audio*

Categoría	**Detalles / Parámetros**
Parámetros espectrales	*n-fft* = 1024, *hop-length* = 256, *win-length* = 512, *mels n-* = 128.
Características de voz	`Resize` a 128×313, `ToTensor`().
Normalización	Z-score: $z = \frac{x-\mu}{\sigma}$.
Tramos de duración	Segmentos de audio de 1, 2 y 3 segundos.

Fuente: Elaboración propia.

División de Elementos

Los datos se dividieron solo en dos conjuntos, ya que la mayoría funcionan haciendo su propia subdivisión, según su naturaleza de caja negra: entrenamiento y validación. Se seleccionará el grupo con mayor equilibrio entre clases emocionales, y en la medida de lo posible, el que tenga menor cantidad representable, considerando gráficas y la desviación estándar.

Almacenamiento de datos y metadatos

Unavezgeneradasycomprobadaslasimágenesdeespectrogramatransformadasentramos temporales, modelos entrenados, complementos de modelos y archivos de control, se almacenaron estos últimos con los conjuntos de metadatos, tales como: ruta del espectrograma, ID del segmento, clase emocional, grupos, características acústicas o predicciones generadas.

4.4 Representación de Datos

Etiquetas y Clases Emocionales

Los conjuntos de datos utilizados para la detección emocional se resumen en la Tabla 7, mostrando sus emociones y características relevantes. Se conservaron las seis emociones

originales de CREMA-D como estándar, también llamadas clases en modelos complejos interpretativos, correspondientes al siguiente orden:

Tabla 7 *Comparación entre conjuntos de datos emocionales*

Dataset	Emociones (ID de clase)	Características
CREMA-D	Anger (0), Disgust (1), Fear (2), Happy (3), Neutral (4), Sad (5).	• Seis emociones discretas balancea-das. • Actuadas, etiquetadas, con buena ca-lidad de grabación.
EmoStim	Interest, Fear, Anxious, Moved, Anger, Ashamed, Warmhearted, Joy, Sad, Satisfied, Surprise, Love, Guilt, Disgust, Disdainful, Calm.	• Múltiples emociones espontáneas. • Se agruparon según representatividady presencia vocal.

Fuente: Elaboración propia.

Extracción de Características

Se utilizaron dos enfoques complementarios. El concepto de **emociones** se emplea para referirse a los elementos básicos para el análisis emocional, incluyendo los espectrogramas redimensionados y el conjunto de características. Los componentes de este enfoque se detallan en la Tabla 8.

Tabla 8 *Componentes del enfoque basado en emociones*

Categoría	Descripción / Detalles
Elemento	• Audio: archivo temporal. • Imagen: representación visual del espectrograma Mellogarítmico. • Transformer: representación de la imagen transformada. • Características de voz: conjunto de características Opens-mile Egmaps.

Predicción	• Accuracy • F1-score • Precision • Recall • Matriz de confusión • Desviación estándar • Validación cruzada
Modelado	• Stem convolucional: para reducción dimensional inicial. • Transformer: para capturar relaciones globales. • Capa densa: para clasificación final.
Hiperparámetros	• Funciones de activación: ReLU, Softmax. • Ajustes empíricos: número de capas, tamaño del patch, nú-mero de cabezas de atención, etc.
Funciones	• Función de pérdida: CrossEntropyLoss. • Optimizador: Adam.

Fuente: Elaboración propia.

A partir de las emociones, al interpretar el embedding, se empieza a usar el término **clases**, elementos complejos y representativos con alta fidelidad. Los componentes asociados a este segundo enfoque se resumen en la Tabla 9.

Tabla 9 *Componentes del enfoque basado en clases*

Categoría	**Descripción / Detalles**
Elemento	• Características explicativas: conjunto de característicasrepresentativas de uno o varios embeddings. • Embedding: representación numérica de características delespectrograma.
Predicción	• Fidelity: medida de parentesco entre modelos. • MSE: error cuadrático medio. • R^2: varianza explicada.

Modelado	• `val-df`: metadatos (ruta, clase, características de entrada). • `y-pred`: predicciones de clases. • `x-emb`: embeddings antes de la capa final. • `vars-used`: características de voz usadas. • `max-depth`: profundidad del árbol.

Fuente: Elaboración propia.

Se dispone de archivos `features`, que contienen metadatos básicos que permiten reducir el reprocesamiento en etapas tempranas, y archivos `combined_features`, que los combinan con representaciones, aumentando la integridad, el control, la gestión y la trazabilidad de los datos. Algunos campos pueden permanecer vacíos o variar según el tipo de segmentación aplicada o la duración del fragmento analizado, los componentes se detallan en la Tabla 10.

Tabla 10 *Metadatos del conjunto de espectrogramas utilizados*

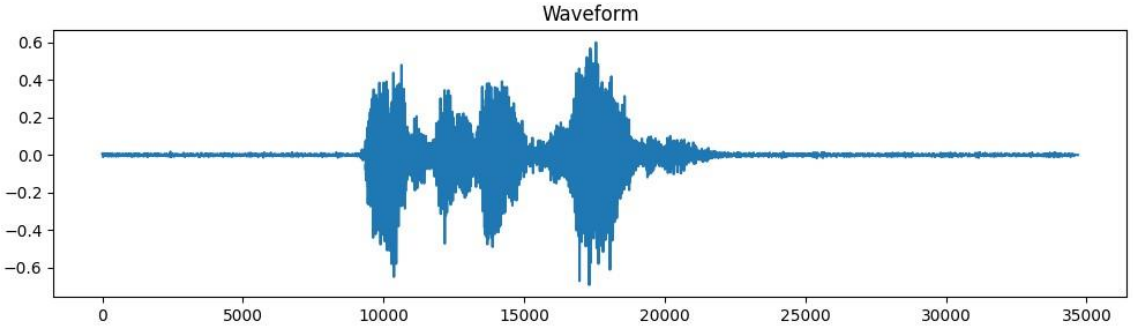
Categoría	**Descripción / Ejemplo**
`spectrogram_path`	Ruta completa del archivo de imagen correspondiente al espectrograma generado. Ejemplo: `/content/cremad/spectrograms/1028_TSI_DIS_XX_seg0.png`
`id`	Identificador único del archivo o segmento de audio analizado. Ejemplo: `1028_TSI_DIS_XX_seg0`.
`dimensions`	Tamaño del espectrograma expresado como `altura × ancho` en píxeles. Ejemplo: `128x153`.
`height / width`	Altura y anchura del espectrograma, respectivamente, en píxeles. Ejemplo: `128 / 153`.
`sampling_rate`	Frecuencia de muestreo del audio original en Hz. Ejemplo: `16000 Hz`.
`window_size`	Tamaño de la ventana de análisis en muestras. Ejemplo: `512`.
`hop_size`	Desplazamiento entre ventanas consecutivas. Ejemplo: `256`.
`n_fft`	Número de puntos de la Transformada Rápida de Fourier (FFT). Ejemplo: `1024`.
`segment_index`	Índice del segmento analizado dentro de un archivo de audio. Ejemplo: `0`.
`segment_duration`	Duración del segmento en segundos. Por defecto sería `1`, pero en este tipo de clips completos no se almacena.
`window_function`	Tipo de ventana aplicada en el análisis. Ejemplo: `Hamming`.

Fuente: Elaboración propia.

Las transformaciones del audio se pueden representar de distintas formas, por ejemplo, la forma de onda (Waveform) se muestra en la Figura 9, mientras que el espectrograma se ilustra en la

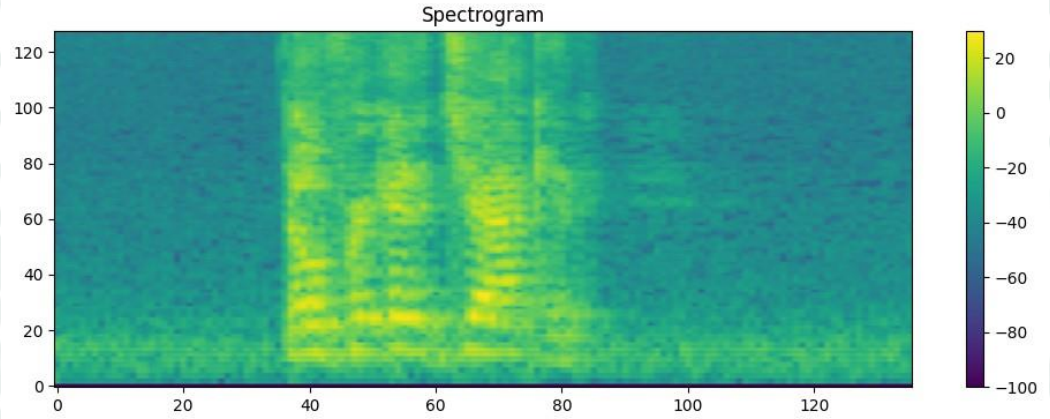
Figura 10. Además, se pueden observar los efectos de la redimensión en la Figura 11 y los parches generados para el Vision Transformer (ViT) en la Figura 12.

Figura 9 *Representacion de audio en forma de onda*



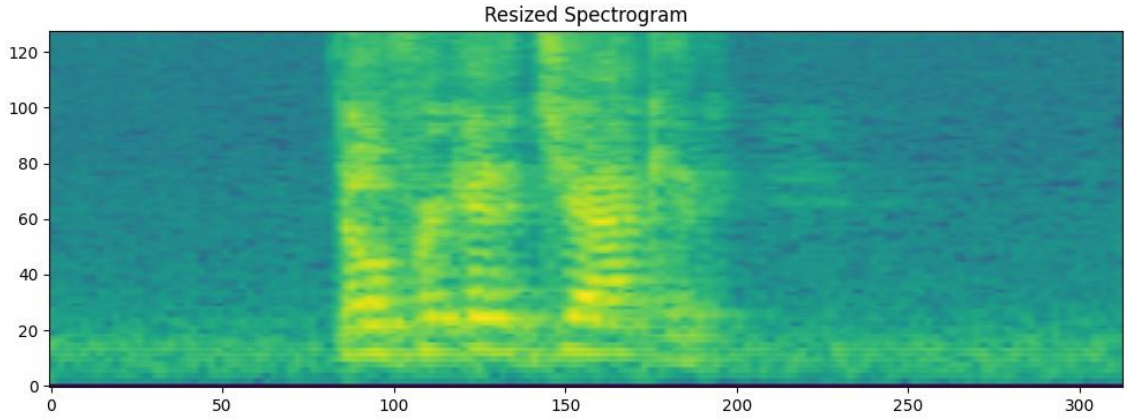
Fuente: Elaboración propia.

Figura 10 *Representacion de audio en espectrograma*



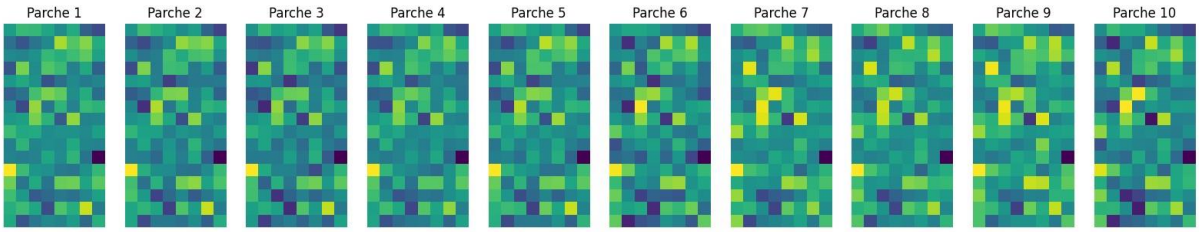
Fuente: Elaboración propia.

Figura 11 *Representacion de audio en redimension del espectrograma*



Fuente: Elaboración propia.

Figura 12 *Representacion de audio en parches Vision Transformer*



Fuente: Elaboración propia.

Análisis, selección y resumen de características vocales y embeddings

En esta sección se presentan visualizaciones clave y ejemplos que ilustran el análisis y la relación entre características vocales y embeddings. Agrupado por temas para facilitar su comprensión y referencia cruzada.

La Figura 13 muestra cómo varían los coeficientes de las 5 características vocales mas relevantes en función de los embeddings.

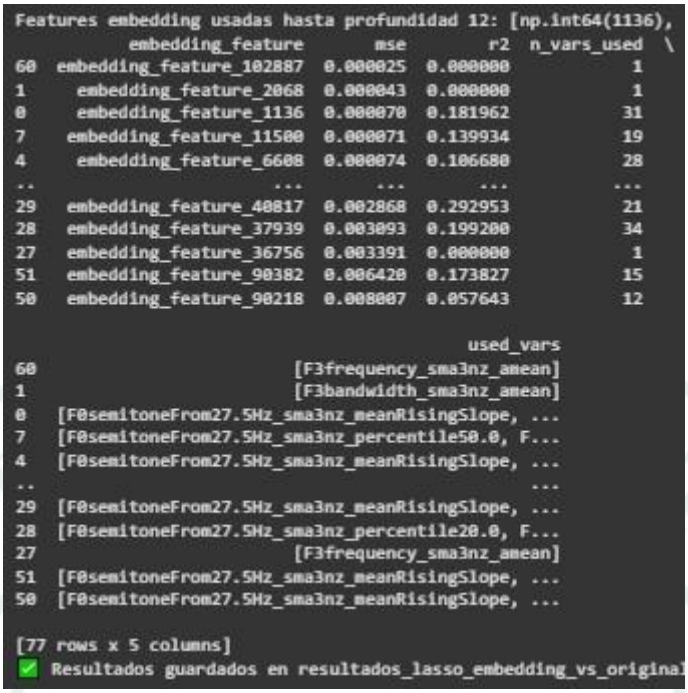
Figura 13 *Coeficiente por característica vocal según embeddings*

```
Top 5 vocal features que predicen feature_10838:
  equivalentSoundLevel_dBp: +0.0214
  spectralFluxV_sma3nz_amean: -0.0108
  F0semitoneFrom27.5Hz_sma3nz_percentile80.0: +0.0099
  mfcc1V_sma3nz_amean: -0.0090
  slopeUV0-500_sma3nz_amean: -0.0076

Top 5 vocal features que predicen feature_14142:
  mfcc2_sma3_amean: +0.0020
  slopeUV500-1500_sma3nz_amean: -0.0010
  mfcc4V_sma3nz_amean: -0.0003
  loudness_sma3_stddevNorm: -0.0003
  mfcc4_sma3_stddevNorm: +0.0001

Top 5 vocal features que predicen feature_21387:
  mfcc1V_sma3nz_amean: +0.0062
  mfcc3_sma3_amean: +0.0032
  slopeUV500-1500_sma3nz_amean: +0.0029
  shimmerLocaldB_sma3nz_amean: -0.0019
  hammarbergIndexV_sma3nz_stddevNorm: -0.0016

Top 5 vocal features que predicen feature_21612:
  slopeUV500-1500_sma3nz_amean: +0.0046
  loudness_sma3_stddevNorm: +0.0031
  slopeUV0-500_sma3nz_amean: +0.0028
  alphaRatioV_sma3nz_amean: -0.0019
  HNRdBACF_sma3nz_amean: -0.0015
```

Fuente: Elaboración propia.

Complementando la anterior, la Figura 14 muestra su comportamiento aislado de los primeros y últimos embeddings en función a valores lasso y características únicas.

Figura 14 *Relación entre características vocales y R^2 con máxima profundidad*

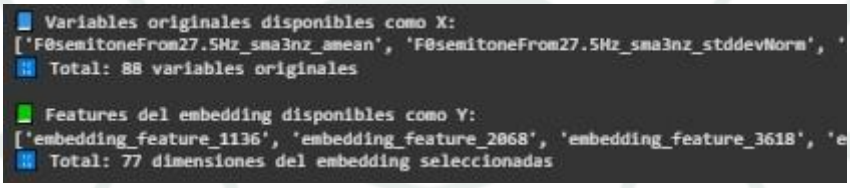


Fuente: Elaboración propia.

Un proceso importante es la actualización automática de dimensiones y su control, junto a su fidelidad, métricas y metadatos, como se ve en Figura 15 haciendo un almacenamiento dummy de características y embeddings,

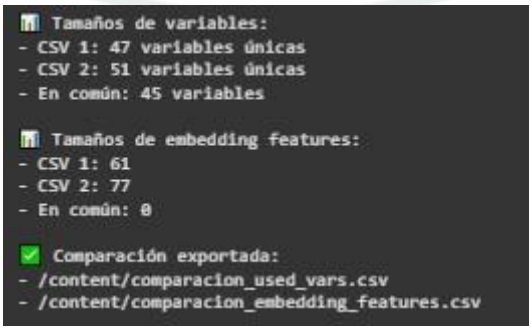
Con la Figura 16 se hace una comprobación y análisis de similitudes, mostrando la ubicación de estos archivos.

Figura 15 *Dimensiones de relación posible y R^2*



Fuente: Elaboración propia.

Figura 16 *Relación entre características vocales y embeddings*



Fuente: Elaboración propia.

La Figura 14 muestran un análisis estadístico más profundo, utilizando R^2 para evaluar la calidad de la relación entre variables.

Figura 17 *Set de características vocales analizadas*

```
    # Guardar en archivo CSV
    pd.Series(columnas).to_csv(f"{nombre_modelo}_feature_names.csv", index=False, header=False
    print(f"✅ Modelo '{nombre_modelo}' tiene {len(columnas)} columnas. Guardado en: {nombre_

/usr/local/lib/python3.11/dist-packages/opensmile/core/smile.py:252: UserWarning: Feature set
  warnings.warn(
✅ Modelo 'eGeMAPSv02' tiene 88 columnas. Guardado en: eGeMAPSv02_feature_names.csv
✅ Modelo 'emobase' tiene 988 columnas. Guardado en: emobase_feature_names.csv
✅ Modelo 'eGeMAPS' tiene 88 columnas. Guardado en: eGeMAPS_feature_names.csv
```

Fuente: Elaboración propia.

Figura 18 *Características vocales menos utilizadas*

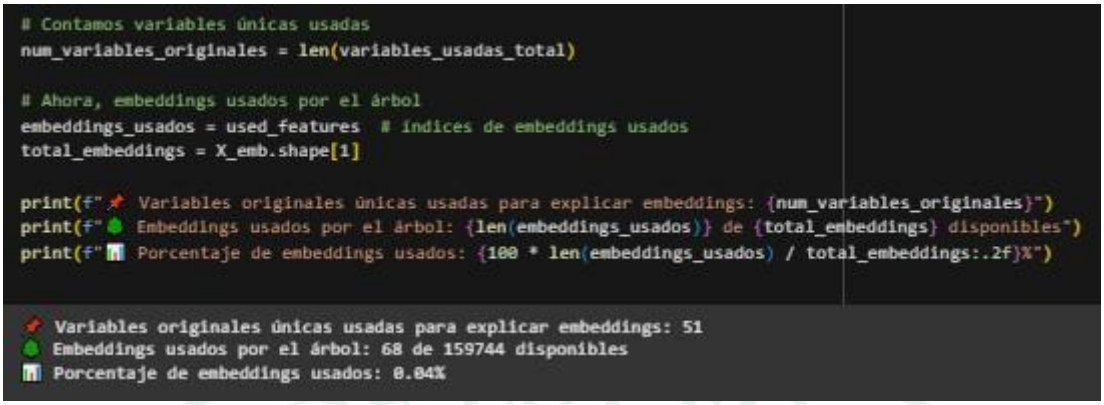
```
# Mostrar primeras filas
print(df_expandido.head(20))

# (Opcional) Guardar en CSV para explorar más fácilmente
df_expandido.to_csv("/content/embedding_vs_variables.csv", index=False)
```

	Embedding	Variable acústica
0	embedding_feature_4283	F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope
0	embedding_feature_4283	mfcc1_sma3_amean
0	embedding_feature_4283	mfcc2_sma3_amean
0	embedding_feature_4283	logRelF0-H1-H2_sma3nz_stddevNorm
0	embedding_feature_4283	F1frequency_sma3nz_amean
0	embedding_feature_4283	F2bandwidth_sma3nz_amean
0	embedding_feature_4283	F2amplitudeLogRelF0_sma3nz_amean
0	embedding_feature_4283	F3frequency_sma3nz_amean
0	embedding_feature_4283	F3bandwidth_sma3nz_amean
0	embedding_feature_4283	slopeV500-1500_sma3nz_stddevNorm
0	embedding_feature_4283	mfcc2V_sma3nz_stddevNorm
0	embedding_feature_4283	mfcc3V_sma3nz_stddevNorm
1	embedding_feature_8601	F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope
1	embedding_feature_8601	F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope
1	embedding_feature_8601	mfcc1_sma3_amean
1	embedding_feature_8601	mfcc4_sma3_stddevNorm
1	embedding_feature_8601	F1frequency_sma3nz_amean
1	embedding_feature_8601	F2frequency_sma3nz_amean
1	embedding_feature_8601	F2bandwidth_sma3nz_amean
1	embedding_feature_8601	F2amplitudeLogRelF0_sma3nz_amean

Fuente: Elaboración propia.

Las Figuras 17 y 18 permiten visualizar en ejemplos preliminares el comportamiento de diferentes sets de caracteristicas con los embeddings.

Figura 19 *Relación entre embeddings utilizados y totales*



Fuente: Elaboración propia.

Finalmente, la Figura 19 presenta un resumen del uso de embeddings en el modelo, comparando los utilizados frente al total disponible.

La llamada de entrenamiento o prueba del sistema utilizará las siguientes métricas y parámetros, como se detalla en el Anexo F.4, al momento de ejecutarse la funcion de control de generacion de baches recibiremos la siguiente información, en caso de estar en la etapa de generación, se mostrarán el avisos correspondientes, junto a un mensaje de validación sobre los elementos encontrados. Además, se presenta la salida del archivo comprimido, que contiene elementos clave como modelos de entrenamiento, DT, características vocales, dummies de audio e imágenes, metadatos y archivos de control. Cada archivo y su proceso de generación son modulares, lo que permite procesar y validar el sistema en cualquier momento, incluyendo ajustes en fidelidad y precisión.

En resumen, las Figuras 13 a 19 constituyen el análisis exploratorio de coeficientes, características y embeddings, respaldado por los anexos que detallan aspectos clave del enfoque, validación y especificaciones utilizadas en los múltiples modelos.

4.5 Arquitectura del Modelo Propuesto

La arquitectura propuesta recibe como entrada un conjunto de audios, segmentados en tramos de tiempo para su análisis, de cada segmento se extraen características vocales mediante OpenSMILE y representaciones espectrales, considerando aspectos de transformación como la dimensionalidad y escala de grises, que permiten capturar variaciones relevantes en los datos.

Las CNN se emplean como extractores principales de características debido a su alta precisión en la clasificación emocional. Se analizan capas intermedias y finales, ya que aportan representaciones informativas que permiten relacionar diferentes dimensiones de las representaciones y evaluar el efecto de hiperparámetros. Complementariamente, se representó el comportamiento de Transformers modelando dependencias temporales y ajustando los datos mediante parches de tiempo, mejorando la precisión por encima del 90%, como indica la literatura.

Para reducir la complejidad del análisis y guiar la toma de decisiones, se utilizan modelos auxiliares de forma iterativa, estos apoyan la selección de algoritmos, tramos de tiempo, conjuntos de características y métricas, aplicando técnicas de reducción de datos como PCA, agrupamiento con DBSCAN, balanceo de clases y estrategias de control de sobreajuste (división entrenamiento/validación/prueba, generación de datos simulados, reorganización de archivos).
Las representaciones profundas se transforman en modelos interpretables mediante DT y Lasso, los árboles permiten identificar patrones emocionales claros en los segmentos de audio, controlando la complejidad mediante el número de nodos y pesos. Lasso facilita la simplificación de modelos y la reducción de datos ruidosos, permitiendo que las representaciones extraídas de CNN y Transformers sean comprensibles para análisis posteriores.
La arquitectura es modular y escalable, aplicable a diferentes tipos de datos, métricas y modelos, la modularidad abarca desde la organización de archivos, sets de características y manejo de anomalías, hasta la carga, almacenamiento y algunas adaptaciones con funciones para modelos entrenados o preentrenados.
Gracias a esto, el sistema genera clasificaciones emocionales precisas e interpretables, asociadas a segmentos específicos del audio, y permite analizar patrones complejos de comportamiento emocional en distintos contextos.

Configuración del Entorno de Entrenamiento

Se utilizarán los cuadernos de Google Colab para ejecutar código en servidores en la nube mediante máquinas virtuales, servicio gratuito o de pago con acceso a recursos computacionales, GPU y TPU con ciertas restricciones. Este entorno es muy utilizado para educación e investigación en modelos de IA, principalmente usando Jupyter y lenguajes como Python, aprovechando recursos compartidos como el almacenamiento en Google Cloud. Según esto, se genera de forma iterativa, por batches, el almacenamiento de datos, la generación de imágenes y la extracción de datos.
Mientras tanto, los diferentes modelos entrenados se pueden almacenar en archivos reinterpretables, de control y reintegración mediante código.

Componentes Técnicos del Sistema

A continuación, se resumen las principales técnicas empleadas en el desarrollo del sistema en la Tabla 11, junto con los procedimientos aplicados y las herramientas o bibliotecas utilizadas para su implementación. Esta descripción permite entender, de forma accesible, cómo se estructuró el procesamiento de datos y la modelación del problema.

Tabla 11 *Técnica, Procedimiento e Instrumento*

Técnica	**Procedimiento**	**Instrumento**

Convolutional Stem	Extracción de patrones visuales en espectrogramas	CNN simple con filtros, ReLU y max-pooling en PyTorch
Vision Transformer	Conversión a espectrograma Mel y etiquetado	Arquitectura Transformer visual de entrada con PyTorch
Decision Tree Classifier	Exploración de reglas embedding	Árbol entrenado con scikit-learn; análisis gráfico por profundidad y hojas
LassoCV	Detección de patrones entre variables cruzadas	Modelo de regresión Lasso, con análisis de coeficientes para selección de variables

Fuente: Elaboración propia.

En este proyecto se usaron librerías clave para procesar datos, construir y evaluar modelos, listadas en la Tabla 12. PyTorch y Torchvision fueron fundamentales para crear y entrenar redes neuronales con imágenes. Pandas y NumPy facilitaron el manejo de datos. Scikit-learn permitió usar DT y regresión Lasso para analizar y explicar los modelos. Librosa y OpenSMILE ayudaron a extraer características del audio. Para visualizar resultados, se emplearon Matplotlib y Seaborn. Además, Joblib se usó para guardar modelos, Kagglehub para gestionar datos y tqdm para mostrar el progreso de los procesos.

Tabla 12 *Librerías y frameworks clave utilizados*

Librería / Framework	**Justificación / Uso en el proyecto**
PyTorch (torch, torch.nn)	Base del desarrollo y entrenamiento de modelos profundos (CNN, Transformer) y extracción de embeddings.
Torchvision.transforms	Preprocesamiento y transformación avanzada de imágenes (espectrogramas).
PIL (Pillow)	Manipulación y carga eficiente de imágenes para el dataset.
Scikit-learn	Modelos interpretativos (DT, LassoCV), evaluación y validación de métricas.
Librosa	Extracción y análisis avanzado de características acústicas y espectrogramas.
OpenSMILE	Extracción de rasgos emocionales específicos para análisis de voz.
Joblib	Serialización y persistencia de modelos para reproducibilidad y despliegue.
Kagglehub	Gestión de datasets y experimentos, acceso a recursos colaborativos.

tqdm	Monitorización en tiempo real de procesos largos, demostrando complejidad y cuidado en la ejecución.
re (expresiones regulares)	Extracción y análisis de clases desde la representación textual del árbol de decisión.

Fuente: Elaboración propia.

Organización de archivos y modelos del pipeline

Los archivos y modelos del pipeline se organizan en carpetas con propósitos específicos, garantizando un flujo ordenado de procesamiento y análisis. La Tabla 13 resume esta estructura.

Tabla 13 *Organización de carpetas y archivos del pipeline*

Carpeta	Contenido / Propósito
`features`	Conjunto de metadatos, emociones e hiperparámetros utilizados en el modelo.
`audio`	Archivosdeaudiosinprocesar,entradaoriginaldelpipeline.
`dummy`	Elementos del conjunto de características de voz o representaciones en formato de imagen.
`checkpoint`	Modelo entrenado basado en CNN, utilizado como punto de control del entrenamiento.
`embeddings`	Representación intermedia generada por el modelo CNN+Transformer.
`detalle-export-text`	Resultados del proceso de búsqueda de hiperparámetros (GridSearch) o algoritmos de fidelidad.
`surrogate`	Estructura del modelo sustituto (Decision Tree), que incluye los valores de `features` analizados.
`results`	Resultados finales obtenidos al aplicar LassoCV sobre los `embeddings`.

Fuente: Elaboración propia.

Esta organización permite rastrear en cada etapa, desde la entrada de audio hasta la obtención de resultados finales, asegurando claridad, reproducibilidad y facilidad de interpretación de los modelos y sus representaciones.

4.6 Implementación y Validación de Modelos

Modelo previo de Árbol de decisión para selección del conjunto de datos y extracción de características

Se utilizó un mapa de calor para evaluar cómo afecta la duración de los clips de audio y el grupo de datos seleccionado en la precisión del modelo.

El eje horizontal indica la duración de los clips en segundos (1, 2 y 3). El eje vertical muestra los distintos grupos de datos utilizados, identificados por nombre, tipo de conjunto (completo o parcial) y número aproximado de clips.

Los grupos ’emotionclips’ contienen grabaciones breves(entre 1 y 3 segundos), mientras que los ’movie clips’ incluyen fragmentos más largos, que pueden extenderse por varios minutos.

Por este motivo, se dividieron en segmentos de duración similar a los demás para poder compararlos adecuadamente.

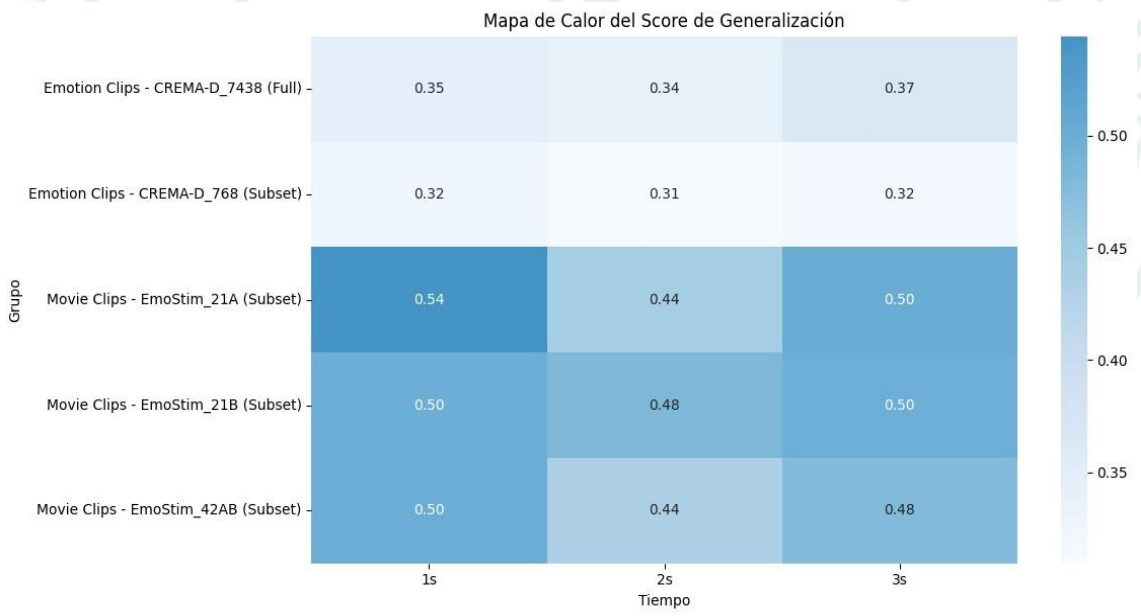
Cada celda del mapa muestra el porcentaje de precisión alcanzado por el modelo en cada combinación, y el color representa visualmente el nivel de rendimiento obtenido, en la Figura 20.

Bajo la premisa de escoger un modelo representativo y balanceado, se implementó un módulo de árbol de decisión para los diferentes datasets y muestras. Se evaluó cómo la duración de estos clips, el conjunto de datos seleccionado y la segmentación temporal afectan la precisión del modelo mediante un mapa de calor. El eje horizontal representa la duración de los clips (1, 2 y 3 segundos), y el eje vertical muestra los distintos grupos de datos utilizados, indicando su nombre, tipo (completo o parcial) y cantidad aproximada de clips.

Los audios provienen de dos fuentes principales: EmoStim (clips breves de pocos segundos, con expresiones emocionales) y CREMA-D (clips de películas, que inducen una emoción en el espectador). Para garantizar una comparación justa, algunos grupos tendrán métricas finales similares como duración total.

Cada celda del mapa de calor indica la precisión obtenida por el modelo para cada combinación de grupo y duración, con una escala de colores que representa visualmente el rendimiento alcanzado.

Figura 20 *Matriz de confusión por grupo*



Fuente: Elaboración propia.

Modelo previo de profundidad óptima del árbol de decisión de precisión y validación

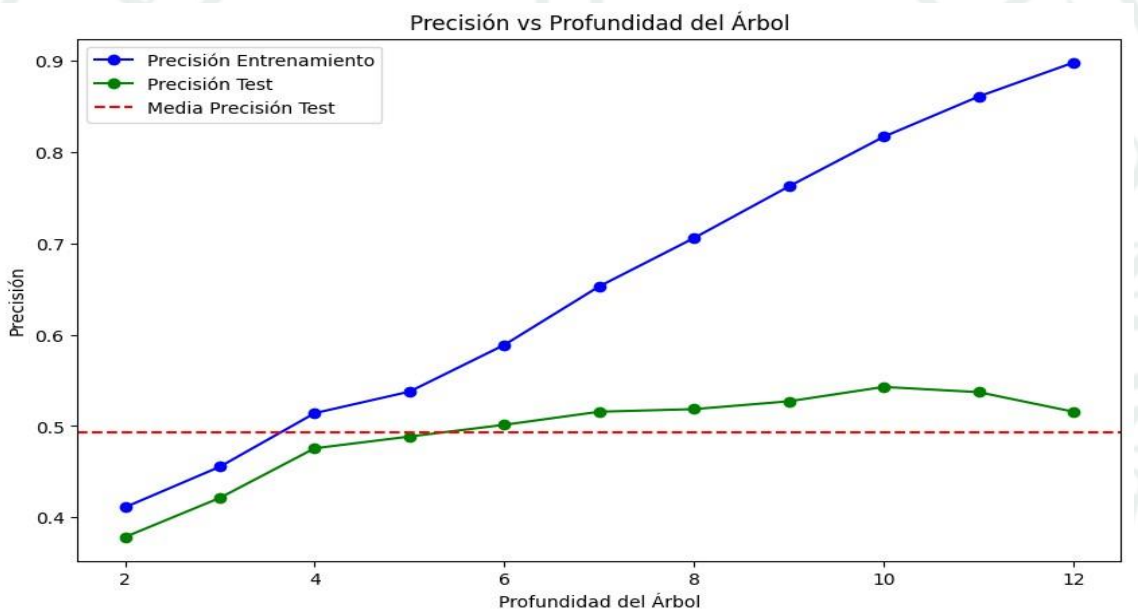
El rendimiento de los modelos de DT en la clasificación de patrones emocionales en la voz considera diferentes profundidades del árbol. De los grupos de experimento definidos en tramos de tiempo, se observa la precisión promedio en los conjuntos de entrenamiento y prueba. En la Figura 21 se observa que:

La precisión en el conjunto de prueba muestra una tendencia a estabilizarse o incluso decrecer ligeramente a partir de cierta profundidad.

Se trazó una línea punteada roja que representa la media de la precisión en el conjunto de prueba, utilizada como referencia para identificar las profundidades óptimas.

Se identificaron rangos de profundidad donde los modelos mantienen una alta precisión de prueba (superior o cercana a la media), con estructuras menos complejas (es decir, menor cantidad de nodos y hojas), típicamente en el intervalo de 5 a 10 niveles de profundidad.

Figura 21 *Punto de equilibrio del modelo de árbol*



Fuente: Elaboración propia.

Modelado de reconocimiento de emociones en voz

El modelo fue entrenado y validado utilizando clips de voz. Para evaluar su desempeño, se aplicaron métricas estándar para clasificación multiclase: precisión general (accuracy), precisión por clase, recall, F1-score y soporte (support).

En el conjunto de validación, el modelo 20250501-210144 alcanzó una validacion del 99.19%. Los resultados detallados por clase emocional se presentan en la Tabla 14, mientras que la evolución del rendimiento durante el entrenamiento se resume en la Tabla 15. Asimismo, se aprecia que la fase de entrenamiento fue eficiente, con una rápida reducción de la pérdida desde un valor inicial de 3.7 hasta 0.21 en las primeras dos épocas, para luego estabilizarse en valores mínimos, 0.0026 en la última época.

La matriz de confusión del modelo Figura 22 confirma lo observado en las métricas previas. La mayoría de los valores se concentran en la diagonal principal, lo que indica que la mayor parte de los casos fueron clasificados correctamente. Esto coincide con el alto desempeño alcanzado

Tabla 14 *Reporte de clasificación por emoción*

Emotion	**Precision**	**Recall**	**F1-score**	**Support**
Anger	0.996	0.988	0.992	254
Disgust	1.000	1.000	1.000	254
Fear	1.000	1.000	1.000	254
Happy	1.000	1.000	1.000	254
Neutral	0.991	1.000	0.995	218
Sad	0.996	0.996	0.996	254
Accuracy			0.997	1488
Macro avg	0.997	0.997	0.997	1488
Weighted avg	0.997	0.997	0.997	1488

Fuente: Elaboración propia.

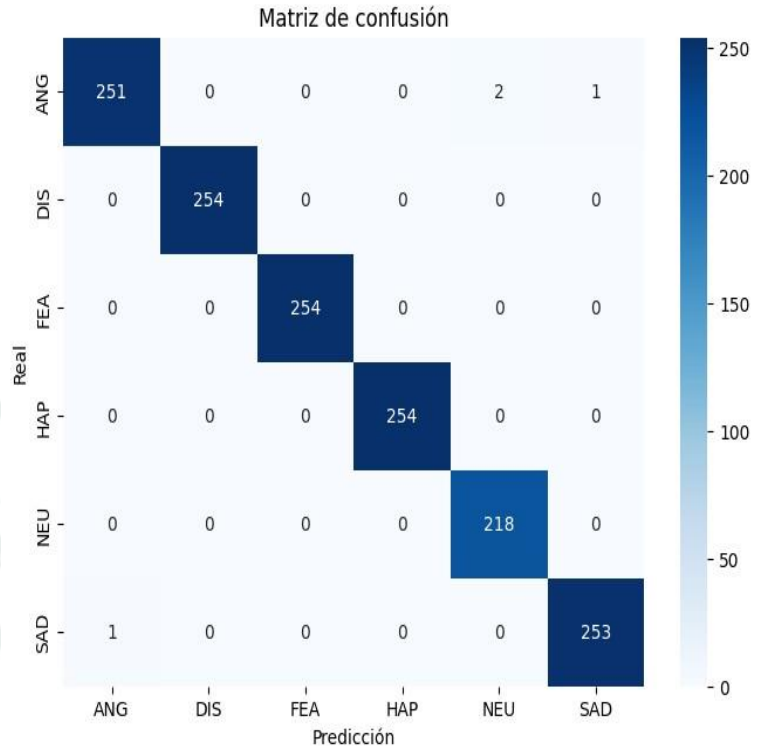
Tabla 15 *Evolución del desempeño por época*

Epoch	**Training Loss**
1/10	3.7067
2/10	0.2114
3/10	0.0752
4/10	0.0355
5/10	0.0226
6/10	0.0132
7/10	0.0087
8/10	0.0046
9/10	0.0034
10/10	0.0026

Fuente: Elaboración propia.

por el modelo, con confusiones mínimas y esperadas.

Figura 22 *Matriz de confusión de la IA*

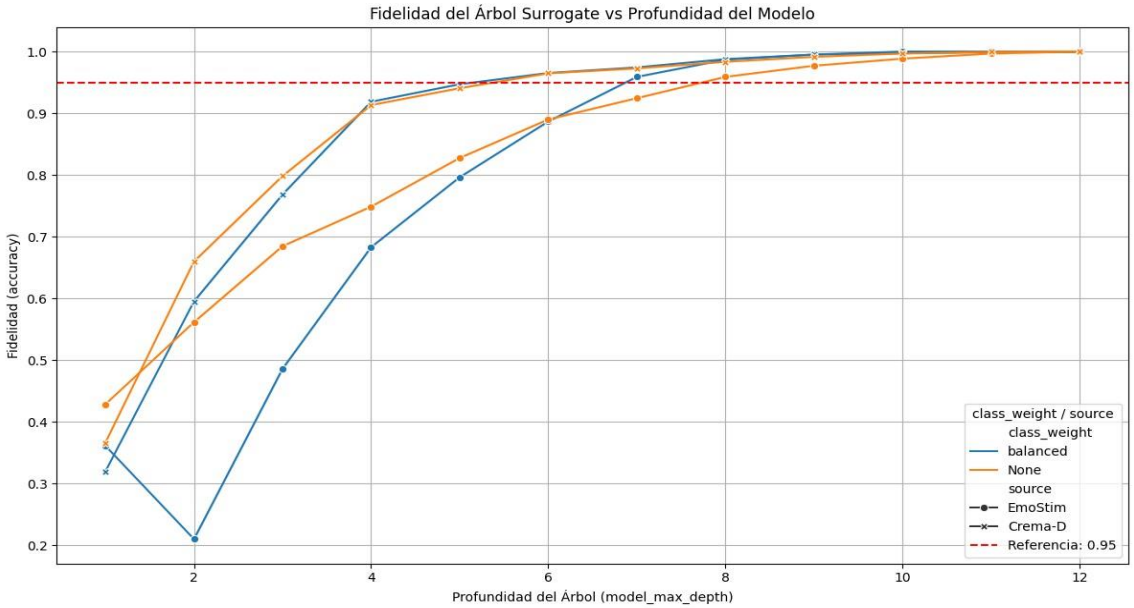


Fuente: Elaboración propia.

Modelado de Surrogate del árbol de decisión explicativos de embedding

El Árbol Surrogate permitió analizar hasta qué punto los embeddings podían explicarse según la profundidad del modelo y la configuración de pesos de clase. Como se muestra en la Figura 23, la fidelidad aumenta con mayor profundidad, alcanzando y superando el umbral de referencia de 0.95 en varias combinaciones. Sin embargo, este incremento también implica que la complejidad crece de forma drástica en cada nivel, lo que reduce el control sobre la interpretación de los embeddings y puede generar dudas sobre si estos aportan un efecto sustancial, en especial cuando se busca explicar emociones básicas con un árbol de decisión.

Figura 23 *Fidelidad del Árbol Surrogate según la profundidad del modelo*



Fuente: Elaboración propia.

Por esta razón, se toma como referencia el primer modelo que supera el umbral de 0.95, ya que representa el mejor equilibrio entre simplicidad e interpretabilidad. En este caso, los primeros en alcanzarlo fueron CREMA-D (sin ajuste de pesos) y EmoStim (con balanceo de clases), ambos con profundidad 8.

La Tabla 16 resume estas configuraciones, donde además de la fidelidad se incluyeron métricas obtenidas con regresión Lasso: el MSE promedio, que indica la desviación media de la predicción, y el R^2 promedio, que refleja qué tan bien se explica la variabilidad de los datos. El R^2 se usó como criterio principal para elegir la mejor variante de cada modelo, dado que ofrece una medida más directa de la capacidad explicativa.

Tabla 16 *Resumen de métricas por configuración*

model	depth	class-weight	fidelity	n-emb	avg-lasso-mse	avg-lasso-r2
CREMA-D	8	None	0.983422	42	0.001642	0.197127
EMO-STIM	8	balanced	0.980296	68	0.000765	0.13398

Fuente: Elaboración propia.

RESULTADOS

A partir de las configuraciones seleccionadas durante el desarrollo de la propuesta metodológica, se analizaron dos modelos surrogate del árbol de decisión explicativo de embeddings, correspondientes a los conjuntos CREMA-D (None) y EMO-STIM (Balanced). Se utilizaron los parámetros definidos durante el diseño experimental, incluyendo la misma profundidad máxima, umbral mínimo de fidelidad y el valor promedio de Lasso R^2, para evaluar la capacidad explicativa de las variables involucradas.

En la Tabla 17 y Tabla 18 se presentan los resultados de los modelos EMO-STIM y CREMAD respectivamente. El avance en los resultados se dividió en diez tramos, definidos por los umbrales de R^2 que marcan los puntos de mayor cambio. En cada tramo se reportan la cantidad de embeddings que superan el umbral, el porcentaje que representan respecto al total y el número de características vocales asociadas. Se indica la utilidad relativa de cada rango. El umbral de referencia se ubica en valores superiores a 0.5. Para más detalles sobre el desempeño completo, consulte el Anexo F.10

Se indican la utilidad relativa de cada rango de R^2.

Tabla 17 *Desempeño general en películas del modelo EMO-STIM (Balanced)*

Umbral R2	Embeddings	Embeddings con fidelidad(%)	Características vocales
0.0000	59	86.76	51
0.0078	53	77.94	51
0.0246	46	67.65	51
0.0955	40	58.82	51
0.1207	33	48.53	51
0.1495	27	39.71	51
0.2022	20	29.41	51
0.2293	14	20.59	50
0.3124	7	10.29	45
0.3935	-	0.00	-

Fuente: Elaboración propia.

Tabla 18 *Desempeño general en entrenamiento del modelo CREMA-D (None)*

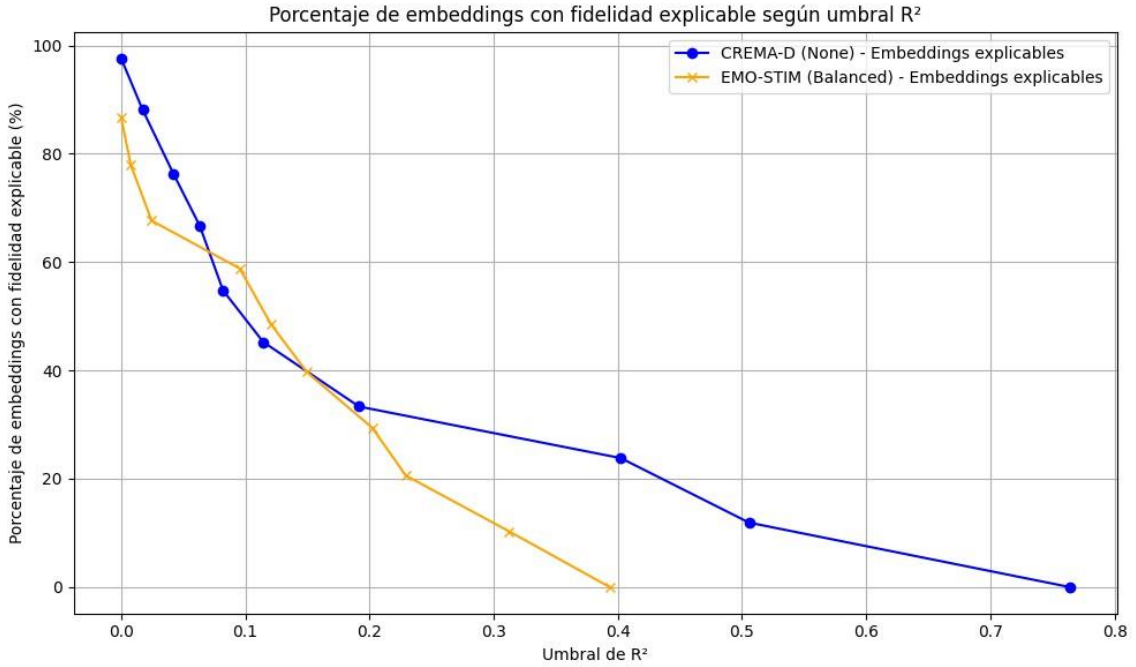
Umbral R2	Embeddings	Embeddings con fidelidad(%)	Características vocales
0.0000	41	97.62	47
0.0174	37	88.10	47
0.0424	32	76.19	47
0.0634	28	66.67	47
0.0821	23	54.76	34
0.1145	19	45.24	34
0.1916	14	33.33	31
0.4024	10	23.81	29

0.5057	5	11.90	20
0.7639	-	0.00	-

Fuente: Elaboración propia.

La Figura 24 muestra el porcentaje de embeddings explicables según el umbral de R^2, permitiendo visualizar el comportamiento de ambos modelos bajo los mismos criterios de análisis.

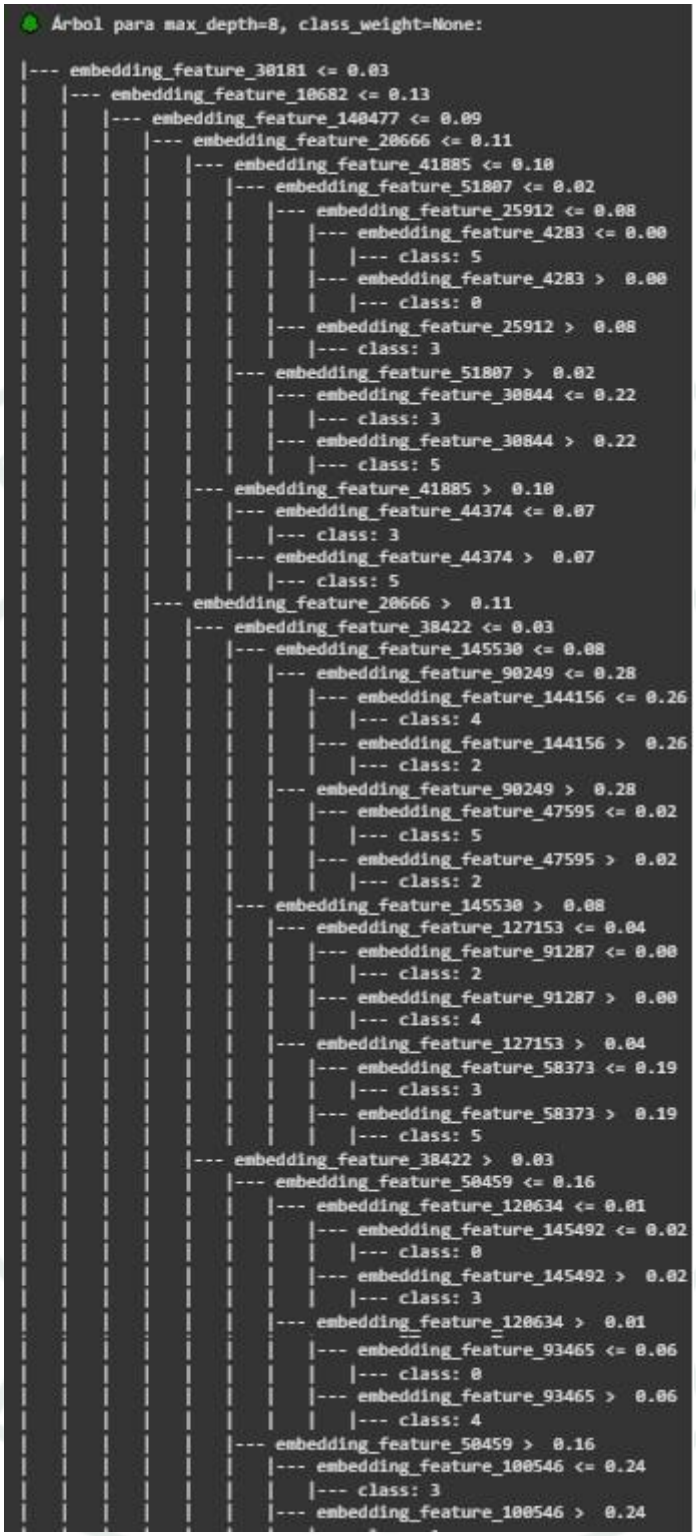
Figura 24 *Explicabilidad de los modelos*



Fuente: Elaboración propia.

En la Figura 25 se muestra el árbol de decisión general, el resto del árbol se encuentra en el Anexo F.9

Figura 25 *Vista del DT CREMA-D (None) (Parte 1)*



Fuente: Elaboración propia.

En la Tabla 19, El análisis considera umbrales R^2 mayores a 0.5, incluyendo la falta de embeddings explicables directamente. El signo - indica que un valor no aplica, mientras que - - señala que el elemento ya ha sido listado en filas anteriores.

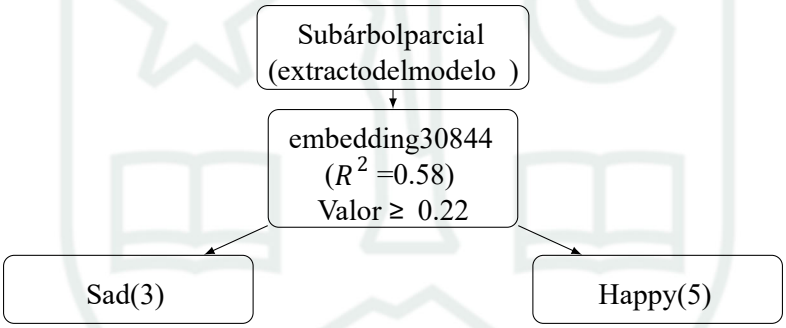
Tabla 19 *Evolución de características explicativas según el umbral* $R^2 > 0{,}5$

Umbral R^2	Emb.	Fid. (%)	Caracteristicas Vocales	Emb. Explicativos
0.764	0	-	-	-
0.651	1	2.38	10	38422
0.635	2	4.76	11	- -, 58373
0.58	3	7.14	19	- -, 30844
0.509	4	9.52	19	- -, 8601
0.503	5	11.9	20	- -, 145492
0.502	6	14.29	21	- -, 96576

Fuente: Elaboración propia.

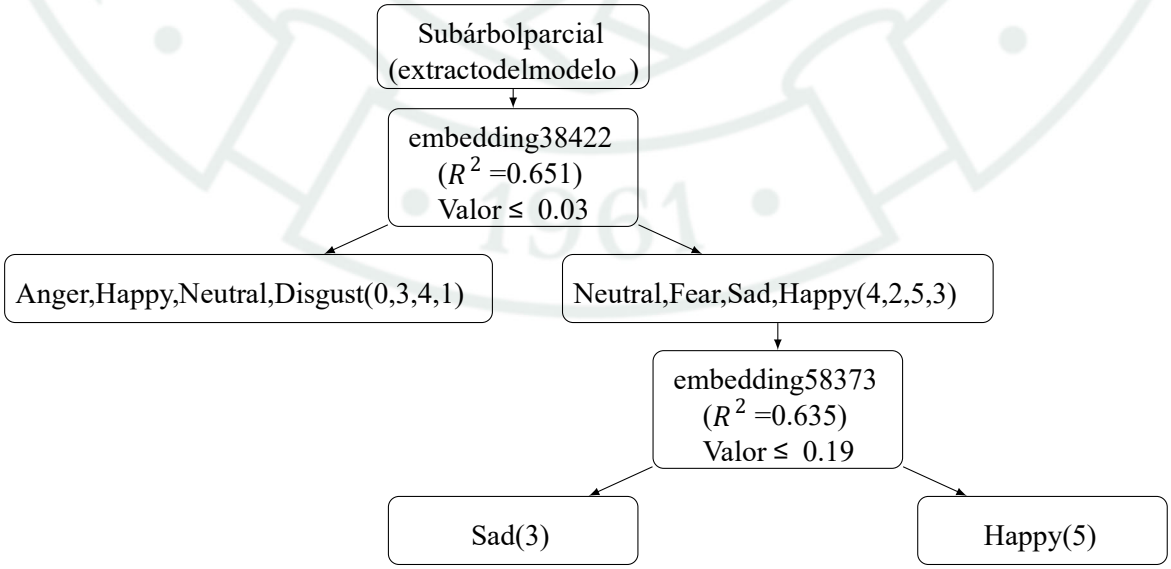
En la Figura 27, en el nivel superior o rama del arbol del embedding 38422 ($R^2 = 0.651$) con multiples clases por un lado 0,3,4,1(Anger, Happy, Neutral, Disgust) y el otro 4,2,5,3(Neutral, Fear, Sad, Happy), ademas el embedding 58373 ($R^2 = 0.635$) una hoja inferior de separación basica de categorías de 3(Sad) y 5(Happy). Que se repite en la Figura 26 con el embedding 30844 ($R^2 = 0.58$) bajo circunstancias todavía más desconocidas de la Tabla 27.

Figura 26 *Subárbol de decisión del embeddings 30844*



Fuente: Elaboración propia.

Figura 27 *Subárbol de decisión de los embeddings 38422 y 58373*



Fuente:Elaboraciónpropia.

La Tabla 20 enumera las características vocales asociadas al embedding 58373, identificado como el embedding explicable completo con el mayor umbral de R^2.

Tabla 20 *Características vocales del embedding 58373*

ID	Característica
1	F0semitoneFrom27.5Hz MeanFallingSlope
2	F0semitoneFrom27.5Hz MeanRisingSlope
3	F1 Bandwidth Amean
4	F1 Frequency Amean
5	F2 AmplitudeRelF0 Amean
6	F2 Bandwidth Amean
7	F3 Frequency Amean
8	EquivalentSoundLevel dBp
9	MFCC1 Amean
10	MFCC4V StddevNorm

Fuente: Elaboración propia.

DISCUSIÓN

El principal aporte de esta investigación radica en el uso de material monodato de entrada, con audio, y datos de salida interpretables en características vocales, que contrasta con los enfoques multimodales más comunes. El reconocimiento multimodal de emociones combina varias fuentes de información para mejorar la precisión y objetividad, con técnicas avanzadas de aprendizaje automático o validación experimental (Pan et al., 2024; Erdem Güler & Patlar Akbulut, 2025; Yan et al., 2024). Fusionando diferentes modalidades como las expresiones faciales, discurso, lenguaje o movimiento corporal, para reducir el riesgo de sobreajuste o subajuste, buscando generar respuestas robustas y comprensibles. En medicina, una revisión sistemática de 44 estudios reporta que el reconocimiento emocional se analiza desde escenarios de aplicación, técnicas multimodales y objetivos temporales, mostrando la evolución tecnológica de la última década (Guo et al., 2024). Desde la perspectiva de la neurociencia afectiva, las emociones humanas se organizan en función del propósito temporal de la situación y tienen bases universales, aunque pueden diferir su respuesta dependiendo de la experiencia personal. Así, las personas con una tendencia negativa (negative affect) son más susceptibles a experimentar emociones como el miedo, la tristeza, la ira y el disgusto. De estas, tanto el enojo como el miedo se consideran respuestas defensivas agresivas. Además, ciertos elementos como la felicidad no se entiende como una emoción en sí misma, sino como el resultado de otros estados emocionales (Panksepp, 1998). En este estudio, se observó un patrón similar: en el embedding 58373 ($R^2 = 0.635$), obviando estados repetidos como resultantes de otros como la neutralidad y felicidad, las emociones negativas con tendencias defensivas se agruparon en dos categorías: una pasiva (miedo, tristeza) y otra activa (enojo, disgusto). Estas emociones se identificaron a través de 10 características vocales, como se muestra en la Tabla 20. Se describen ejemplos de este tipo en el Anexo G, donde se presentan observaciones exploratorias basadas en análisis subjetivos y datos no oficiales.

En cuanto al uso de clips de películas como fuente de datos, si bien su desempeño inicial fue superior en algunos modelos de entrenamiento, se observó una caída inesperada en la explicabilidad, especialmente en los valores de R^2. Esto coincide con antecedentes donde las películas se han empleado en contextos exploratorios, mostrando que su validez depende de criterios de selección estrictos y de tamaños reducidos de muestra (Michelini et al., 2019; Fernández Megías et al., 2011; Michelini et al., 2015).

El uso de muestras de películas para el reconocimiento emocional se fundamenta en la gramática cinematográfica y en la experiencia afectiva de los directores, lo que se conoce como entendimiento afectivo. En un análisis de 2040 escenas de 36 películas de Hollywood, se obtuvo un 45% de precisión sin utilizar vectores de escenas afectivas (SAV) y un 65% al incorporarlos. Al combinar estos vectores con elementos prosódicos basados en las emociones básicas de Ekman (felicidad, sorpresa, ira, tristeza, miedo y asco), la precisión aumentó al 78.1%; estos resultados se estandarizan las características de los clips desde perspectivas cinematográficas,

psicológicas y cognitivas, aplicando un modelo de valencia y activación (Valence and Arousal, VA); sin embargo, el análisis presenta desafíos relevantes, como los errores de clasificación híbridos (manuales y automáticos), la presencia de casos ambiguos o borderline, la subjetividad emocional, por ejemplo, la distinta tolerancia individual al miedo, y la sensibilidad a cambios emocionales sutiles; también influyen factores como las características prosódicas o psicofisiológicas de la voz, la mayor riqueza informativa del audio frente al video, el tipo de escena (monólogos, múltiples hablantes o silencios), la segmentación óptima de 2, 4 u 8 segundos, y la concatenación de segmentos emocionales. Estos aspectos representan tanto ventajas como limitaciones que se han refinado a lo largo de más de 50 años de investigación (Wang & Cheong, 2006). En los experimentos, la muestra sin categorías afectivas alcanzó un coeficiente de determinación R^2 = 0.39 en 14 películas y 21 escenas, mientras que el entrenamiento con la batería de Ekman, explicado mediante embeddings de características vocales, alcanzó un R^2 = 0.76. Según los modelos auxiliares DT, la segmentación más efectiva fue de 1 segundo, las características más relevantes fueron las prosódicas, y la concatenación de segmentos no mostró mejoras significativas.

Además, en el Anexo G, en notas exploratorias perceptuales, que no estan comprobadas, se aprovecharon todas las metricas disponibles como las emociones percibidas o la intensidad de las peliculas, el comportamiento general de ejemplos de medios contemporaneos de gusto personal, en especial los que su comportamiento fue inusual o contradictorios a la literatura o lo visto en la batería de peliculas. Las clasificaciones combinadas sugeridas, se podrian asociar a cambios o grados emocionales de riqueza, variedad, emocionalidad, complejidad, transitoriedad o choque emocional(Emocion percibida vs expresada); en diferentes narrativas como monólogos, multiples lineas o naturalidad, usando distribución simple o borderline.

Se enfatiza la aparente relación con estudios sobre la selección de características representativas del habla usando redes neuronales densas (DNN) junto con árboles extremadamente aleatorizados (ET) para reconocimiento de emociones, alcanzando un 61.8% de Recall Promedio No Ponderado (UAR). Este estudio se enfocó en charlas en diferentes idiomas, usando los corpus IEMOCAP (inglés) y FAU Aibo (alemán), generando elementos intermedios representativos de MFCC o SDC para identificar características con i-vectors (Heracleous et al., 2019). Este caso sobresale por manejar características de voz de forma similar a la propuesta, aunque en aquel estudio los elementos intermedios solo fueron almacenados sin procesamiento adicional. No es posible una comparación directa dado que son elementos más simples que no suelen verse afectados por el idioma, como fue comprobado en las técnicas CNN empleadas, y las métricas no son comparables al usar bases de datos, pesos y categorías diferentes, ya que el R^2 encontrado aquí fue del 76% frente al 61.8% UAR, manejando cinco emociones básicas (Anger, Disgust, Fear, Happy, Sad) y una clase sin clasificar (Neutral), contra tres emociones básicas (Angry, Empathic, Joyful) y dos sin clasificar (Neutral, Rest). Esto requeriría una reestructuración del tipo DT y objetivo, aunque es factible dada la modularidad del sistema.

Complementando el área de DNN con estructuras SVM de decisión, otro estudio sobre señales de voz en emociones profundas busca extraer elementos distintivos mejor que métodos tradicionales en SVM y DNN-SVM (Sun et al., 2019). Sin embargo, no es posible realizar un análisis directo debido a diferencias en métricas y categorías: cinco emociones (angry, happy, fear, surprise, sad) y una sin clasificación (neutral).

Otros estudios no emplean el mismo enfoque interpretativo, pero sí clasificación con datos distintos en reconocimiento emocional. Que combinan CNN con clasificación para mejorar la calidad técnica de la señal (Vieira et al., 2020), y otro que propone arquitecturas híbridas para preprocesamiento y extracción de datos de procesamiento de lenguaje natural (NPL) en textos, sobre lenguaje malicioso (Charbuty & Abdulazeez, 2021). Estos estudios muestran que las técnicas especializadas en clasificación dimensional son necesarias, aunque no dividen específicamente en dimensiones relacionadas con emociones básicas, sino más bien con niveles de intensidad.

Un estudio reciente en revisión utilizó los conjuntos de características ComParE y EmoBase con OpenSMILE para la predicción de emociones, mostrando mejoras respecto a los clasificadores tradicionales (Marín Torrent, 2023). Si bien estos conjuntos ofrecieron resultados aceptables y similares, su rendimiento no superó al obtenido con GeMAPS. Este Set de Opensmile es la base estándar reconocida para medir emociones en voz, alineado con la comunidad científica, ademas de aumenta la validez de la replicación en estudios de emociones y voz, al igual que un estandar para el uso de embeddings en rasgos clásicos, al usar formantes de fonación y articulación, sin olvidar el vínculo con los mecanismos fisiológicos de producción vocal(Eybenetal.,2016).Este comportamiento es común en este tipo de tareas, donde la elección del conjunto de características impacta considerablemente los resultados, donde conjuntos compactos y especializados superar configuraciones más extensas al evitar redundancias y sobreajuste, especialmente util cuando el análisis detallado de la selección de características resulta computacionalmente costoso y complejo.

Por otro lado, analizar el enfoque evidencia la importancia de la elección de algoritmos. Puede lograrse resultados más interpretables y de latencia baja de procesamiento en 24 ms, al centrarse en características acústicas y no necesariamente en embeddings semánticos complejos, fusionando MFCC + entrenamiento auto-supervisado (Self-Supervised Learning, SSL) en discursos y texto, (Deeb et al., 2025). Aunque trabajos recientes aún en revisión, presenta un enfoque relevante de interpretabilidad, como su aplicación sensible y elementos relevantes, como los aportados por los rasgos energéticos en las emociones y datasets, concluyendo que los embeddings capturan parte de la información, pero eso no significa que sean lo más importante para el modelo. Lo que realmente puede marcar la diferencia es la información que podemos entender e interpretar (Dixit et al., 2024). Este estudio se distingue de esos enfoques que priorizan exclusivamente maximizar la precisión en el reconocimiento emocional. Partiendo de un modelo con una combinación de parámetros que fomentan alta precisión, el objetivo fue evaluar si los datos utilizados contenían patrones emocionales significativos y comprensibles.

Para ello, se priorizó el uso de herramientas y bases de datos comprobadas como OpenSMILE, CREMA-D y EMO-STIM, enfocándose en modelos explicativos, como DT y LassoCV, así como su capacidad para generar representaciones de datos. El enfoque adoptado busca identificar patrones de compatibilidad más allá del mero desempeño del modelo, como detectar integración en múltiples disciplinas, el desempeño frente a otras técnicas de explicabilidad, la correlación de los grupos de variables acústicas o su interrelación en distintos contextos sociales, cognitivos y de perspectiva.

CONCLUSIONES

1. Se logró diseñar e implementar un sistema capaz de reconocer emociones y explicar los factores que influyen en los patrones de voz, logrando la precisión esperada; incluso después del diseño híbrido completo, se lograron resultados superiores en versiones simplificadas de CNN junto con comportamiento simulado transformador, reduciendo complejidad. Las pruebas con películas, se mostraron inicialmente superiores en el entrenamiento. Pese a la complejidad de adaptación de la muestra, el uso de bases de datos preparadas con métodos auxiliares simplificadores permitió verificar pautas propias de la gramática cinematográfica. De los siete embeddings con (R^2 >0.5), seis se explican mediante características vocales. En total se identificaron 21 características vocales únicas, lo que facilita su manejo. El mejor embedding (R^2 = 0.651) agrupó las emociones en dos tendencias relacionadas con la afectividad defensiva, pasiva y activa, explicadas por solo 10 características vocales.

2. Se comprobó que los clips de películas son adecuados para representar emociones auditivas, empleando la base de datos CREMA-D. Su comportamiento fue coherente con la gramática cinematográfica, considerando criterios como predominancia hablada, curación, segmentación y balanceo. Los modelos auxiliares DT mostraron mejores resultados con muestras pequeñas y balanceadas, destacando el subgrupo A (21 películas) segmentada en 1 segundo y 54% de precisión, incluso sobre el grupo AB (42 películas). Se verificó la fiabilidad del sistema, entrenado con la batería Ekman y la base CREMA-D, al etiquetar y generar correctamente todas las representaciones emocionales. El modelo DT surrogate alcanzó fidelidad superior al 0.98 con profundidad 8 y sin balanceo, mostrando un equilibrio óptimo entre simplicidad e interpretabilidad. Además, se identificó que, según el objetivo del estudio, incluir categorías o etiquetas afectivas adicionales durante la preparación de las bases de datos podría mejorar la precisión general del sistema.

3. Las características vocales más relevantes se encontraron en el set de 88 del conjunto eGeMAPSv02 de OpenSMILE, estándar en estudios emocionales, incluido CREMA-D. Al evaluar sus variantes (GeMAPS, eGeMAPS, Emobase y ComParE), no se observaron mejoras métricas ni reducción superior al punto estandarizado. El modelo DT interpretativo seleccionó 51 en la muestra y 47 en el entrenamiento, de las cuales 21 mostraron correlación fuerte (>0.5) y 10 una correlación aún mayor 0.651. Estas características reflejan variaciones prosódicas, frecuenciales, acústicas y temporales que describen las emociones en la voz.

4. El sistema alcanzó una precisión general del 99.7%, con altos valores de recall y F1score por clase. Este resultado se atribuye a la combinación de técnicas de la literatura y al ajuste de parámetros como la dimensionalidad, escala de grises, manejo y segmentación de datos, así como el uso de embeddings y modelos auxiliares. Durante el entrenamiento, la pérdida disminuyó de 3.7 a 0.21 en las dos primeras épocas y se estabilizó en 0.0026 en la última.

5. El modelo DT surrogate explicó de forma significativa las características vocales en relación con los embeddings emocionales del modelo híbrido. Con eGeMAPSv02, la fidelidad de LassoCV superó el 98%, y la DT alcanzó 97.62% en muestra y 86.76% en entrenamiento. Los coeficientes de determinación R^2 de 0.764 y 0.394 son relevantes y comparables con estudios de mayor escala, así como los valores de MSE 0.00165 y 0.00138, y RMSE 0.0406 y 0.0372, respectivamente. El análisis mostró patrones emocionales detectables incluso sin analizar la reacción del público. Los subárboles parciales evidenciaron una estructura jerárquica binaria multinivel con un coeficiente cuadrado conjunto >= 0.635, lo que confirma la organización interna de las representaciones emocionales.

6. Se logró modelar un DT con fidelidad del 99.7%, respecto al modelo híbrido, considerado además inherentemente interpretativo en el marco XAI. El DT surrogate alcanzó fidelidad superior a 0.98 equilibrando simplicidad e interpretabilidad con profundidad 8 y umbral (>= 0.95). LassoCV permitió un entrenamiento estable (max.iter=100000). La principal área de mejora identificada fue la automatización de la interpretación de reglas embedding, actualmente analizada manualmente. Se destaca que la simplificación del Transformer en la arquitectura híbrida no comprometió la precisión, lo que sugiere que futuras iteraciones pueden priorizar interpretabilidad y eficiencia sobre complejidad arquitectónica. También se sugiere incluir etiquetas afectivas adicionales y priorizar escenas con monólogos o clasificaciones borderline, según el objetivo analítico.

RECOMENDACIONES Y TRABAJOS FUTUROS

1. Aplicación de la metodología en contextos controlados de experimentación: se sugiere el uso de espacios donde las variables externas estén reguladas, de manera que los resultados puedan atribuirse con confianza a la intervención aplicada. La principal ventaja de los audios sin NPL es que son poco invasivos y muy flexibles en distintos idiomas, lo que facilita su adaptación sin mayores complicaciones. Esto permite profundizar en aspectos que requieren mayor profundización, como la intensidad, complejidad, choque, neutralidad, variedad, emocionalidad y reacciones emocionales, siendo especialmente útil en aplicaciones sociales.

2. Validación utilizando características musicales: otra recomendación importante es validar la reproductibilidad de esta estrategia utilizando conjuntos de características musicales, como las que proporciona OpenSMILE. Estudios previos han demostrado que sets de características como GeMAPS son excelentes candidatas para ser utilizadas en investigaciones emocionales, ya que al aplicarlas se ha observado un rendimiento mejorado tanto en el análisis de la voz como en la música, identificando una carga emocional más rica en la muestra musical (Atmaja & Akagi, 2020). Esto ayudaría a fortalecer la comprensión de las emociones y a ofrecer un enfoque más completo en su estudio.

3. Explorar la reacción afectiva del público ante el doblaje: una línea de investigación futura interesante sería la clasificación afectiva a partir del reconocimiento emocional y etiquetas de reacción, como la disponible en la muestra original. Lenguas poco representadas, como el punjabi, han demostrado ser eficaces para el análisis emocional. Usando una batería estándar de emociones básicas y frases neutras (Kaur & Singh, 2021), es posible obtener resultados prometedores y compatibles con los requisitos del estudio. Investigaciones centradas en este tipo de idiomas, además de fomentar una mayor integración cultural, amplían el alcance del análisis emocional y su aplicación práctica en plataformas digitales, como el marketing y las redes sociales (Sharma et al., 2023). Esto abre nuevas posibilidades, como la delimitación del contenido en frases, lo cual podría reducir la complejidad computacional y facilitar la clasificación afectiva. Por ejemplo, al detectar el impacto emocional de escenas dobladas, es posible analizar emociones implícitas y determinar su carácter pasivo o activo, lo que permite múltiples aplicaciones.

4. Exploración de emociones complejas en el reconocimiento emocional: finalmente, en trabajos futuros sería valioso explorar las metodologias en emociones complejas, aunque esta investigación se centró principalmente en emociones básicas, se desconoce los beneficios y estrategias, como las de optimización, que podrían haber permitido una exploración más rápida y potente. En tareas de reconocimiento emocional utilizando

arquitecturas DCNN, proponen una metodología que mejora la extracción de características, reduciendo el cos-
to computacional sin sacrificar precisión, al utilizar ET provenientes del DT (Heracleous et al., 2019).

5. Analizar la generación de datos procesados por Lasso con Shapley: como trabajo futuro, consolidar los datos obtenidos mediante Lasso con el apoyo de Shapley puede facilitar su aplicación práctica y su utilización en contextos sensibles dentro de enfoques de reconocimiento emocional del discurso (Speech Emotion Recognition, SER) (Nfissi et al., 2024).

REFERENCIAS

Afzal, S., Ali Khan, H., Jalil Piran, M., & Weon Lee, J. (2024). A Comprehensive Survey on Affective Computing: Challenges, Trends, Applications, and Future Directions. *IEEE Access*, *12*, 96150-96168. https://doi.org/10.1109/access.2024.3422480

Alhussein, G., Ziogas, I., Saleem, S., & Hadjileontiadis, L. J. (2025). Speech emotion recognition in conversations using artificial intelligence: a systematic review and meta-analysis. *Artificial Intelligence Review*, *58*(7). https://doi.org/10.1007/s10462-025-11197-8

Arun, A., Rallabhandi, I., Hebbar, S., Nair, A., & Jayashree, R. (2021). Emotion Recognition in Speech Using Machine Learning Techniques. *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 01-07. https://doi.org/10.1109/icccnt51525.2021.9580028

Atmaja, B. T., & Akagi, M. (2020). On The Differences Between Song and Speech Emotion Recognition: Effect of Feature Sets, Feature Types, and Classifiers. *2020 IEEE REGION 10 CONFERENCE (TENCON)*, 968-972. https://doi.org/10.1109/tencon50793.2020.9293852

Bhanbhro, J., Memon, A. A., Lal, B., Talpur, S., & Memon, M. (2025). Speech Emotion Recognition: Comparative Analysis of CNN-LSTM and Attention-Enhanced CNN-LSTM Models. *Signals*, *6*(2), 22. https://doi.org/10.3390/signals6020022

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, *5*(4), 377-390. https://doi.org/10.1109/taffc.2014.2336244

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20-28. https://doi.org/10.38094/jastt20165

Deeb, B. M., Savchenko, A. V., & Makarov, I. (2025). Enhancing Emotion Recognition in Speech Based on Self-Supervised Learning: Cross-Attention Fusion of Acoustic and Semantic Features. *IEEE Access*, *13*, 56283-56295. https://doi.org/10.1109/access.2025.

3554454

Dixit, S., Low, D. M., Elbanna, G., Catania, F., & Ghosh, S. S. (2024). Explaining Deep Learning Embeddings for Speech Emotion Recognition by Predicting Interpretable Acoustic Features. https://doi.org/10.48550/arXiv.2409.09511

Elngar, A. A., Arafa, M., Fathy, A., Moustafa, B., Mahmoud, O., Shaban, M., & Fawzy, N. (2021). Image Classification Based On CNN: A Survey. *Journal of Cybersecurity and Information Management*, PP. 18-50. https://doi.org/10.54216/jcim.060102

Erdem Güler, S., & Patlar Akbulut, F. (2025). Multimodal Emotion Recognition: Emotion Classification Through the Integration of EEG and Facial Expressions. *IEEE Access*, *13*, 24587-24603. https://doi.org/10.1109/access.2025.3538642

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, *7*(2), 190-202. https://doi.org/10.1109/taffc.2015.2457417

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast opensource audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*, 1459-1462. https://doi.org/10.1145/1873951.1874246

Fernández Megías, C., Pascual Mateos, J. C., Soler Ribaudi, J., & García Fernández-Abascal, E. (2011). Validación española de una batería de películas para inducir emociones. *Psicothema*, *23*(4), 778-785. Consultado el 10 de noviembre de 2025, desde https://www. psicothema.com/pdf/3956.pdf

Figueredo J.N., J. N., & Castillo J.A., J. A. (2016). Evaluación de desórdenes vocales en profesionales que usan su voz como herramienta de trabajo. Occupational Voice Quick Screening. *Ciencias de la Salud*, *14*(especial), 97-112. https://doi.org/10.12804/revsalud14. especial.2016.07

González Mares, M. (2019, enero). *Hernández-Sampieri, R. & Mendoza, C (2018). Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta* (Vol. 10). Universidad Nacional Autonoma de Mexico.

https://doi.org/10.22201/fesc.20072236e.2019.10.18.6 Goudbeek, M., & Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, *128*(3), 1322-1336. https://doi.org/10.1121/1.3466853

Gross, J. J., & Levenson, R. W. (1995). Emotion elicitation using films. *Cognition & Emotion*, *9*(1), 87-108. https://doi.org/10.1080/02699939508408966

Gross, J. J., & Levenson, R. W. (1997). Hiding feelings: The acute effects of inhibiting negative and positive emotion. *Journal of Abnormal Psychology*, *106*(1), 95-103. https://doi.org/10.1037//0021-843x.106.1.95

Guo, R., Guo, H., Wang, L., Chen, M., Yang, D., & Li, B. (2024). Development and application of emotion recognition technology — a systematic literature review. *BMC Psychology*, *12*(1). https://doi.org/10.1186/s40359-024-01581-4

Guyer, J. J., Briñol, P., Vaughan-Johnston, T. I., Fabrigar, L. R., Moreno, L., & Petty, R. E. (2021). Paralinguistic Features Communicated through Voice can Affect Appraisals of Confidence and Evaluative Judgments. *Journal of Nonverbal Behavior*, *45*(4), 479-504. https://doi.org/10.1007/s10919-021-00374-2

Harar, P., Galaz, Z., Alonso-Hernandez, J. B., Mekyska, J., Burget, R., & Smekal, Z. (2018). Towards robust voice pathology detection: Investigation of supervised deep learning, gradient boosting, and anomaly detection approaches across four databases. *Neural Computing and Applications*, *32*(20), 15747-15757. https://doi.org/10.1007/s00521-018-3464-7

Heracleous, P., Mohammad, Y., & Yoneyama, A. (2019). Deep Convolutional Neural Networks for Feature Extraction in Speech Emotion Recognition. En *Human-Computer Interaction. Recognition and Interaction Technologies* (pp. 117-132). Springer International Publishing. https://doi.org/10.1007/978-3-030-22643-5_9

Idrogo Zamora, D. I., & Asenjo-Alarcón, J. A. (2021). Relación entre inteligencia emocional y rendimiento académico en estudiantes universitarios peruanos. *Revista Investigación de*

Psicología, (26), 69-79. https://doi.org/10.53287/ryfs1548js42x

Islam, K., & ElSayed, Z. (2024). Speech-Based Emotion Recognition and PTSD Detection through Machine and Deep Learning. *International Journal of Computer Engineering in Research Trends*, *11*(3), 46-53. https://doi.org/10.22362/ijcert/2024/v11/i3/v11i306

Johnson, D. S., Hakobyan, O., Paletschek, J., & Drimalla, H. (2025). Explainable AI for Audio and Visual Affective Computing: A Scoping Review. *IEEE Transactions on Affective Computing*, *16*(2), 518-536. https://doi.org/10.1109/taffc.2024.3505269

Joint Commission International. (2021). Communicating Clearly and Effectively to Patients: How to Overcome Common Communication Challenges in Health Care. Consultado el 10 de noviembre de 2025, desde https://www.jointcommissioninternational.org/ what-we-offer/publications/white-papers/communicating-clearly-and-effectively-topatients/

Josephine Mary Juliana, M., Sudha, G. F., & Nakkeeran, R. (2023). Diagnosis of Post Traumatic Stress Disorder through Speech-Based Emotion Modelling with 1D CNN, 1-6. https: //doi.org/10.1109/icacic59454.2023.10435066

Käsermann, M.-L., Altorfer, A., Foppa, K., Jossen, S., & Zimmermann, H. (2000). The study of emotional processes in communication: I. Measuring emotionalization in everyday faceto-face communicative interaction. *Behavior Research Methods, Instruments & Computers*, *32*(1), 33-46. https://doi.org/10.3758/bf03200786

Kashef, R. (2022). ECNN: Enhanced convolutional neural network for efficient diagnosis of autism spectrum disorder. *Cognitive Systems Research*, *71*, 41-49. https://doi.org/10. 1016/j.cogsys.2021.10.002

Kaur, K., & Singh, P. (2021). Punjabi Emotional Speech Database: Design, Recording and Verification. *International Journal of Intelligent Systems and Applications in Engineering*, *9*(4), 205-208. https://doi.org/10.18201/ijisae.2021473641

Kawade, R., Konade, R., Majukar, P., & Patil, S. (2022). Speech Emotion Recognition Using 1D CNN-LSTM Network on Indo-Aryan Database. *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*,

1288-1293. https://doi.org/10.1109/icicict54557.2022.9917635

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access*, *7*, 117327-117345. https://doi.org/10.1109/access.2019.2936124

Landgraf, Z. (2021, septiembre). AI Basics. Consultado el 10 de noviembre de 2025, desde https://ai4health.io/wp-content/uploads/2021/10/Artificial-Intelligence-Basics.pdf

Li,H.,Li,J.,Liu,H.,Liu,T.,Chen,Q.,&You,X.(2024).MelTrans:Mel-SpectrogramRelationship-Learning for Speech Emotion Recognition via Transformers. *Sensors*, *24*(17), 5506. https://doi.org/10.3390/s24175506

Li, J., Zhang, X., Huang, L., Li, F., Duan, S., & Sun, Y. (2022). Speech Emotion Recognition Using a Dual-Channel Complementary Spectrogram and the CNN-SSAE Neutral Network. *Applied Sciences*, *12*(19), 9518. https://doi.org/10.3390/app12199518

Liscombe, J. J. (2007). *Prosody and speaker state: paralinguistics, pragmatics, and proficiency* [Tesis doctoral] [AAI3285119]. Columbia University. https://www.cs.columbia.edu/nlp/theses/jackson_liscombe.pdf

Lok, E. J., & Cooper, D. (2025). CREMA-D Dataset [[Conjunto de datos]. Kaggle. ODC Attribution License (ODC-By)].

Lu, N., & Hao, L. (2023). Deep Learning Based Emotion Recognition Algorithm for Digital Music Speech. *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)*, 1-6. https://doi.org/10.1109/nmitcon58196.2023.10276005

Luengo, I., Navas, E., Hernáez, I., & Sánchez, J. (2005). Reconocimiento automático de emociones utilizando parámetros prosódicos. *Procesamiento del Lenguaje Natural*, *35*(35), 13-20. https://www.redalyc.org/articulo.oa?id=515751735002

Marghescu, B., Toma, Ş.-A., Morogan, L., & Bica, I. (2023a). Speech Emotion Recognition for Emergency Services. *2023 International Conference on Speech Technology and HumanComputer Dialogue (SpeD)*, 105-110. https : / / doi . org / 10 . 1109 / sped59241 . 2023 . 10314876

Marghescu, B., Toma, Ş.-A., Morogan, L., & Bica, I. (2023b). Speech Emotion Recognition for Emergency Services, 105-110. https://doi.org/10.1109/sped59241.2023.10314876

Marín Torrent, I. (2023, julio). *Uso de la Codificación ComParE en la Detección de Emociones*

[Unpublished]. Consultado el 10 de noviembre de 2025, desde https://oa.upm.es/75485/

Michelini, Y., Acuña, I., & Godoy, J. C. (2015). CARACTERÍSTICAS DE LA EXPERIENCIA EMOCIONAL INDUCIDA MEDIANTE FRAGMENTOS DE PELÍCULAS EN UNA MUESTRA DE JÓVENES ARGENTINOS. *Interdisciplinaria: Revista de Psicología y Ciencias Afines*, *32*(2). https://doi.org/10.16888/interd.2015.32.2.10

Michelini, Y., Acuña, I., Guzmán, J. I., & Godoy, J. C. (2019). LATEMO-E: A Film Database to Elicit Discrete Emotions and Evaluate Emotional Dimensions in Latin-Americans. *Temas em Psicologia*, *27*(2), 473-490. https://doi.org/10.9788/tp2019.2-13

Mohammadi, G., Vuilleumier, P., & Somarathna, R. (2023, mayo). EmoStim Dataset. https://doi.org/10.26037/yareta:usqhnbgauvgx3ggf3bqbrwtjsa

Mustafa, H. H., Darwish, N. R., & Hefny, H. A. (2024). Automatic Speech Emotion Recognition: a Systematic Literature Review. *International Journal of Speech Technology*, *27*(1), 267-285. https://doi.org/10.1007/s10772-024-10096-7

Nfissi, A., Bouachir, W., Bouguila, N., & Mishara, B. (2024). Unveiling hidden factors: explainable AI for feature boosting in speech emotion recognition. *Applied Intelligence*, *54*(11–12), 7046-7069. https://doi.org/10.1007/s10489-024-05536-5

Oritsegbemi, O. (2023). Human Intelligence versus AI: Implications for Emotional Aspects of Human Communication. *Journal of Advanced Research in Social Sciences*, *6*(2), 76-85. https://doi.org/10.33422/jarss.v6i2.1005

Pan, J., Fang, W., Zhang, Z., Chen, B., Zhang, Z., & Wang, S. (2024). Multimodal Emotion Recognition Based on Facial Expressions, Speech, and EEG. *IEEE Open Journal of Engineering in Medicine and Biology*, *5*, 396-403. https://doi.org/10.1109/ojemb.2023.3240280

Panksepp, J. (1998, septiembre). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University PressNew York, NY. https://doi.org/10.1093/oso/9780195096736.001.0001

Project Management Institute. (2013). The High Cost of Low Performance: The Essential Role of Communications. https://www.pmi.org/learning/library/en-2013-pulse-high-costlow-performance-13512

Riehle, A., Braun, B. I., & Hafiz, H. (2013). Improving Patient and Worker Safety: Exploring Opportunities for Synergy. *Journal of Nursing Care Quality*, *28*(2), 99-102. https://doi.org/10.1097/ncq.0b013e3182849f4a

Sharma, P., Rani, S., & Singh, P. (2023). Emotion Detection in Punjabi Audio: A CNN-Based Sentimental Analysis, 855-858. https://doi.org/10.1109/iciip61524.2023.10537761

Somarathna, R., Vuilleumier, P., & Mohammadi, G. (2024). EmoStim: A Database of Emotional FilmClipsWithDiscreteandComponentialAssessment.*IEEETransactionsonAffective Computing*, *15*(3), 1202-1212. https://doi.org/10.1109/taffc.2023.3328900

Stegmann, G. M., Hahn, S., Liss, J., Shefner, J., Rutkove, S. B., Kawabata, K., Bhandari, S., Shelton, K., Duncan, C. J., & Berisha, V. (2020). Repeatability of Commonly Used SpeechandLanguageFeaturesforClinicalApplications.*DigitalBiomarkers*,*4*(3),109-122. https://doi.org/10.1159/000511671

Sun, L., Zou, B., Fu, S., Chen, J., & Wang, F. (2019). Speech emotion recognition based on DNN-decision tree SVM model. *Speech Communication*, *115*, 29-37. https://doi.org/10.1016/j.specom.2019.10.004

Totaro, P. (2021). Emotion, rationality, and social identity: a theoretical–methodological proposal for a cognitive approach. *Cognitive Processing*, *22*(4), 579-592. https://doi.org/10. 1007/s10339-021-01030-9

Trinh Van, L., H. Nguyen, Q., & Dao Thi Le, T. (2022). Emotion Recognition with Capsule Neural Network. *Computer Systems Science and Engineering*, *41*(3), 1083-1098. https://doi.org/10.32604/csse.2022.021635

Turković, B., Talović, M., Hodžić, A., Ormanović, Š., & Ćirić, A. (2022). Communication and Emotion at Workplace – Systematic Review. *International Journal of Education and Teaching*, *2*(1), 69-76. https://doi.org/10.51483/ijedt.2.1.2022.69-76

Ubur, S. D., & Gracanin, D. (2025). Narrative Review of Emotional Expression Support in XR: Psychophysiology of Speech-to-Text Interfaces. https://doi.org/10.48550/arXiv.2405.13924

UNESCO. (2015, enero). Education for All Global Monitoring Report 2015: Education for All 2000-2015: Achievements and challenges. https://doi.org/10.54676/lbsf6974

Vardhan, J. V., Kalyan Chakravarti, Y., & Chand, A. J. (2024). Emotion Recognition by Facial Expressions and Speech Using Deep Learning. *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 1-7. https://doi.org/10.1109/icaect60202.2024.10469626

Vargas Mamani, J. D., & Villanueva Villena, V. M. (2022). Inteligencia emocional y estrés laboral en personal de salud del Centro de Salud Miraflores, Arequipa 2021 [Consultado el 10 de noviembre de 2025]. Consultado el 10 de noviembre de 2025, desde https: //repositorio.ucsm.edu.pe/handle/20.500.12920/11758

Vieira, S. T., Rosa, R. L., & Rodríguez, D. Z. (2020). A Speech Quality Classifier based on Tree-CNN Algorithm that Considers Network Degradations. *Journal of communications software and systems*, *16*(2), 180-187. https://doi.org/10.24138/jcomss.v16i2.1032

Wang, H. L., & Cheong, L.-F. (2006). Affective understanding in film. *IEEE Transactions on Circuits and Systems for Video Technology*, *16*(6), 689-704. https://doi.org/10.1109/ tcsvt.2006.873781

Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Mansor, H., Kartiwi, M., & Ismail, N. (2020). Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks. *2020 6th International Conference on Wireless and Telematics (ICWT)*, 1-6. https://doi.org/10.1109/icwt50448.2020.9243622

World Health Organization. (2022). Mental health at work: Policy brief. Consultado el 10 de noviembre de 2025, desde https://www.who.int/publications/i/item/9789240057944

Yan, J., Li, P., Du, C., Zhu, K., Zhou, X., Liu, Y., & Wei, J. (2024). Multimodal Emotion Recognition Based on Facial Expressions, Speech, and Body Gestures. *Electronics*, *13*(18),

3756. https://doi.org/10.3390/electronics13183756

Yu, B., & Kumbier, K. (2018). Artificial intelligence and statistics. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), 6-9. https://doi.org/10.1631/fitee.1700813

Zaman, M., & Hassan, A. (2021). Fuzzy Heuristics and Decision Tree for Classification of Statistical Feature-Based Control Chart Patterns. *Symmetry*, *13*(1), 110. https://doi.org/10. 3390/sym13010110

Zhang, Q., Yang, Y., Ma, H., & Wu, Y. N. (2019). Interpreting CNNs via Decision Trees. *2019 IEEE/CVFConferenceonComputerVisionandPatternRecognition(CVPR)*,6254-6263. https://doi.org/10.1109/cvpr.2019.00642

Zlatintsi, A., Koutras, P., Evangelopoulos, G., Malandrakis, N., Efthymiou, N., Pastra, K., Potamianos, A., & Maragos, P. (2017). COGNIMUSE: a multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. *EURASIP Journal on Image and Video Processing*, *2017*(1). https://doi.org/10.1186/ s13640-017-0194-1

ANEXO

Anexo A: Glosario

Comunicación emocional: Capacidad de expresar y entender emociones usando la voz, gestos o expresiones.

Patrones de audio vocal: Rasgos del sonido de la voz (como tono, ritmo e intensidad) que reflejan emociones.

CNN (Redes Neuronales Convolucionales): Tipo de modelo de AI que analiza imágenes o sonidos (como espectrogramas) para detectar patrones.

Convolucional: Un token de una operación matemática entre matrices.

Transformers: Modelo de AI que entiende secuencias (como la voz) y capta relaciones entre sonidos a lo largo del tiempo.

Embeddings: Representaciones numéricas de datos (como fragmentos de voz) para que puedan ser procesados por un modelo de AI.

CREMA-D: Base de datos con voces de actores que expresan emociones controladas. **EMO-STIM**: Conjunto de escenas de películas usadas para estudiar cómo las personas perciben emociones.

Espectrograma Mel: Imagen que muestra cómo cambian los sonidos de la voz a lo largo del tiempo, según cómo los percibe el oído humano.

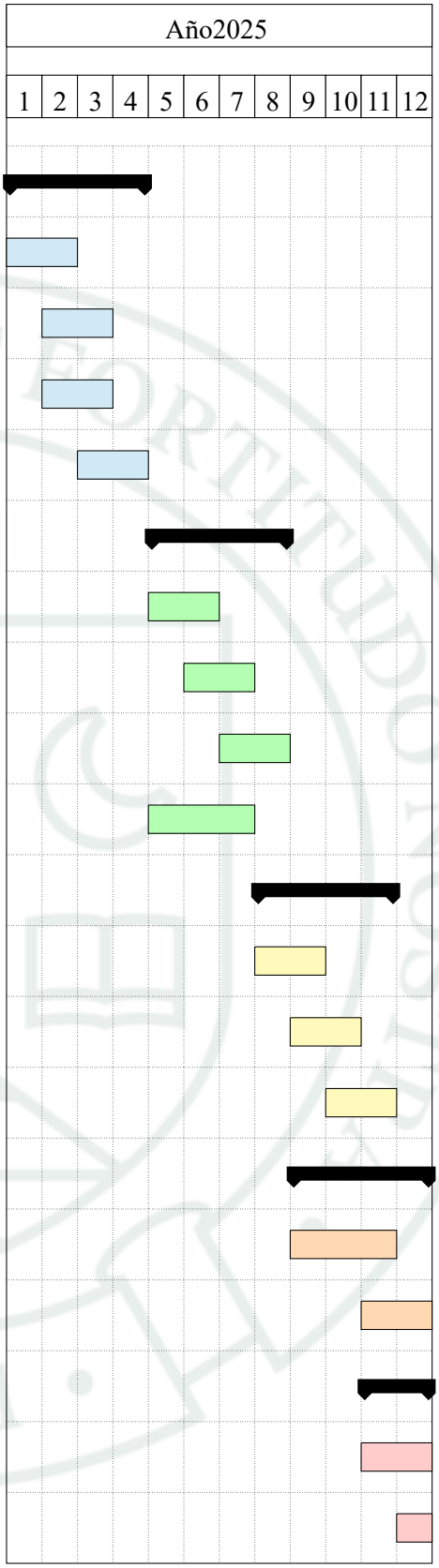
Interpretabilidad: Grado en que se puede entender cómo un modelo de AI toma decisiones.

Modelointerpretativosurrogate:Modelo simpleque imita auno complejopara entender cómo toma decisiones.

Fidelidad del modelo: Qué tan bien un modelo explicativo representa al modelo original.

Anexo B: Plan de trabajo

Figura 28 *Plan de trabajo*

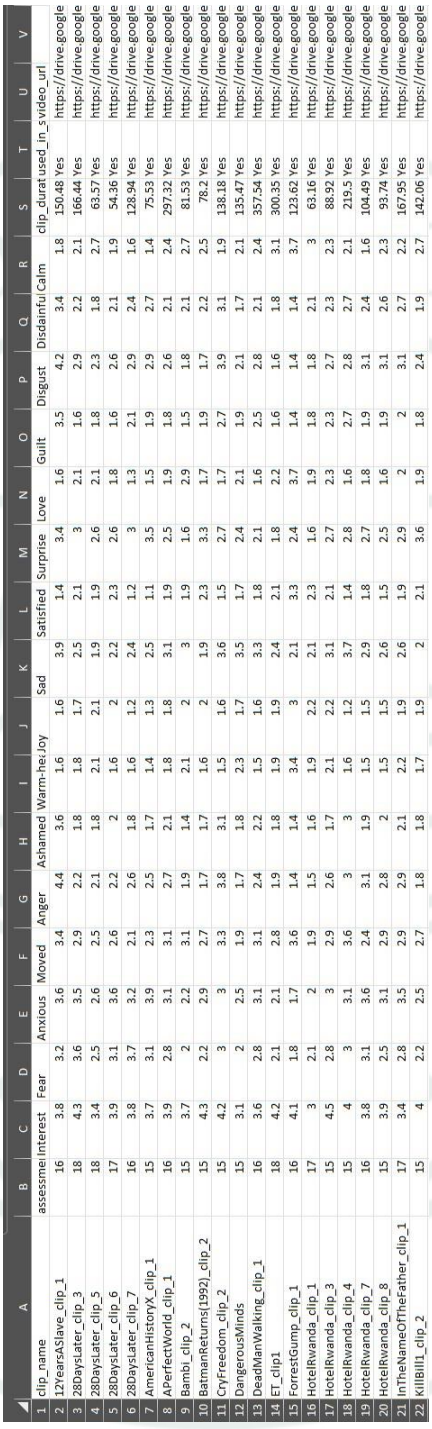


Fuente: Elaboración propia.

Cronograma detallado del proyecto, estructurado por fases clave: definición, desarrollo del modelo, integración, evaluación y cierre. Se describen las actividades distribuidas a lo largo del año 2025, incluyendo la selección de datos, desarrollo y entrenamiento de modelos, integración de módulos de IA y análisis de resultados. Ofrece una visión clara y secuencial para una gestión eficiente del tiempo.

Anexo C: Muestra original (grupo A)

Figura 29 *Muestra original (grupo A)*



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	clip_name	assessme	Interest	Fear	Anxious	Moved	Anger	Ashamed	Warm-he	Joy	Sad	Satisfied	Surprise	Love	Guilt	Disgust	Disainfu	Calm	clip_durat	used_in_s	video_url	
2	12YearsASlave_clip_1	16	3.8	3.2	3.6	3.4	4.4	3.6	1.6	1.6	3.9	1.4	3.4	1.6	3.5	4.2	3.4	1.8	150.48	Yes	https://drive.google	
3	28DaysLater_clip_3	18	4.3	3.6	3.5	2.9	2.2	1.8	1.8	1.7	2.5	2.1	3	2.1	1.6	2.9	2.2	2.1	166.44	Yes	https://drive.google	
4	28DaysLater_clip_5	18	3.4	2.5	2.6	2.5	2.1	1.8	2.1	2.1	1.9	1.9	2.6	2.1	1.8	2.3	1.8	2.7	63.57	Yes	https://drive.google	
5	28DaysLater_clip_6	17	3.9	3.1	3.6	2.6	2.2	2	1.6	2	2.2	2.3	2.6	1.8	1.6	2.6	2.1	1.9	54.36	Yes	https://drive.google	
6	28DaysLater_clip_7	16	3.8	3.7	3.2	2.1	2.6	1.8	1.6	1.2	2.4	1.2	3	1.3	2.1	2.9	2.4	1.6	128.94	Yes	https://drive.google	
7	AmericanHistoryX_clip_1	15	3.7	3.1	3.9	2.3	2.5	1.7	1.4	1.3	2.5	1.1	3.5	1.5	1.9	2.9	2.7	1.4	75.53	Yes	https://drive.google	
8	APerfectWorld_clip_1	16	3.9	2.8	3.1	3.1	2.7	2.1	1.8	1.8	3.1	1.9	2.5	1.9	1.8	2.6	2.1	2.4	297.32	Yes	https://drive.google	
9	Bambi_clip_2	15	3.7	2	2.2	3.1	1.9	1.4	2.1	2	3	1.9	1.6	2.9	1.5	1.8	2.1	2.7	81.53	Yes	https://drive.google	
10	BatmanReturns(1992)_clip_2	15	4.3	2.2	2.9	2.7	1.7	1.7	1.6	2	1.9	2.3	3.3	1.7	1.9	1.7	2.2	2.5	78.2	Yes	https://drive.google	
11	CryFreedom_clip_2	15	4.2	3	3	3.3	3.8	3.1	1.5	1.6	3.6	1.5	2.7	1.7	2.7	3.9	3.1	1.9	138.18	Yes	https://drive.google	
12	DangerousMinds	15	3.1	2	2.5	1.9	1.7	1.8	2.3	1.7	3.5	1.7	2.4	2.1	1.9	2.1	1.7	2.1	135.47	Yes	https://drive.google	
13	DeadManWalking_clip_1	16	3.6	2.8	3.1	3.1	2.4	2.2	1.5	1.6	3.3	1.8	2.1	1.6	2.5	2.8	2.1	2.4	357.54	Yes	https://drive.google	
14	ET_clip1	18	4.2	2.1	2.1	2.8	1.9	1.8	1.9	1.9	2.4	2.1	1.8	2.2	2.2	1.6	1.8	3.1	300.35	Yes	https://drive.google	
15	ForrestGump_clip_1	16	4.1	1.8	1.7	3.6	1.4	1.4	3.4	3	2.1	3.3	2.4	3.7	1.4	1.4	1.4	3.7	123.62	Yes	https://drive.google	
16	HotelRwanda_clip_1	17	3	2.1	2	1.9	1.5	1.6	1.9	2.2	2.1	2.3	1.6	1.9	1.8	1.8	2.1	3	63.16	Yes	https://drive.google	
17	HotelRwanda_clip_3	15	4.5	2.8	3	2.9	2.6	1.7	2.1	2.2	3.1	2.1	2.7	2.3	2.3	2.7	2.3	2.3	88.92	Yes	https://drive.google	
18	HotelRwanda_clip_4	15	4	3	3.1	3.6	3	3	1.6	1.2	3.7	1.4	2.8	1.6	2.7	2.8	2.7	2.1	219.5	Yes	https://drive.google	
19	HotelRwanda_clip_7	16	3.8	3.1	3.6	2.4	3.1	1.9	1.5	1.5	2.9	1.8	2.7	1.8	1.9	3.1	2.4	1.6	104.49	Yes	https://drive.google	
20	HotelRwanda_clip_8	15	3.9	2.5	3.1	2.9	2.8	2	1.5	1.5	2.6	1.5	2.5	1.6	1.9	3.1	2.6	2.3	93.74	Yes	https://drive.google	
21	InTheNameOfTheFather_clip_1	17	3.4	2.8	3.5	2.9	2.9	2.1	2.2	1.9	2.6	1.9	2.9	2	2	3.1	2.7	2.2	167.95	Yes	https://drive.google	
22	KillBill1_clip_2	15	4	2.2	2.5	2.7	1.8	1.8	1.7	1.9	2	2.1	3.6	1.9	1.8	2.4	1.9	2.7	142.06	Yes	https://drive.google	

Fuente: Tomado de FilmClipsDetails del dataset EmoStim (Mohammadi et al., 2023)

Fragmento del dataset original con tiempos, uso de clips y múltiples clasificaciones emocionales. Incluye tanto emociones expresadas como reacciones internas de interés o evaluación, aunque muchas no fueron aprovechadas en el modelo final. Este registro permite trazar relaciones entre expresión y reacción emocional, más allá de las categorías seleccionadas.

Anexo D: Características vocales completa de CREMA-D (None)

El habla cambia constantemente,se mide con características acústicas organizadas por impacto y rango (Stegmann et al., 2020). En la tabla se destacan la frecuencia fundamental (F0), vinculada a la vibración de las cuerdas vocales; los coeficientes MFCC, que reflejan energía en distintas bandas; y los formantes F1, F2 y F3, que indican picos de resonancia en el tracto vocal.

Tabla 21 *Características vocales completa de CREMA-D (None) (Parte 1)*

ID	Característica	ID	Característica
1	F0semitoneFrom27.5Hz.sma3nz.meanFallingSlope	24	logRelF0-H1-A3.sma3nz.amean
2	F0semitoneFrom27.5Hz.sma3nz.meanRisingSlope	25	logRelF0-H1-H2.sma3nz.amean
3	F0semitoneFrom27.5Hz.sma3nz.pctlrange0-2	26	logRelF0-H1-H2.sma3nz.stddevNorm
4	F0semitoneFrom27.5Hz.sma3nz.percentile20.0	27	loudnessPeaksPerSec
5	F0semitoneFrom27.5Hz.sma3nz.percentile50.0	28	loudness.sma3.meanFallingSlope
6	F0semitoneFrom27.5Hz.sma3nz.percentile80.0	29	loudness.sma3.meanRisingSlope
7	F0semitoneFrom27.5Hz.sma3nz.stddevFallingSlope	30	loudness.sma3.stddevRisingSlope
8	F0semitoneFrom27.5Hz.sma3nz.stddevRisingSlope	31	mfcc1V.sma3nz.amean
9	F1amplitudeLogRelF0.sma3nz.amean	32	mfcc1.sma3.amean
10	F1bandwidth.sma3nz.amean	33	mfcc1.sma3.stddevNorm
11	F1frequency.sma3nz.amean	34	mfcc2V.sma3nz.amean

Fuente: Elaboración propia.

Tabla 22 *Características vocales completa de CREMA-D (None) (Parte 2)*

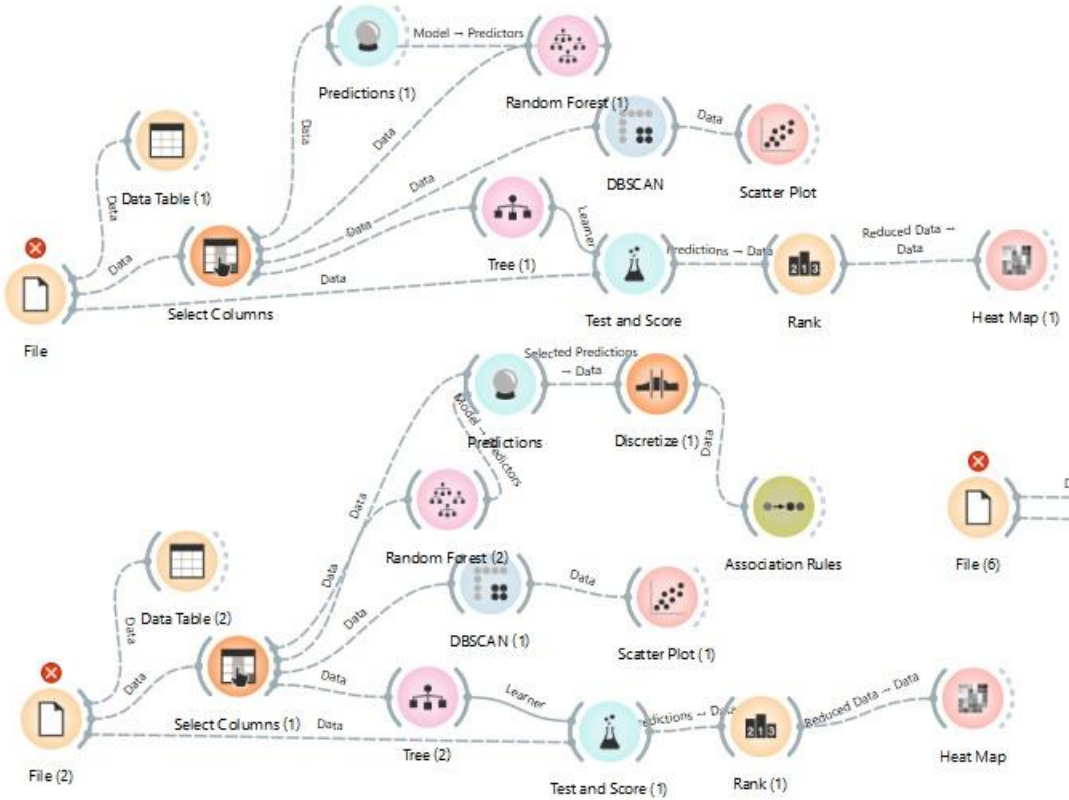
ID	Característica	ID	Característica
12	F2amplitudeLogRelF0.sma3nz.amean	35	mfcc2V.sma3nz.stddevNorm
13	F2bandwidth.sma3nz.amean	36	mfcc2.sma3.amean
14	F2frequency.sma3nz.amean	37	mfcc2.sma3.stddevNorm
15	F3amplitudeLogRelF0.sma3nz.amean	38	mfcc3V.sma3nz.amean
16	F3bandwidth.sma3nz.amean	39	mfcc3V.sma3nz.stddevNorm
17	F3frequency.sma3nz.amean	40	mfcc3.sma3.amean
18	HNRdBACF.sma3nz.stddevNorm	41	mfcc3.sma3.stddevNorm
19	alphaRatioUV.sma3nz.amean	42	mfcc4V.sma3nz.amean
20	alphaRatioV.sma3nz.amean	43	mfcc4V.sma3nz.stddevNorm
21	equivalentSoundLevel.dBp	44	mfcc4.sma3.amean
22	hammarbergIndexUV.sma3nz.amean	45	mfcc4.sma3.stddevNorm
23	hammarbergIndexV.sma3nz.amean	46	slopeV0-500.sma3nz.stddevNorm
		47	slopeV500-1500.sma3nz.stddevNorm

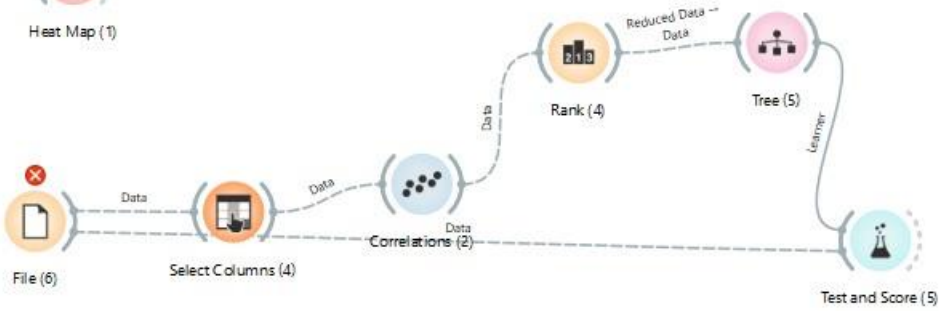
Fuente: Elaboración propia.

Anexo E: Flujo de trabajo en Orange

Se analiza el comportamiento de los datos de los grupos A y B, conformados por 21 clips, utilizando modelos auxiliares que permiten obtener una visión general del conjunto de datos. Este análisis sirve como base para seleccionar la arquitectura más adecuada. Para ello, se emplean métricas y técnicas como PCA, DBSCAN y Random Forest. Posteriormente, según el tipo de datos, se generan visualizaciones y resultados como rankings, mapas de calor, análisis de correlación, matrices y modelos de asociación o predicción.

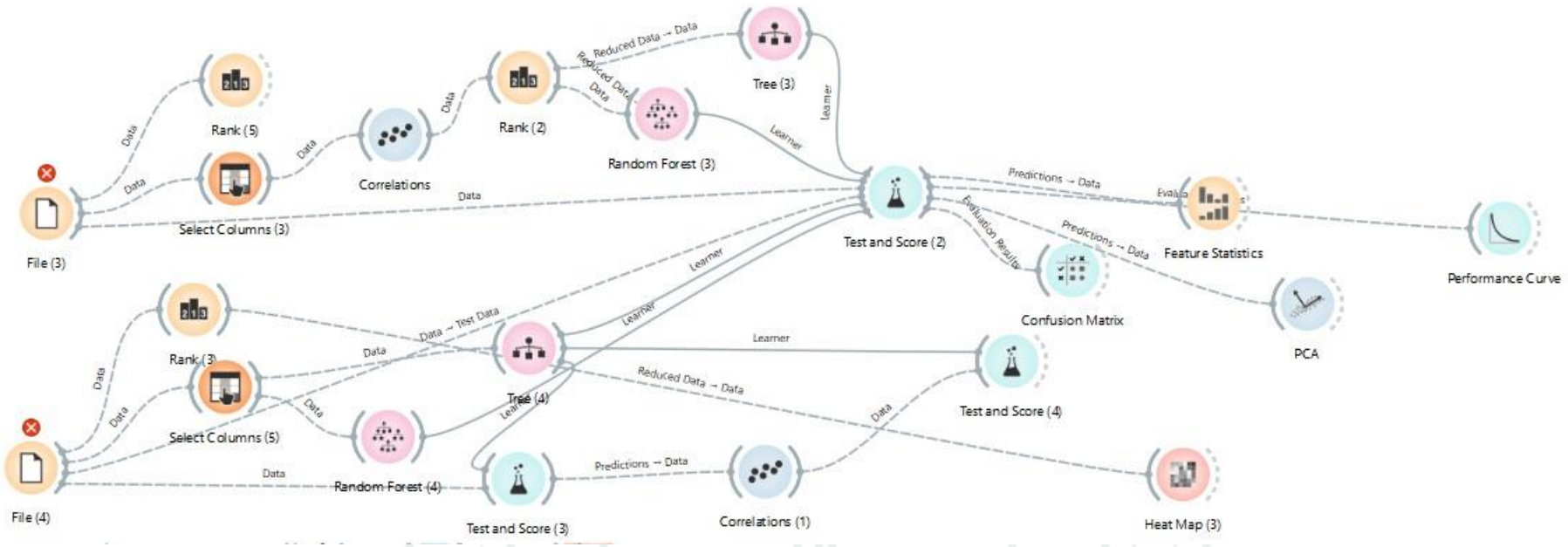
Figura 30 *Flujo de trabajo en Orange: verificación de calidad de datos (Parte 1)*





Fuente: Elaboración propia.

Figura 31 *Flujo de trabajo en Orange: verificación de calidad de datos (Parte 2)*



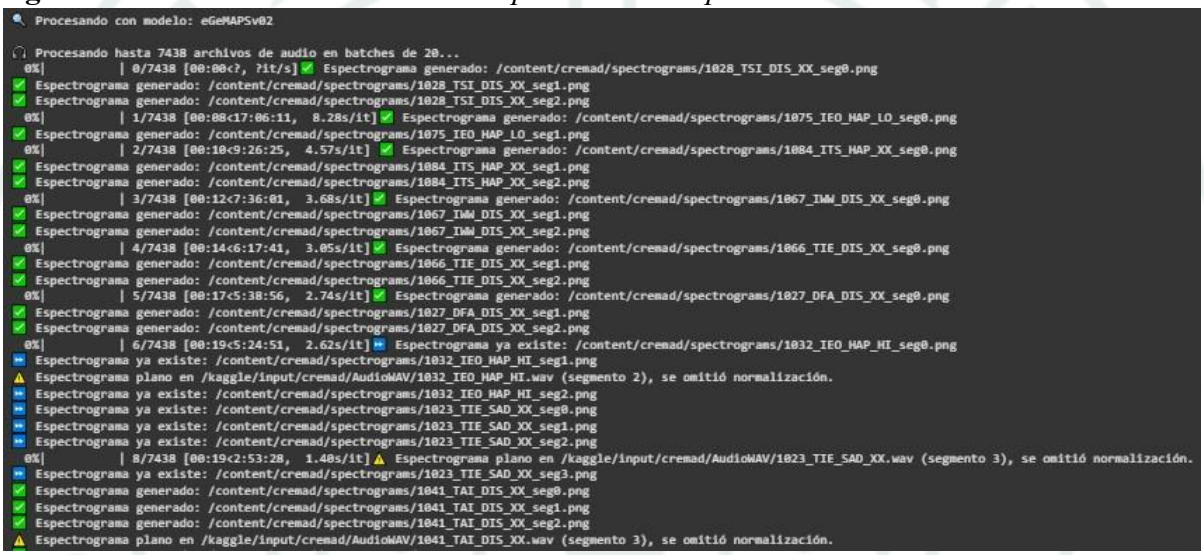
Fuente: Elaboración propia.

Anexo F: Analisis y diseño de la implementación

Se presenta el análisis y diseño de la implementación del sistema, desde la generación en baches, validación de archivos y procesamiento modular, hasta el análisis del árbol de decisión. Se incluyen ejemplos de características vocales clave, submodelos explicativos y los resultados completos de desempeño para CREMA-D y EMO-STIM. El contenido está organizado de forma secuencial y visualmente documentado.

Anexo F.1 Procesamiento en baches después de interrupción

Figura 32 *Procesamiento en baches después de interrupción*

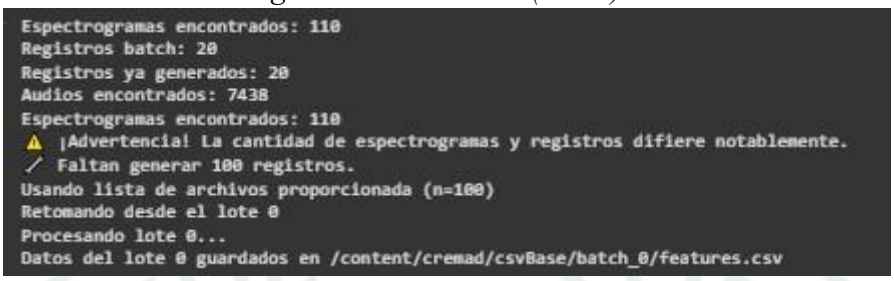


Fuente: Elaboración propia.

Detalla el manejo automático del sistema tras interrupciones, permitiendo reanudar procesos sin pérdida de datos.

Anexo F.2 Función de control de generación en lotes tch)

Figura 33 *Función de control de generación en lotes (batch)*

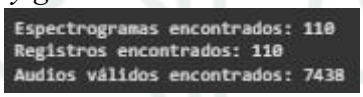


Fuente: Elaboración propia.

Describe cómo se controla la creación de lotes de entrenamiento/prueba, incluyendo los parámetros utilizados y su ejecución programada.

Anexo F.3 Función de validación y generación de archivos de entrenamiento

Figura 34 *Función de validación y generación de archivos de entrenamiento*

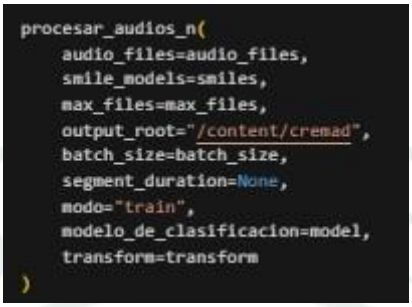


Fuente: Elaboración propia.

Resume en 3 variables si el contenido se genero correctamente entre imagenes, archivos csv y audios completos o en clips.

Anexo F.4 Especificaciones de llamada en entrenamiento

Figura 35 *Especificaciones de llamada en entrenamiento*



```
procesar_audios_n(
    audio_files=audio_files,
    smile_models=smiles,
    max_files=max_files,
    output_root="/content/cremad",
    batch_size=batch_size,
    segment_duration=None,
    modo="train",
    modelo_de_clasificacion=model,
    transform=transform
)
```

Fuente: Elaboración propia.

Muestra los parámetros técnicos y configuraciones necesarias al invocar el modelo durante la fase de entrenamiento.

Anexo F.5 Descompresión de archivos entrenados

Figura 36 *Descompresión de archivos entrenados*

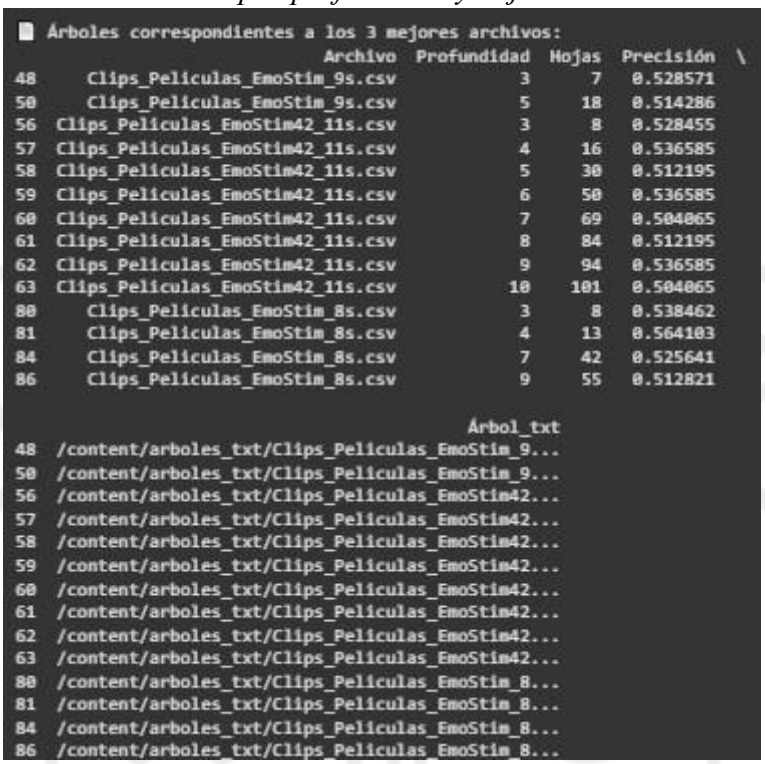
```
!zip -r spectrograms.zip /content/content/

  adding: content/content/ (stored 0%)
  adding: content/content/surrogate/ (stored 0%)
  adding: content/content/surrogate/surrogate_model1.pkl (deflated 75%)
  adding: content/content/surrogate/features_surrogate_features1.csv (deflated 65%)
  adding: content/content/surrogate/surrogate_model.pkl (deflated 76%)
  adding: content/content/surrogate/surrogate_tree.txt (deflated 89%)
  adding: content/content/surrogate/surrogate_tree1.txt (deflated 88%)
  adding: content/content/surrogate/.ipynb_checkpoints/ (stored 0%)
  adding: content/content/dummy/ (stored 0%)
  adding: content/content/dummy/eGeMAPSv02_feature_names.csv (deflated 80%)
  adding: content/content/dummy/dummy.wav (deflated 100%)
  adding: content/content/dummy/emobase_feature_names.csv (deflated 87%)
  adding: content/content/dummy/eGeMAPS_feature_names.csv (deflated 80%)
  adding: content/content/results/ (stored 0%)
  adding: content/content/results/resultados_lasso_embedding_vs_nuevo_dataset_Crema.csv (deflated 91%)
  adding: content/content/results/resultados_lasso_embedding_vs_nuevo_dataset_Emo.csv (deflated 91%)
  adding: content/content/results/resultados_lasso_embedding_vs_nuevo_dataset_EmoCompare.csv (deflated 75%)
  adding: content/content/results/resultados_lasso_embedding_vs_nuevo_dataset_CremaCompare.csv (deflated 78%)
  adding: content/content/embeddings/ (stored 0%)
  adding: content/content/embeddings/crema_d/ (stored 0%)
  adding: content/content/embeddings/crema_d/X_emb.npy (deflated 73%)
  adding: content/content/embeddings/crema_d/val_df.csv (deflated 62%)
  adding: content/content/embeddings/crema_d/.ipynb_checkpoints/ (stored 0%)
  adding: content/content/embeddings/crema_d/y_pred.npy (deflated 91%)
  adding: content/content/embeddings/emo_stim/ (stored 0%)
  adding: content/content/embeddings/emo_stim/X_emb.npy (deflated 62%)
  adding: content/content/embeddings/emo_stim/val_df.csv (deflated 62%)
  adding: content/content/embeddings/emo_stim/.ipynb_checkpoints/ (stored 0%)
  adding: content/content/embeddings/emo_stim/y_pred.npy (deflated 90%)
  adding: content/content/detalle/ (stored 0%)
  adding: content/content/detalle/detalle_export_text_por_depth(Emo).csv (deflated 96%)
  adding: content/content/detalle/detalle_export_text_por_depth(Cremad).csv (deflated 95%)
  adding: content/content/checkpoints/ (stored 0%)
  adding: content/content/checkpoints/modelo_final_99_20250501_213336.pth (deflated 7%)
```

Fuente: Elaboración propia.

Presenta el contenido estructurado del archivo comprimido que contiene modelos, árboles, metadatos y salidas del sistema.

Anexo F.6 Modelo auxiliar: tramos por profundidad y hojas del DT

Figura 37 *Modelo auxiliar: tramos por profundidad y hojas del DT*



Árboles correspondientes a los 3 mejores archivos:

	Archivo	Profundidad	Hojas	Precisión	\
48	Clips_Peliculas_EmoStim_9s.csv	3	7	0.528571	
50	Clips_Peliculas_EmoStim_9s.csv	5	18	0.514286	
56	Clips_Peliculas_EmoStim42_11s.csv	3	8	0.528455	
57	Clips_Peliculas_EmoStim42_11s.csv	4	16	0.536585	
58	Clips_Peliculas_EmoStim42_11s.csv	5	30	0.512195	
59	Clips_Peliculas_EmoStim42_11s.csv	6	50	0.536585	
60	Clips_Peliculas_EmoStim42_11s.csv	7	69	0.504065	
61	Clips_Peliculas_EmoStim42_11s.csv	8	84	0.512195	
62	Clips_Peliculas_EmoStim42_11s.csv	9	94	0.536585	
63	Clips_Peliculas_EmoStim42_11s.csv	10	101	0.504065	
80	Clips_Peliculas_EmoStim_8s.csv	3	8	0.538462	
81	Clips_Peliculas_EmoStim_8s.csv	4	13	0.564103	
84	Clips_Peliculas_EmoStim_8s.csv	7	42	0.525641	
86	Clips_Peliculas_EmoStim_8s.csv	9	55	0.512821	

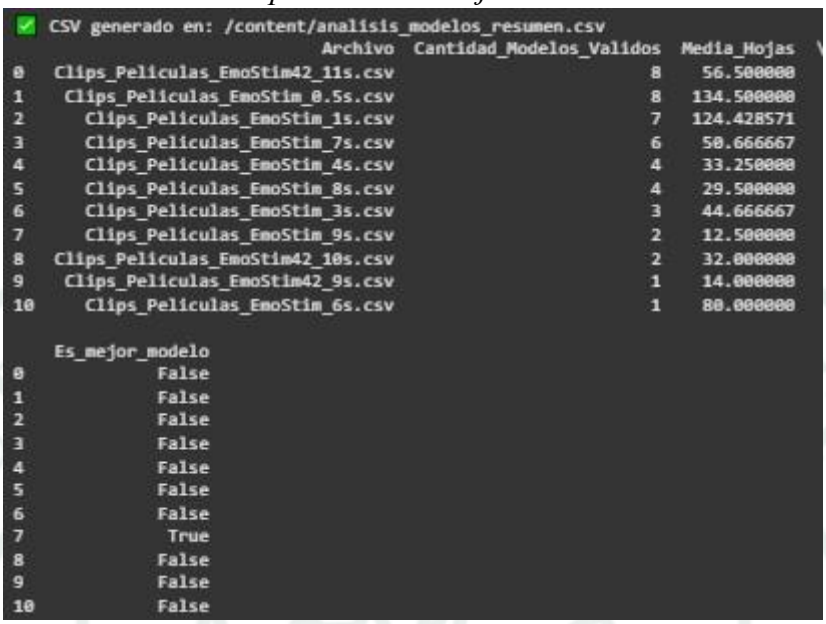
	Árbol_txt
48	/content/arboles_txt/Clips_Peliculas_EmoStim_9...
50	/content/arboles_txt/Clips_Peliculas_EmoStim_9...
56	/content/arboles_txt/Clips_Peliculas_EmoStim42...
57	/content/arboles_txt/Clips_Peliculas_EmoStim42...
58	/content/arboles_txt/Clips_Peliculas_EmoStim42...
59	/content/arboles_txt/Clips_Peliculas_EmoStim42...
60	/content/arboles_txt/Clips_Peliculas_EmoStim42...
61	/content/arboles_txt/Clips_Peliculas_EmoStim42...
62	/content/arboles_txt/Clips_Peliculas_EmoStim42...
63	/content/arboles_txt/Clips_Peliculas_EmoStim42...
80	/content/arboles_txt/Clips_Peliculas_EmoStim_8...
81	/content/arboles_txt/Clips_Peliculas_EmoStim_8...
84	/content/arboles_txt/Clips_Peliculas_EmoStim_8...
86	/content/arboles_txt/Clips_Peliculas_EmoStim_8...

Fuente: Elaboración propia.

Analiza la estructura del árbol de decisión en tramos de varios segundos, agrupando tramos según profundidad y número de hojas.

Anexo F.7 Modelo auxiliar: tramos por media de hojas del DT

Figura 38 *Modelo auxiliar: tramos por media de hojas del DT*



```
✅ CSV generado en: /content/analisis_modelos_resumen.csv
                        Archivo  Cantidad_Modelos_Validos  Media_Hojas  \
0   Clips_Peliculas_EmoStim42_11s.csv                        8    56.500000
1    Clips_Peliculas_EmoStim_0.5s.csv                        8   134.500000
2      Clips_Peliculas_EmoStim_1s.csv                        7   124.428571
3      Clips_Peliculas_EmoStim_7s.csv                        6    50.666667
4      Clips_Peliculas_EmoStim_4s.csv                        4    33.250000
5      Clips_Peliculas_EmoStim_8s.csv                        4    29.500000
6      Clips_Peliculas_EmoStim_3s.csv                        3    44.666667
7      Clips_Peliculas_EmoStim_9s.csv                        2    12.500000
8   Clips_Peliculas_EmoStim42_10s.csv                        2    32.000000
9    Clips_Peliculas_EmoStim42_9s.csv                        1    14.000000
10     Clips_Peliculas_EmoStim_6s.csv                        1    80.000000

    Es_mejor_modelo
0             False
1             False
2             False
3             False
4             False
5             False
6             False
7              True
8             False
9             False
10            False
```

Fuente: Elaboración propia.

Alternativa de agrupación basada en la media de hojas utilizadas por los embeddings, útil para estimar complejidad.

Anexo F.8 Ejemplo de características más utilizadas

Figura 39 *Ejemplo de características más utilizadas*

```
# Paso 4 (opcional): Normalizar por cantidad total de embeddings
df_conteo_vars['proporcion'] = df_conteo_vars['veces_usada'] / len(df_bien_explicado)

# Mostrar top 15 variables más frecuentes
print(df_conteo_vars.head(100))
```

	veces_usada	proporcion
F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope	5	1.0
F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope	5	1.0
F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope	5	1.0
F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope	5	1.0
mfcc1_sma3_amean	5	1.0
mfcc2_sma3_amean	5	1.0
mfcc3_sma3_amean	5	1.0
mfcc4_sma3_amean	5	1.0
F1bandwidth_sma3nz_amean	5	1.0
equivalentSoundLevel_dBp	5	1.0
F3bandwidth_sma3nz_amean	5	1.0
F3frequency_sma3nz_amean	4	0.8
F2frequency_sma3nz_amean	4	0.8
F2bandwidth_sma3nz_amean	4	0.8
F1frequency_sma3nz_amean	4	0.8
loudnessPeaksPerSec	4	0.8
logRelF0-H1-A3_sma3nz_amean	4	0.8
slopeV500-1500_sma3nz_stddevNorm	4	0.8
logRelF0-H1-H2_sma3nz_amean	3	0.6
logRelF0-H1-H2_sma3nz_stddevNorm	3	0.6
mfcc4_sma3_stddevNorm	3	0.6
mfcc3_sma3_stddevNorm	3	0.6
F1amplitudeLogRelF0_sma3nz_amean	3	0.6
alphaRatioUV_sma3nz_amean	3	0.6
hammarbergIndexUV_sma3nz_amean	3	0.6
mfcc2_sma3_stddevNorm	3	0.6
mfcc2V_sma3nz_amean	3	0.6
HNRdBACF_sma3nz_amean	3	0.6
mfcc4V_sma3nz_amean	2	0.4
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	2	0.4
mfcc1V_sma3nz_amean	2	0.4
F2amplitudeLogRelF0_sma3nz_amean	1	0.2
F3amplitudeLogRelF0_sma3nz_amean	1	0.2
loudness_sma3_stddevFallingSlope	1	0.2
mfcc3V_sma3nz_amean	1	0.2
F0semitoneFrom27.5Hz_sma3nz_percentile20.0	1	0.2
loudness_sma3_meanFallingSlope	1	0.2
mfcc1_sma3_stddevNorm	1	0.2
HNRdBACF_sma3nz_stddevNorm	1	0.2
hammarbergIndexV_sma3nz_amean	1	0.2
slopeV0-500_sma3nz_stddevNorm	1	0.2
mfcc3V_sma3nz_stddevNorm	1	0.2
mfcc4V_sma3nz_stddevNorm	1	0.2

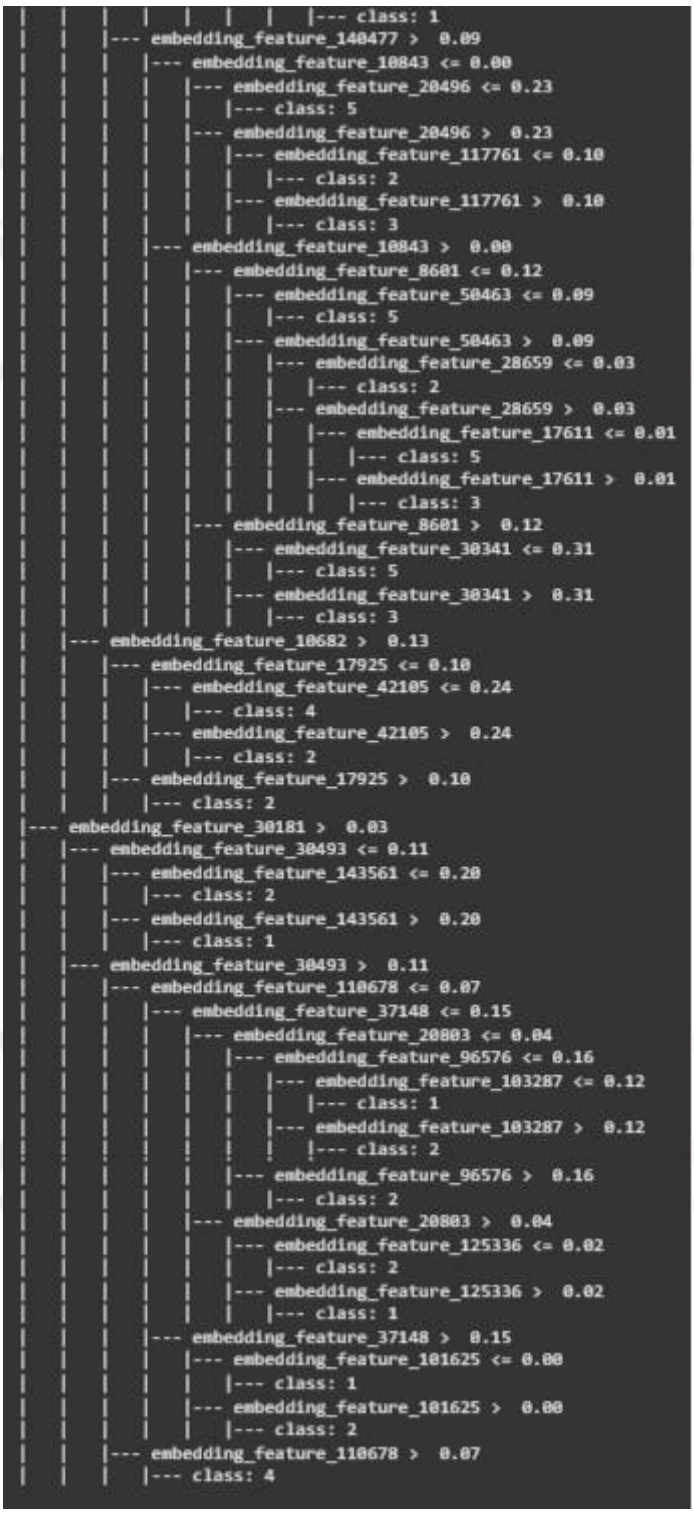
Fuente: Elaboración propia.

Muestra un conjunto representativo de las características vocales que más se repitieron en los modelos evaluados.

Anexo F.9 Ejemplo de vista del DT (Parte 2)

Figura 40 *Ejemplo de vista del*

DT (Parte 2)



Fuente: Elaboración propia.

Segunda parte del árbol de decisión general, donde se visualizan subramas adicionales que no aparecen en la figura principal.

Anexo F.10 Desempeño completo del modelo CREMA-D (None)

Figura 41 *Desempeño completo del modelo CREMA-D (None)*

	Umbral R2	Cantidad explicada	Porcentaje explicado (%)	Modelo
0	0.000	59	86.76	balanced
1	0.000	58	85.29	balanced
2	0.001	57	83.82	balanced
3	0.001	56	82.35	balanced
4	0.004	55	80.88	balanced
5	0.006	54	79.41	balanced
6	0.008	53	77.94	balanced
7	0.008	52	76.47	balanced
8	0.010	51	75.00	balanced
9	0.012	50	73.53	balanced
10	0.015	49	72.06	balanced
11	0.019	48	70.59	balanced
12	0.022	47	69.12	balanced
13	0.023	46	67.65	balanced
14	0.034	45	66.18	balanced
15	0.041	44	64.71	balanced
16	0.083	43	63.24	balanced
17	0.087	42	61.76	balanced
18	0.087	41	60.29	balanced
19	0.093	40	58.82	balanced
20	0.097	39	57.35	balanced
21	0.102	38	55.88	balanced
22	0.108	37	54.41	balanced
23	0.113	36	52.94	balanced
24	0.116	35	51.47	balanced
25	0.120	34	50.00	balanced
26	0.120	33	48.53	balanced
27	0.123	32	47.06	balanced
28	0.126	31	45.59	balanced
29	0.127	30	44.12	balanced
30	0.128	29	42.65	balanced
31	0.130	28	41.18	balanced
32	0.144	27	39.71	balanced
33	0.151	26	38.24	balanced
34	0.182	25	36.76	balanced
35	0.184	24	35.29	balanced
36	0.185	23	33.82	balanced
37	0.185	22	32.35	balanced
38	0.188	21	30.88	balanced
39	0.202	20	29.41	balanced
40	0.203	19	27.94	balanced
41	0.206	18	26.47	balanced
42	0.214	17	25.00	balanced
43	0.214	16	23.53	balanced
44	0.222	15	22.06	balanced
45	0.225	14	20.59	balanced
46	0.230	13	19.12	balanced
47	0.235	12	17.65	balanced
48	0.256	11	16.18	balanced
49	0.282	10	14.71	balanced
50	0.292	9	13.24	balanced
51	0.297	8	11.76	balanced
52	0.310	7	10.29	balanced
53	0.316	6	8.82	balanced
54	0.340	5	7.35	balanced
55	0.346	4	5.88	balanced
56	0.377	3	4.41	balanced
57	0.380	2	2.94	balanced
58	0.388	1	1.47	balanced
59	0.394	0	0.00	balanced

Fuente: Elaboración propia.

Resultados del modelo CREMA-D bajo balanceo, en distintos tramos de R^2 , mostrando el porcentaje de embeddings efectivos por nivel.

Anexo F.11 Desempeño completo del modelo EMO-STIM (Balanced)

Figura 42 *Desempeño completo del modelo EMO-STIM (Balanced)*

	Umbral R2	Cantidad explicada	Porcentaje explicado (%)	Modelo
0	0.000	41	97.62	None
1	0.001	40	95.24	None
2	0.002	39	92.86	None
3	0.013	38	90.48	None
4	0.015	37	88.10	None
5	0.019	36	85.71	None
6	0.022	35	83.33	None
7	0.032	34	80.95	None
8	0.039	33	78.57	None
9	0.042	32	76.19	None
10	0.047	31	73.81	None
11	0.057	30	71.43	None
12	0.062	29	69.05	None
13	0.062	28	66.67	None
14	0.064	27	64.29	None
15	0.066	26	61.90	None
16	0.068	25	59.52	None
17	0.071	24	57.14	None
18	0.082	23	54.76	None
19	0.082	22	52.38	None
20	0.089	21	50.00	None
21	0.093	20	47.62	None
22	0.096	19	45.24	None
23	0.120	18	42.86	None
24	0.120	17	40.48	None
25	0.136	16	38.10	None
26	0.181	15	35.71	None
27	0.187	14	33.33	None
28	0.202	13	30.95	None
29	0.216	12	28.57	None
30	0.254	11	26.19	None
31	0.286	10	23.81	None
32	0.417	9	21.43	None
33	0.420	8	19.05	None
34	0.471	7	16.67	None
35	0.502	6	14.29	None
36	0.503	5	11.90	None
37	0.509	4	9.52	None
38	0.580	3	7.14	None
39	0.635	2	4.76	None
40	0.651	1	2.38	None
41	0.764	0	0.00	None

Fuente: Elaboración propia.

Desempeño general del modelo EMO-STIM sin balanceo, también organizado por tramos y vinculado a sus características vocales.

Anexo G: Notas exploratorias sobre percepción emocional en medios(No comprobada)

Durante pruebas informales con audios de interés personal se observó un patrón llamativo: algunas escenas particularmente enganchantes o cargadas emocionalmente se comportaban distinto a los de la batería de películas, en especial con la tristeza y disgusto, mostraban ya sea niveles de activación o carácter contrarios a lo esperado; la intensidad se mantenía, variaba mínimamente, transicionaba velozmente o era completamente errática; percibiendo algo de afinidad con narrativas como monólogos, múltiples hablantes o en ambientes considerados más naturales, en algunos casos las transiciones eran tan rápidas que podrían ser inexistentes.

El episodio ’Free Churro’ de BoJack Horseman, nominado al Primetime Emmy Award for Outstanding Animated Program, usó una técnica común en el entretenimiento conocida como ’episodio botella’, que se caracteriza por personajes que conversan o hacen monólogos en escenarios simples. Las críticas la destacan por su gran carga emocional y por la actuación vocal. La escena del prólogo con el monólogo del padre abusivo, desde el análisis con el modelo auxiliar (Figura 3) se detectó inconsistencias como que una emocion que fue casi absoluta y el carácter de comportamiento fue contrario al esperado por la afectividad, y si bien no es evidente esta consistencia al escucharla, podemos estar identificando los cambios percibidos mas que los expresados. Algo similar se observó en episodios de Hunter x Hunter, con otro monólogo contradictorio, por una combinación de disgusto pasivo que los críticos perciben como tristeza. En transmisiones de videojuegos competitivos amicales el tono alegre, donde solía aparecer el enojo pasivo, los espectadores reportaron mayor interés por largos tramos. Además, analizando medios más experimentales, que podrían contener un ambiente más natural, ya sea por estar basados en entrevistas reales como en el podcast de Midnight Gospel en un contexto sensible y personal, que involucran exponerse al público, o con streamers que suelen subir un contenido con la menor cantidad de edición de forma diaria por más de una década, su comportamiento emocional bastante complejo puede sugerir de forma involuntaria una relación entre expresiones emocionales inusuales y el atractivo de la escena.

De forma interesante, los embeddings con mayor rendimiento se comportan con afectividad, los elementos que más se repiten en las clasificaciones son de felicidad y tristeza, el comportamiento de la neutralidad es uno de los más útiles, incluso frente a grupos bien definidos, esto plantea la duda de si algunos de estos embeddings, más que contraponer emociones, miden su grado de activación o presencia. Sutilmente, este aparenta clasificar las emociones expresadas sin entrenamiento previo, pero este comportamiento puede no estar tan lejos de ser comprendido, ya sea por ayuda del contexto multidisciplinario o por la fusión de elementos.

Aunque no formó parte del análisis oficial, esta observación podría servir para formular hipótesis en estudios futuros o complementarios sobre la percepción en medios.