ESCUELA DE INGENIERÍA DE FUENLABRADA

Grado en Título de Grado

Trabajo Fin de Grado

Clasificación de emociones en grabaciones de voz mediante técnicas de Machine Learning

por

Willians Paul Pico Quiroz

Tutora: Sara García De Villa
Co-tutora: Elena Aparicio Esteve

Año académico 2023/2024

Dedicado a Modesta y Néstor

Agradecimientos

Me gustaría empezar agradeciendo a mis tutoras Sara García de Villa y Elena Aparicio Esteve, por su gran orientación y sobretodo disponibilidad a lo largo de estos meses al tener que compaginar este Trabajo de Fin de Grado con mi trabajo laboral.

Y también quiero agradecer a mis padres, a mis abuelos y a mi hermano por su constante apoyo y confianza en mi, ya que sin ellos me hubiera rendido hace mucho y no lo habría conseguido, gracias por todo.

Abstract

This TFG focuses on the classification of voice recordings using Machine Learning techniques. Using the supervised learning algorithm, we aim to achieve an accurate and effective model through data labeling. Some alternatives that can serve as support material are presented, such as *GeMAPS* and *GSU Praat Tools* for feature extraction and *Oxford Vocal Sounds* tools for database selection. As mentioned, supervised learning is chosen as a model for this work, where some of the proposed methods are RF, LDA and k-NN. The database is given from the training of two speech corpora formed by positive, negative and neutral words and pseudo-words, which are classified by six types of emotions with values according to thirteen acoustic features or parameters. Before proceeding to the application of the methods, the database will undergoes database processing. And after this, six types of experiments are performed with the intention of seeing if modifying certain aspects of the database will lead to a higher *Accuracy*. Given the results, for the experiment in which all the characteristics are maintained and the number of emotions is reduced to 3, the LDA method is the most feasible since it presents a higher *Accuracy*, with a value of 66 %.

Resumen

Este TFG se centra en la clasificación de grabaciones de voz mediante técnicas de Machine Learning. Usando el algoritmo de aprendizaje supervisado, se busca conseguir un modelo preciso y efectivo a través del etiquetado de datos. Se presentan algunas alternativas que pueden servir como material de apoyo, como por ejemplo en lo referido a la extracción de características destacan *GeMAPS* y *GSU Praat Tools* y en lo referido a selección de una base de datos, se presenta como un apoyo las herramientas de *Oxford Vocal Sounds*. Como se ha mencionado, para este trabajo se elige el aprendizaje supervisado como modelo, donde algunos de los métodos propuestos son el RF, el LDA y el k-NN. La base de datos viene dada de la formación de dos corpus de voz formada por palabras y pseudo-palabras positivas, negativas y neutrales, la cuales están clasificadas por seis tipos de emociones con valores según trece características o parámetros acústicos. Antes de proceder a la aplicación de los métodos, la base de datos pasa por un procesado de base de datos. Y, tras esto, se realizan seis tipos de experimentos con la intención de comprobar si modificando ciertos aspectos de la base de datos se consigue una mayor tasa de acierto o *Accuracy*. Dados los resultados, para el experimento en el que se mantienen todas las características y se reduce el número de emociones a 3, es el método LDA el más factible ya que presenta una mayor *Accuracy*, con un valor del 66 %.

Índice general

Índice de Figuras

Índice de Tablas

Abreviaciones

DBSCAN *Density-Based Spatial Clustering of Applications with Noise*. 24

DL *Deep Learning*. 59

ERCRVSTPP *Emotion Recognition and Confidence Ratings predicted by Vocal Stimulus Type and Prosodic Parameters*. 32

k-NN k-Nearest Neighbors. 3, 4, 20, 22–24, 36, 42, 43, 45, 47, 49–54, 57, II, III, IX

LDA Análisis Discriminante Lineal. 3, 4, 20–22, 36, 42, 43, 45–58, II, III, IX

MFCCs *Coeficientes Cepstrales de Frecuencia Mel*. 17

ML *Machine Learning*. 1, 4, 9, 15–18, 20, 57, 59, 60, 62

ODS Objetivos de Desarrollo Sostenible. 60, 61, XII

RF Random Forest. 3, 4, 7, 20, 21, 36, 42–47, 49–57, II, III, IX

TFG Trabajo de Fin de Grado. 1, 3, 4, 8, 15, 18, 26, 32, 34, 35, 39, 55–62, II, III, X, XII

ZCR *Zero-Crossing Rate*. 17

Capítulo 1

Introducción

Este capítulo aborda la introducción a la clasificación de emociones en grabaciones de voz mediante técnicas de *Machine Learning* (ML). La primera Sección 1.1 pone en contexto como se deriva la clasificación de emociones y la motivación de detectarlas y clasificarlas. La siguiente Sección 1.2, explica los principales objetivos que se buscan en este Trabajo de Fin de Grado (TFG). Y por último, la Sección 1.3 indica cómo está estructurada la memoria, asimismo, explicando brevemente cada capítulo.

1.1. Motivación

En la actualidad, la detección y clasificación de emociones en grabaciones de voz mediante técnicas de ML se ha convertido en una herramienta esencial en diversos ámbitos, como la atención al cliente, la salud mental, etc. Por ejemplo, la capacidad de identificar y clasificar emociones permite a las empresas mejorar la calidad del servicio al cliente al comprender mejor sus necesidades y emociones en tiempo real, luego en el ámbito de la salud, facilita el diagnóstico y seguimiento de trastornos emocionales. Este enfoque tiene el potencial para abrir nuevas posibilidades para comprender y mejorar las relaciones humanas a través del análisis de la voz.

La capacidad de comprender correctamente y responder adecuadamente a las emociones de otras personas desempeña un papel importante en las interacciones sociales cotidianas [1], [2]. Esto se puede observar en la Figura 1.1 utilizándose siete características para diferenciar distintos tipos de estado afectivo, donde las tres primeras columnas *Intensity*, *Duration* y el grado de *synchronization* reflejan el grado en que los sistemas fisiológicos del cuerpo se alteran y trabajan juntos de manera coordinada en respuesta a una situación específica. La focalización en el acontecimiento se refiere a la probabilidad de que el estado afectivo sea desencadenado por un objeto, acontecimiento o situación específicos. La elicitación apreciativa se refiere al grado en que el tipo de reacción se debe

a la evaluación subjetiva del significado del acontecimiento para una persona, dados los motivos y objetivos momentáneos o los valores a más largo plazo. Por último, la rapidez del cambio se refiere a la rapidez con la que puede cambiar el estado (inicio, desplazamiento y cambio en la calidad), mientras que el impacto en la conducta se refiere a la fuerza del impacto del estado afectivo respectivo en las respuestas fisiológicas, la expresión motora y las tendencias a la acción. Por otro lado, las filas indican tres grandes clases de estados afectivos que sería emociones y estrés, estados de ánimo y posturas interpersonales y referencias/actitudes y disposiciones afectivas. Las emociones y el estrés son reacciones bastante breves pero intensas que producen cambios en el comportamiento. Los estados de ánimo y las posturas interpersonales rara vez son generados por acontecimientos. Por último, las preferencias/actitudes son evaluaciones afectivas a largo plazo de objetos o personas que tienen una intensidad baja y relativamente poco impacto en el comportamiento, porque los factores situacionales suelen ser determinantes más fuertes del comportamiento. Cabe destacar también, que los puntos indican el valor de la característica en función del estado afectivo, habiendo tres niveles: bajo, medio y alto [2].

En la comunicación verbal, por ejemplo, los seres humanos no se limitan a considerar lo que dicen sus interlocutores (es decir, el significado semántico), sino también cómo transmiten la información hablada (por ejemplo, el tono alto/bajo de su voz). Un término que engloba todas estas cualidades vocales del habla es prosodia (es decir, tono de voz). Se ha demostrado que la prosodia puede apoyar las interpretaciones correctas de los enunciados independientemente de la comprensión lingüística [3], [4], con estudios que informan que las tasas de reconocimiento de las emociones son significativamente más altas que el azar [5–11].

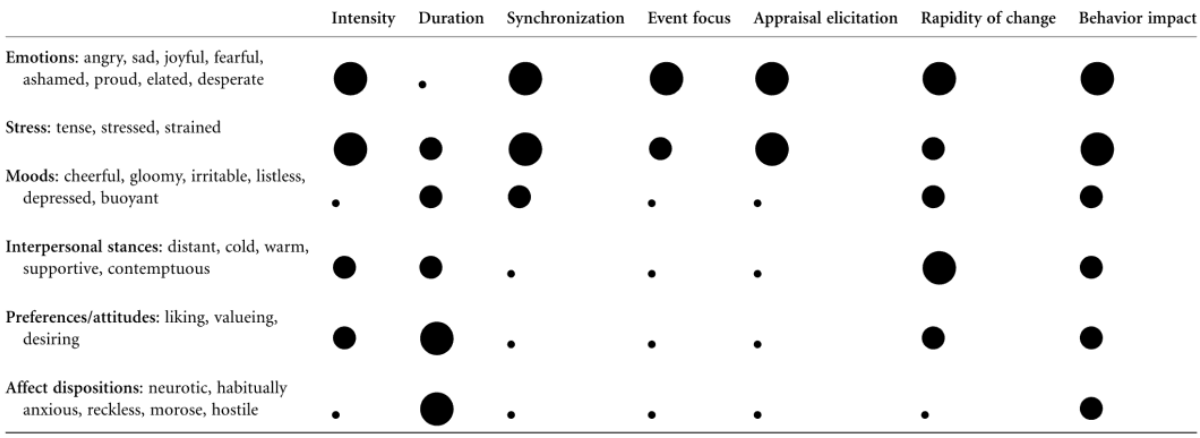


Figura 1.1: Delimitación de las características de diseño de los distintos estados afectivos [2].

Además, se ha argumentado que la metacognición, la capacidad de supervisar activamente y reflexionar sobre el propio rendimiento, influye en los juicios de precisión en tareas de reconocimiento de emociones [12–14]. Para comprender mejor los mecanismos

subyacentes al reconocimiento de emociones a partir de la voz, se han examinado cómo los diferentes tipos de estímulos vocales y sus atributos acústicos influyen en el reconocimiento de emociones y las calificaciones de confianza de los oyentes. En su empeño por evaluar el reconocimiento de emociones a partir de la prosodia, los investigadores crearon una amplia variedad de materiales de estímulo. Sin embargo, como las emociones no se expresan en el mismo grado en cada palabra de una frase, se ha sugerido que estos estímulos de larga duración podrían contener una mayor variación y ruido en la señal. Asimismo, la extracción de parámetros acústicos el tono, el volumen, el tempo y la calidad (voz acentuada/respirada) son los rasgos paralingüísticos más relevantes que emplean los hablantes al expresar emociones [15]. Una serie de estadísticas sobre estos rasgos paralingüísticos ha revelado que los parámetros relacionados con el tono o frecuencia fundamental (F0) (p. ej., mínimo, máximo, media, fluctuación), la energía/amplitud (p. ej., volumen, brillo), temporales (p. ej., duración) y parámetros de calidad (p. ej., relación armónicos-ruido [HNR]) se encuentran entre los candidatos más importantes para los correlatos prosódicos de la emoción en el habla [16], [17].

Por lo que en este contexto, este TFG tiene como propósito el empleo de diversos métodos de aprendizaje supervisado: (Random Forest (RF), Análisis Discriminante Lineal (LDA), k-Nearest Neighbors (k-NN)) para llevar a cabo la clasificación de emociones (*fear*, *disgust*, *happy*, *sad*, *neutral* y *anger*) y utilización de características (*Duration*, *PeakTime*, *Amp(dB)*, *PeakAmp*, *MinF0*, *MaxF0*, *MeanF0*, *StDevF0*, *Jitter*, *Shimmer*, *MaxHNR*, *MeanHNR* y *StDevHNR*) en grabaciones de voz. Para esto se emplea una base de datos que contiene señales de voz en forma de palabras y pseudo-palabras en entornos controlados [18]. Estas señales están modeladas con los corpus de voz *Anna* [19] y *Magdeburg Prosody Corpus* [20]. Se realiza previamente un procesado de la base de datos: eliminación de los datos NaN (Not a Number), importación de las bibliotecas necesarias para cada método manteniendo los parámetros por defecto, la estandarización de los datos, la obtención de las características más importantes y la división de los datos en conjuntos de entrenamiento y prueba. Tras esto, finalmente se llevan a cabo seis experimentos: 1) se mantienen todas las características y emociones, 2) mantener todas las características y filtrar por 3 emociones, 3) omisión de la característica *Duration* de las demás características y filtrar por las 3 emociones, 4) obtener las características más importantes con valores entre 0,1 y menores a 0,15, 5) obtener las características más importantes con valores entre 0,15 y menores a 0,2 y 6) obtener las características más importantes con valores iguales y mayores a 0,2. Los resultados obtenidos en estos experimentos se analizan utilizando métricas de clasificación y de matriz de confusión obteniendo como mejor resultado el método LDA, con una *Accuracy* de $0,666$.

1.2. Objetivos

Este TFG se basa en el análisis para la clasificación de emociones en grabaciones de voz mediante técnicas de ML. Por lo tanto, los principales objetivos que persigue este TFG son los siguientes:

- Clasificar emociones mediante métodos de aprendizaje supervisado a partir de grabaciones de voz provenientes de una base de datos. Los métodos empleados para alcanzar este objetivo son: RF, LDA y k-NN.
- Preprocesar la base de datos y aplicarle cada método del análisis supervisado a cada experimento.
- Realizar un estudio comparativo entre los 6 experimentos realizados y sus métricas obtenidas para elegir el método adecuado para la propuesta de este TFG.

1.3. Estructura de la Memoria

La memoria de este TFG se ha estructurado en seis capítulos divididos en dos grandes bloques. El primer bloque consta de los Capítulos 1, 2 y 3 y proporciona el contexto y las herramientas utilizadas para el desarrollo de este trabajo. El segundo bloque está compuesto por los Capítulos 4, 5 y 6 y recoge la metodología seguida, resultados obtenidos y conclusiones de este TFG, respectivamente. A continuación se presenta un resumen de cada uno de los capítulos que componen esta memoria:

- El Capítulo 1 introduce la motivación y los objetivos de este TFG, enfocado en la clasificación de emociones en grabaciones de voz mediante técnicas de ML, así como un breve resumen de los puntos tratados en la memoria.
- El Capítulo 2 realiza una revisión de la literatura científica y tecnológica relacionada con el análisis y la clasificación de emociones.
- El Capítulo 3 incluye el marco teórico del algoritmo basado en ML que se usará, el cual es el aprendizaje supervisado, destacando sus principios, aplicaciones y métodos más utilizados. También incluye una breve explicación del aprendizaje no supervisado.
- El Capítulo 4 presenta tanto la base de datos utilizada como la descripción de las características y emociones utilizadas, el procesado de la base de datos. Por último, se definen cada uno de los experimentos realizados y las métricas empleadas para su evaluación.

- El Capítulo 5 recoge los resultados obtenidos de cada uno de los experimentos descritos en el Capítulo 4, así como un análisis y discusión de estos.
- El Capítulo 6 incluye las conclusiones obtenidas, el impacto del estudio y las principales limitaciones así como las futuras líneas de investigación.

Capítulo 2

Estado del arte

La clasificación de sonidos según emociones es una disciplina especializada en el campo del procesamiento de señales acústicas que busca identificar y categorizar las expresiones emocionales contenidas en el audio. Este campo de estudio se centra en la detección automática de estados emocionales, como la alegría, la tristeza, el miedo o la ira, a partir de las características acústicas presentes en la voz. A través del análisis de las señales acústicas, los clasificadores de sonidos emocionales emplean algoritmos de aprendizaje automático para distinguir entre diferentes estados emocionales, donde esta clasificación es útil para comprender y responder adecuadamente a la información emocional transmitida a través del audio.
En esta sección, se explican diferentes métodos en los que se basa un clasificador de emociones que emplea sonidos de voz. En primer lugar, se detallan los sistemas de extracción de características, en este caso GeMAPS y GSU Praat Tools, y por último, se presentan las bases de datos, como la Oxford Vocal Sounds, para la investigación en el procesamiento de señales de voz y reconocimiento del habla. Además, también mencionarán distintos trabajos previos que se utilizan para el reconocimiento de emociones, donde expliquen las diferentes técnicas que usan para llegar al mismo objetivo.

2.1. Trabajos previos de reconocimiento de emociones

Los siguientes trabajos han ayudado a sentar las bases para el desarrollo de algoritmos y modelos que pueden analizar y clasificar las expresiones emocionales en el habla con una precisión cada vez mayor, los cuales se van a describir de forma técnica.

En [21] se describe un estudio que aborda las dificultades que enfrentan los niños con Trastorno del Espectro Autista (TEA) para detectar y expresar emociones, lo que resulta en problemas de comunicación y funcionamiento social. Se propone un juego educativo

hablado, utilizando técnicas de Aprendizaje Automático, para ayudar a estos niños a identificar y expresar emociones. El juego se enfoca en cuatro estados emocionales (felicidad, tristeza, ira y neutralidad) y consta de dos niveles de dificultad creciente. El sistema integra un reconocedor multilingüe (inglés y alemán) de emociones basado en el tono de voz. Se realiza un procesamiento para el reconocimiento de emociones mediante el análisis del tono de voz, incluyendo la extracción de características (animaciones silenciosas, caras de personas reales y el tono de voz) y el uso de un clasificador de RF. Se menciona el uso de conjuntos de datos en inglés y alemán para entrenar el modelo. Se proporciona una tasa promedio de precisión de clasificación del 72 % para el clasificador. En las siguientes Figuras 2.1 y 2.2 se puede observar el diagrama de flujo que sigue la aplicación y los procesos de entrenamiento y predicción que aplica:

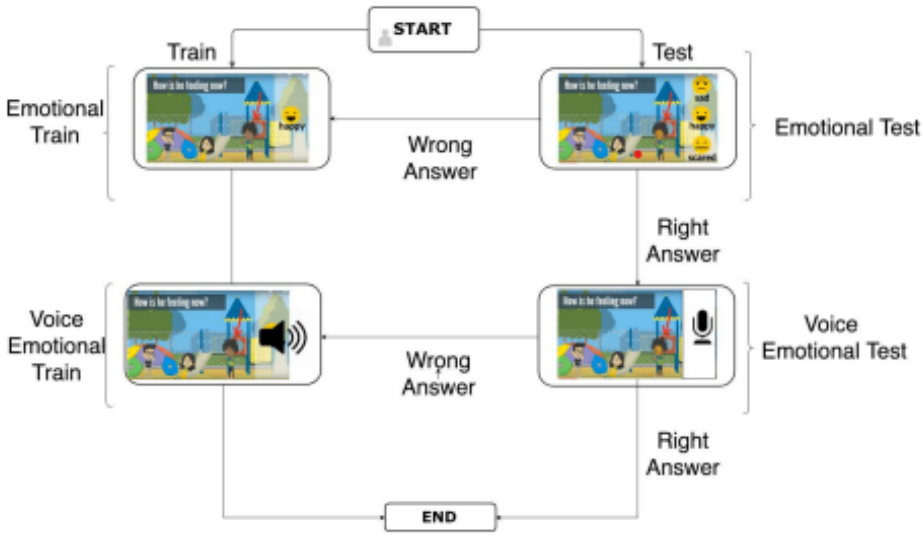


Figura 2.1: Diagrama de flujo de la aplicación. El juego se enfoca en cuatro estados emocionales, dos de train y dos de test. Consta de dos niveles de dificultad creciente. El sistema integra un reconocedor multilingüe (inglés y alemán) de emociones basado en el tono de voz [21].

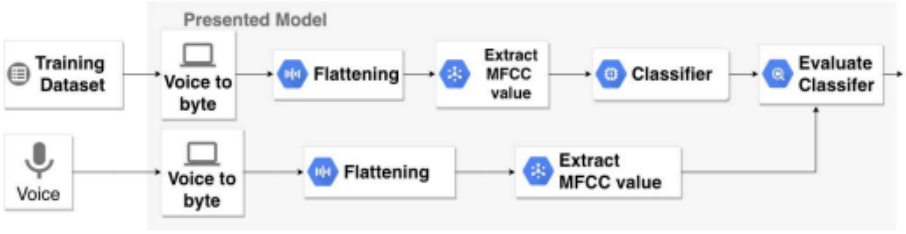


Figura 2.2: Procesos de entrenamiento y predicción. Se realiza un procesamiento para el reconocimiento de emociones mediante el análisis del tono de voz, incluyendo la extracción de características y el uso de un clasificador [21].

Por otro lado está el Reconocimiento de Emociones Basado en Transformada de Onda Compleja de Doble Árbol y Aprendizaje Automático", que es un enfoque que combina técnicas de procesamiento de señales y aprendizaje automático para identificar y clasificar emociones en señales de voz [22]. Este método utiliza la transformada de onda compleja de doble árbol para extraer características relevantes de las señales acústicas. Esta es una técnica de descomposición de señales que permite analizar tanto la información en el dominio del tiempo como en el dominio de la frecuencia con alta resolución. Al aplicar esta transformada a las señales de voz, se pueden obtener características que capturan detalles finos en diferentes escalas de tiempo y frecuencia, lo que resulta en una representación más rica y descriptiva de la señal. Tras ello, estas características se utilizan como entrada para algoritmos de aprendizaje automático, como clasificadores de vectores de soporte o redes neuronales, que son entrenados para reconocer y clasificar emociones específicas, como alegría, tristeza o enfado, basándose en patrones identificados en las características extraídas. Este enfoque ha demostrado ser prometedor en la identificación precisa de emociones en señales de voz, lo que lo hace relevante para aplicaciones en campos como la interacción humano-computadora, la psicología computacional y la salud mental.

También cabe a destacar que en el articulo "Decoding Emotions From EEG Responses Elicited by Videos Using Machine Learning Techniques on Two Datasets- [23], se hace referencia a un estudio que se centra en la aplicación de técnicas de aprendizaje automático para descifrar las emociones humanas a partir de respuestas de EEG (electroencefalografía) provocadas por la visualización de videos. Aunque no se utiliza voz, se exploran otras señalees para la clasificación de emociones con los algortimos que se analizan en este TFG. La investigación se llevó a cabo utilizando dos conjuntos de datos diferentes: SEED y DEAP. Se menciona que SVM tiene el mejor rendimiento para DEAP y MLP para SEED. Estos clasificadores, junto con su configuración, se pueden observar en la Figura 2.3. En este estudio, los investigadores recopilaron datos de EEG de participantes que observaron una serie de videos diseñados para evocar respuestas emocionales específicas. Estos videos podrían provocar emociones como felicidad, tristeza, miedo, entre otras. Luego, utilizaron técnicas de aprendizaje automático para analizar los patrones en las señales de EEG y determinar qué emociones estaban experimentando los participantes en función de estas respuestas cerebrales. Los resultados del estudio demostraron la viabilidad de utilizar el EEG como una herramienta para detectar y clasificar las emociones humanas con una precisión significativa. Además, al utilizar dos conjuntos de datos diferentes, los investigadores pudieron validar y comparar la efectividad de sus enfoques en diferentes contextos y poblaciones. Este estudio tiene importantes implicaciones en diversos campos, como la psicología, la neurociencia y la informática. Por ejemplo, podría ayudar a mejorar la comprensión de cómo procesamos y experimentamos las emociones, así como a desarrollar tecnologías de asistencia emocional basadas en el EEG. En resumen, "Decoding Emotions From EEG Responses Elicited by Videos Using Machine Learning Techniques

on Two Datasets.ºfrece una contribución significativa al campo emergente de la detección de emociones utilizando datos fisiológicos y técnicas de aprendizaje automático.

Classifier	Parameter details
KNN	K = {3,5}
SVM	Kernel: RBF Decision function: One-vs-One
MLP-v1	2 hidden layers with 100 and 50 nodes dropout rate: 0.1
MLP-v2	2 hidden layers with 500 and 300 nodes dropout rate: 0.2
MLP-v3	3 hidden layers with 2000, 1000 and 500 nodes dropout rate: 0.1

Figura 2.3: Detalles sobre los clasificadores utilizados, donde se puede observar los parámetros que se aplican a cada clasificador [23].

En el siguiente artículo fusiona la educación y la tecnología de aprendizaje automático [24]. En este campo, los investigadores buscan utilizar algoritmos de ML para comprender mejor las expresiones faciales de los estudiantes durante el proceso de aprendizaje. Esta investigación se centra en cómo las expresiones faciales pueden ser indicadores de emociones, niveles de atención y comprensión en el aula. Al emplear técnicas avanzadas de ML, como redes neuronales convolucionales, se pueden analizar grandes volúmenes de datos de expresiones faciales para identificar patrones y correlaciones significativas (este proceso lo podemos observar en la Figura 2.4. Esto podría conducir a la creación de sistemas de retroalimentación en tiempo real para los educadores, permitiéndoles adaptar sus métodos de enseñanza según las necesidades individuales de los estudiantes. En última instancia, este enfoque podría mejorar la eficacia del proceso educativo al personalizar la experiencia de aprendizaje y fomentar un entorno más receptivo y colaborativo en el aula.

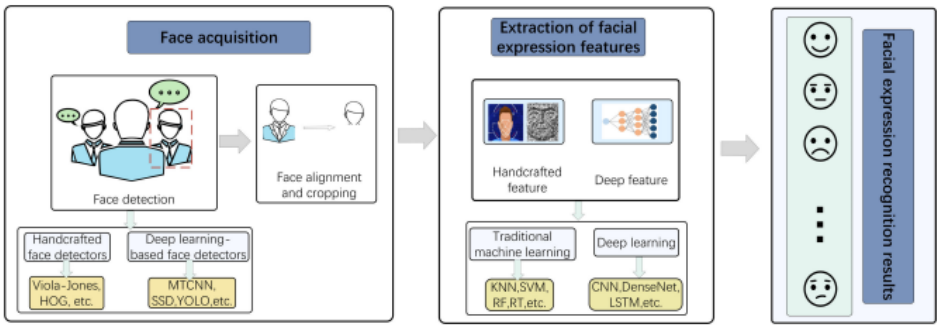


Figura 2.4: Marco estándar para el reconocimiento de la expresión facial basado en el aprendizaje automático [24].

Por último, en [25], se menciona un estudio que se centra en la aplicación de modelos de aprendizaje automático para reconocer el estrés y la ansiedad a partir de señales de

actividad cardíaca. Asimismo, se enfoca también en evaluar la robustez de estos modelos en diferentes condiciones y entornos, considerando la variabilidad natural de las señales fisiológicas y los posibles artefactos o interferencias en los datos. Los investigadores analizan la efectividad de los modelos de aprendizaje automático en la detección de estrés y ansiedad bajo diversas circunstancias, como cambios en la postura corporal, actividad física, o variaciones en las condiciones ambientales. Además, el estudio también examina la influencia de diferentes características extraídas de las señales de actividad cardíaca en el rendimiento de los modelos de aprendizaje automático. Esto incluye parámetros como la variabilidad de la frecuencia cardíaca, la forma de onda de los latidos cardíacos, y otros indicadores de la actividad fisiológica asociada con el estrés y la ansiedad. La explicación del proceso se puede observar en la Figura 2.5. El conjunto de datos de señales de emoción continuamente anotadas (CASE) se recopiló en un experimento con 30 participantes, 15 hombres y 15 mujeres, con ocho modalidades de sensores que captaban señales fisiológicas. Se utilizan señales ECG y BVP capturadas mediante un convertidor analógico-digital de 16 bits. El conjunto de datos incluye anotaciones recopiladas de un joystick que los participantes usan para autoinformar su estado emocional durante el experimento. Este joystick proporciona lecturas de excitación y valencia a 20 Hz y se anota con símbolos del maniquí de autoevaluación (SAM). Durante el experimento, los participantes ven vídeos que evocan emociones como diversión, aburrimiento, relajación y miedo. El conjunto de datos WESAD proporciona datos fisiológicos de 15 participantes sometidos a un protocolo experimental que incluye estados de referencia, diversión, meditación y estrés, inducido por una prueba social de estrés Trier. También incluye datos autoinformados basados en cuestionarios y SAM. Las modalidades de sensores incluyen datos ECG muestreados a 700 Hz desde un dispositivo de pecho y señales BVP muestreadas a 64 Hz desde un dispositivo de muñeca.

2.2. Sistemas de extracción de características

En el ámbito de procesamiento de señales acústicas, la extracción de características juega un papel fundamental para descifrar información útil de las señales sonoras. Estas señales, que pueden provenir de diversas fuentes como voz, música, ambiente, entre otras, contienen una gran cantidad de información que puede ser aprovechada para una amplia gama de aplicaciones, desde reconocimiento de habla hasta monitorización ambiental.

Por otro lado, la extracción de características de señales acústicas puede enfrentar una serie de desafíos, desde la alta dimensionalidad y la variabilidad intrínseca de las señales hasta la presencia de ruido y la selección de características adecuadas. Superar estas problemáticas requiere un enfoque cuidadoso y la aplicación de técnicas avanzadas de procesamiento de señales y aprendizaje automático.

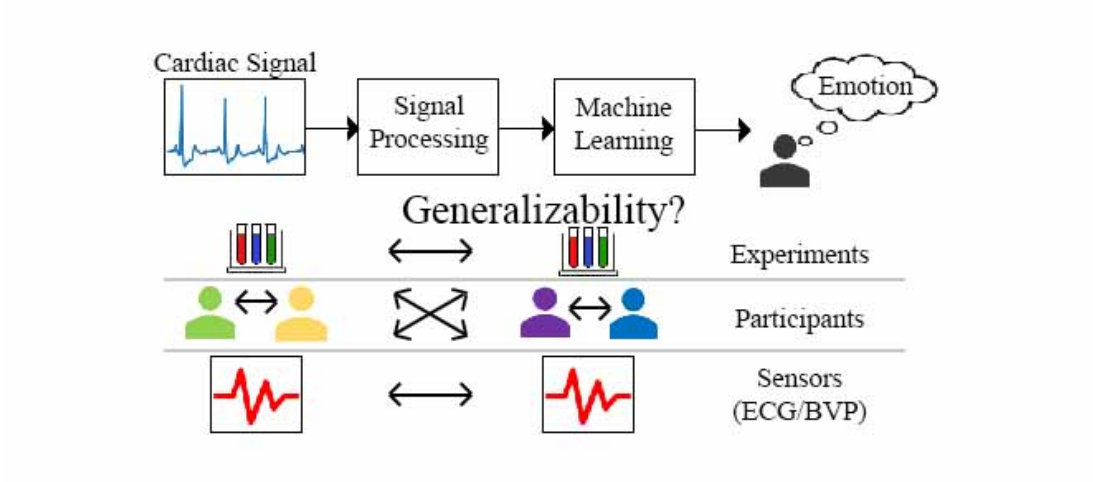


Figura 2.5: Flujo explicativo de cómo, a través de a influencia de diferentes características extraídas de las señales de actividad cardíaca en el rendimiento de los modelos de aprendizaje automático, se puede analizar la efectividad detección de estrés y ansiedad bajo diversas circunstancias, como cambios en la postura corporal, actividad física, o variaciones en las condiciones ambientales [25].

Hay diferentes alternativas de sistemas de extracción de características, por ejemplo GeMaps y GSU Praat Tools.

2.2.1. *GeMAPS (The Geneva Minimalistic Acoustic Parameter Set)*

Es un conjunto de parámetros acústicos desarrollado por el Laboratorio de Procesamiento de Señales Multimedia (MSP Lab) en la Universidad de Ginebra [26]. Estos parámetros están diseñados para capturar características acústicas relevantes en señales de voz hablada.

Incluye una variedad de parámetros, como la frecuencia fundamental, la intensidad de la voz, la duración de los segmentos de habla, así como características espectrales como la energía en diferentes bandas de frecuencia y la frecuencia de pico. Estas características capturan diferentes aspectos de la señal de voz, como el tono y la energía. Además, permite adquirir otras características más especificas de la voz, como la prosodia, referida a los elementos no verbales de la comunicación oral, aquellos aspectos que trascienden las palabras y se transmiten a través de la entonación, el ritmo, el tono y el énfasis. Todas estas características son importantes para el análisis de emociones y otras tareas de procesamiento de voz.

Destaca también su capacidad para capturar tanto aspectos temporales como espectrales de la señal de voz. Esto significa que puede proporcionar información sobre la melodía, el ritmo y el énfasis en el habla, así como características relacionadas con la calidad vocal

y la articulación.

En cuanto a las señales de audio utilizada en el desarrollo y evaluación de GeMAPS, se han utilizado varias bases de datos de voz, incluidas la base de datos de expresión emocional de Geneva (GEMEP), la base de datos de expresión emocional de Berlin (EMO-DB) y la base de datos de emociones de Viena (Vienna Emotions Database). Estas bases de datos contienen grabaciones de voz no etiquetdas de hablantes realizando diferentes expresiones emocionales, que se utilizan para extraer y validar los parámetros acústicos de GeMAPS. En la Figura 2.6 se puede observar la asignación de categorías para las emociones.

Corpus	Activation		Valence	
	low	high	negative	positive
FAU AIBO	-		NEG	IDL
TUM AVIC	loi1	loi2, loi3	loi1	loi2, loi3
EMO-DB	boredom, disgust, neutral, sadness	anger, fear, happiness	angry, sad	happy, neutral, surprise
GEMEP	pleasure, relief, interest, irritation, anxiety, sadness	joy, amusement, pride, hot anger, panic fear, despair	hot anger, panic fear, despair, irritation, anxiety, sadness	joy, amusement, pride, pleasure, relief, interest
SING	neutral, tenseness, sadness, tenderness, calm/serenity, condescension	fear, passionate love, animated joy, triumphant pride, anger	fear, tense arousal, anger, sadness, condescension	neutral, passionate love, animated joy, triumphant pride, tenderness, calm/serenity
VAM	q2, q3	q1, q4	q3, q4	q1, q2

Figura 2.6: Asignación de categorías de emociones específicas del conjunto de datos a etiquetas de activación binarias (bajo/alto) y etiquetas de valencia binarias(negativo/positivo) [26].

2.2.2. *GSU Praat Tools*

Estas herramientas fueron desarrolladas por el Grupo de Investigación en Fonética y Fonología de la Universidad Estatal de Georgia (GSU). Están dirigidas a investigadores, lingüistas y profesionales que trabajan en el análisis de señales de voz [27].

Este conjunto de herramientas consiste en rutinas basadas en texto escritas en el lenguaje de *scripting* de Praat, al cual se accede mediante menús y comandos integrados en la interfaz gráfica de Praat. Esta herramienta se muestra en la Figura 2.7 el modelo.

Está diseñado con el propósito de asistir a los usuarios de Praat en la exploración de conjuntos específicos de sonidos de interés. Además, posibilita realizar diversas operaciones como edición, filtrado, reescalado o modificación de estos sonidos. Asimismo, facilita la ejecución de análisis cuantitativos, cuyos resultados son registrados en archivos de datos fácilmente interpretables. Estos archivos pueden ser leídos sin dificultad en hojas de cálculo o programas estadísticos.

Las GSU Praat Tools ofrecen la posibilidad de extraer una amplia variedad de datos útiles para comprender las propiedades acústicas y fonéticas de las señales de voz. Entre

estos datos se encuentran las características acústicas, como la frecuencia fundamental (F0), la intensidad de la voz, la duración de segmentos de habla y medidas espectrales que detallan la distribución de energía en diferentes frecuencias. Además, proporcionan información prosódica, abarcando aspectos como el ritmo, la entonación y el énfasis en la voz.

Un ejemplo de aplicación de las GSU Praat Tools sería en la investigación lingüística o en el procesamiento de lenguaje natural. Por ejemplo, se supone que un lingüista está interesado en analizar las diferencias prosódicas entre dos dialectos de una lengua. Podría utilizar estos scripts para extraer características prosódicas como la duración de los segmentos de habla y la entonación característica de cada dialecto a partir de grabaciones de hablantes nativos. Luego, podría comparar estas características entre los dialectos para identificar patrones distintivos de entonación y ritmo que diferencian a uno del otro.

Figura 2.7: Interfaz gráfica del GSU Praat Tools [27].

2.3. Bases de Datos

2.3.1. La base de datos Oxford Vocal (OxVoc) Sounds

Esta base de datos, desarrollada por el Laboratorio de Neurociencia del Habla y la Audición de la Universidad de Oxford, contiene una amplia colección de grabaciones de

sonidos vocales de alta calidad [28].

Incluye una variedad de muestras de voz, que van desde fonemas y palabras individuales hasta secuencias de habla continua. Estos ejemplos se pueden ver en la tabla de la Figura 2.8. Estas grabaciones abarcan diferentes variedades de idiomas y acentos, lo que permite a los investigadores explorar la variabilidad lingüística y la percepción auditiva en diversos contextos culturales y lingüísticos.

Una característica destacada de OxVoc Sounds es su cuidadoso diseño experimental y su rigurosa anotación fonética. Cada grabación está meticulosamente etiquetada con información detallada sobre el hablante, el contexto de la grabación y la transcripción fonética precisa. Esto proporciona a los investigadores un marco sólido para realizar experimentos controlados y analizar los datos de manera sistemática.

Esta base de datos se utiliza ampliamente en estudios sobre la percepción del habla, el reconocimiento de patrones vocales, la adquisición del lenguaje y la neurociencia del habla y la audición. Los investigadores pueden acceder a OxVoc Sounds para diseñar experimentos, desarrollar modelos computacionales y realizar investigaciones empíricas que profundicen nuestra comprensión de cómo procesamos y comprendemos los sonidos vocales en el cerebro humano.

Por último, cabe destacar que la base de datos OxVoc Sounds es una herramienta invaluable para la investigación en el campo de la percepción auditiva y la cognición del habla, proporcionando una amplia gama de grabaciones de sonidos vocales cuidadosamente anotadas y diseñadas para abordar una variedad de preguntas científicas sobre el procesamiento del lenguaje humano.

Stimulus category	Number of stimuli	F_0 (Hz), *M* (*SD*)	Number of vocal bursts, *M* (*SD*)	Mean burst duration (s), *M* (*SD*)
Infant cry	21	445.54 (84.81)	1.90 (1.09)	1.02 (0.50)
Infant neutral	25	347.34 (122.34)	1.88 (0.97)	0.93 (0.46)
Infant laugh	18	348.31 (87.95)	3.22 (1.40)	0.41 (0.22)
Adult cry	19	368.22 (94.83)	2.11 (0.32)	0.55 (0.130)
Adult neutral	30	228.13 (57.88)	1.00 (0.00)	0.91 (0.20)
Adult laugh	30	348.83 (104.12)	4.27 (1.78)	0.37 (0.29)
Animal distress	30	439.28 (101.53)	1.63 (0.76)	1.01 (0.42)
Total	173			

Values presented are averaged across stimuli within each stimulus category individual category.

Figura 2.8: Parámetros físicos básicos de los estímulos de vocalización en la base de datos OxVoc [28].

Capítulo 3

Marco Teórico

En este capítulo se definen los sistemas de captura, métodos y parámetros de voz que se utilizan en el desarrollo de este TFG. En la Sección 3.1 se explican teóricamente los sistemas de captura de audio. En la Sección 3.2 se proporciona una introducción detallada a los métodos de ML que se usarán, los cuales son el aprendizaje supervisado y no supervisado, destacando sus principios, aplicaciones y algoritmos más utilizados. Y en la Sección 3.3 se explican las características propias de las señales de voz y las emociones propuestas para este TFG.

3.1. Sistemas de captura de audio

Los sistemas de captura de audio utilizados en ML están diseñados para recolectar, procesar y analizar datos de audio. Estos sistemas son fundamentales para aplicaciones de reconocimiento de voz o detección de emociones por ejemplo. Asimismo, estos sistemas combinan hardware especializado, como micrófonos y convertidores analógico-digitales, con software sofisticado que incluye técnicas de preprocesamiento, extracción de características y algoritmos de aprendizaje automático.

3.1.1. Componentes de captura

En este apartado se muestran los distintos tipos de componentes de captura de audio que se pueden utilizar, donde destacan principalmente los *micrófonos*, las *interfaces de audio* y los *entornos de grabación*.

Los micrófonos son los dispositivos primarios para la captura de audio. En la Figura 3.1 se puede observar un ejemplo de captura de audio mediante el uso de micrófonos. La calidad del micrófono influye directamente en la calidad de los datos capturados, donde

según el tipo puede ofrecer más o menos sensibilidad al igual que robustez ante ruidos ambientales.

Figura 3.1: Ejemplo de captura de audio mediante el uso de micrófonos [29].

Luego las interfaces de audio, donde estos se complementan con los micrófonos, permitiendo conectarlos a ordenadores u otros sistemas de procesamiento. Las interfaces de audio convierten las señales analógicas de los micrófonos en señales digitales que pueden ser procesadas por algoritmos de ML.

Por otro lado también destacan los entornos de grabación. En la Figura 3.2 se puede observar un ejemplo de tres entornos utilizados para grabar. El entorno donde se captura el audio es una parte importante. Las condiciones acústicas, como la reverberación y el ruido de fondo, pueden afectar la calidad de los datos. Se suelen utilizar estudios de grabación, salas anecoicas o configuraciones controladas para minimizar estos efectos.

Figura 3.2: Ejemplo de tres entornos utilizados para grabar, donde el primero sería una oficina, el segundo una cafetería y el tercero una sala insonorizada [29].

3.1.2. Procesamiento de audio y extracción de características

El procesamiento de señales de audio es el primer paso en cualquier sistema de captura de audio para ML. Este proceso incluye la adquisición, preprocesamiento y análisis de señales de audio.

La adquisición de audio [30] se refiere a la captura de ondas sonoras mediante dispositivos como, por ejemplo, los micrófonos. La señal analógica capturada se convierte en una

señal digital mediante un convertidor analógico-digital (ADC). Y el muestreo se realiza a una frecuencia adecuada para asegurar que se conserve la mayor cantidad de información posible. Luego, el preprocesamiento implica la limpieza y preparación de la señal de audio para su análisis que consta de tres pasos: filtrado de ruido, normalización y segmentación.

Por otro lado, la extracción de características es otro paso importante para convertir los datos de audio en un formato adecuado para los algoritmos de ML. Se cuenta con varias técnicas para poder realizarlo, por ejemplo, *la transformada de Fourier*, el *espectrograma*, *Coeficientes Cepstrales de Frecuencia Mel* (MFCCs), *Zero-Crossing Rate* (ZCR) [29].

La transformada de Fourier convierte la señal de tiempo en el dominio de la frecuencia, permitiendo el análisis de las componentes frecuenciales de la señal. El espectrograma es la representación visual de las frecuencias del audio a lo largo del tiempo, en la Figura 3.3 se puede observar varios ejemplos de espectro en diferentes entornos utilizando un tipo de dispositivo de grabación. Aparte, el espectro se calcula mediante la transformada de Fourier de corto tiempo (STFT). Los MFCCs proporcionan una representación compacta de la envolvente del espectro de potencia de la señal de audio. Son una de las características más utilizadas en el reconocimiento de voz. Y el ZCR es el número de veces que la señal cruza el eje cero en un período de tiempo dado, útil para la clasificación de sonidos.

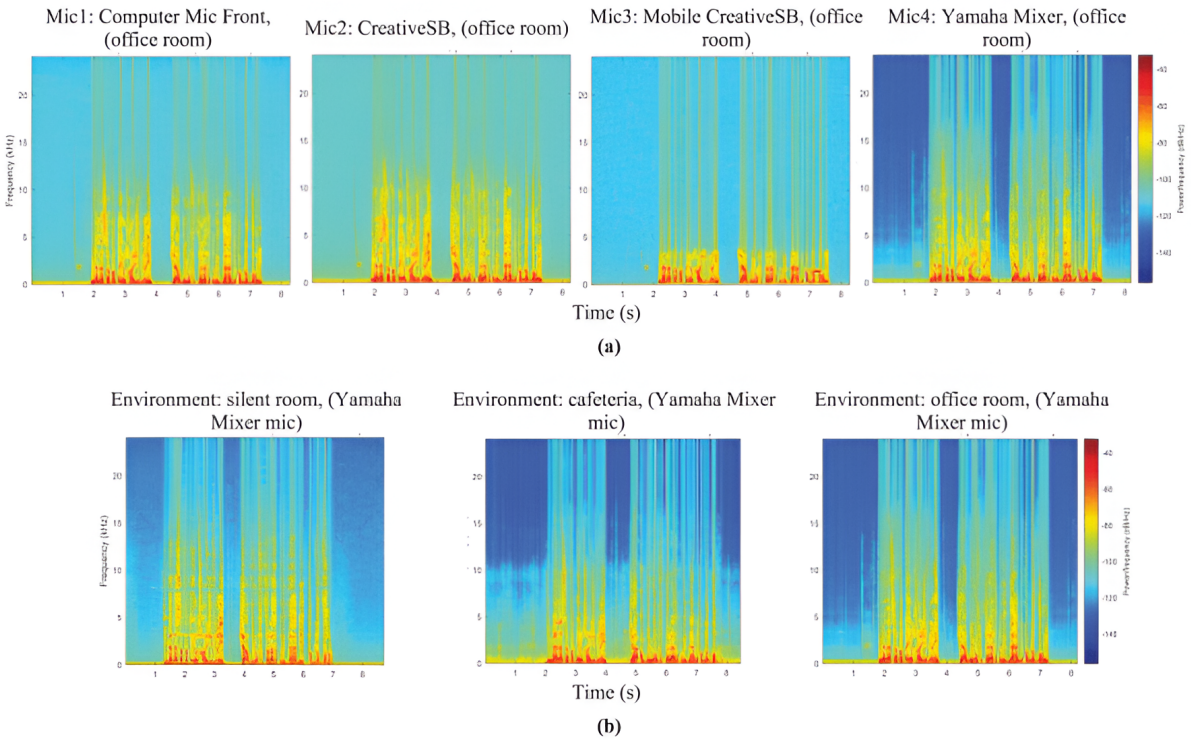


Figura 3.3: Ejemplos de espectrogramas de un hablante que utiliza la misma frase, a) grabados en un entorno (oficina) utilizando cuatro dispositivos de grabación diferentes, b) grabados utilizando un tipo de dispositivo de grabación (mezclador Yamaha) en tres entornos diferentes. [29].

3.2. Métodos de Machine Learning

El ML es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las aplicaciones a aprender y mejorar automáticamente a partir de la experiencia. Para comprender esta actividad, es importante conocer la clasificación de los tipos de algoritmos de aprendizaje de ML, que son el aprendizaje supervisado, el cual se ha elegido para este TFG, y el no supervisado. El aprendizaje supervisado requiere de datos etiquetados y se centra en hacer predicciones, mientras que el aprendizaje no supervisado no requiere etiquetas, se centra en explorar patrones y relaciones. En muchos casos se combinan para obtener una comprensión más completa de datos.

3.2.1. Aprendizaje supervisado

El aprendizaje supervisado es una técnica de aprendizaje automático en la que se entrena un modelo utilizando un conjunto de datos etiquetados [31]. Los datos etiquetados consisten en un conjunto en el que se conoce la respuesta deseada para cada muestra, es decir, incluye una muestra de entradas junto con las respuestas esperadas asociadas. El objetivo principal del aprendizaje supervisado es aprender una función que asigne entradas a salidas, de modo que el modelo pueda hacer predicciones precisas sobre datos sin etiquetar. En este proceso, la entrada se denomina característica y la salida deseada es la etiqueta, esto podemos verlo en la Figura 3.4. Durante el entrenamiento, el modelo aprende a asociar características específicas con etiquetas específicas.

Algunos conceptos básicos relacionados con el aprendizaje supervisado incluyen características y etiquetas, así como una variedad de algoritmos de aprendizaje supervisado. Por ejemplo, entre estos algoritmos se encuentran la regresión lineal y la regresión logística, que se utilizan para problemas de regresión y clasificación respectivamente. Además, el SVM es útil para problemas de clasificación y regresión. Los árboles de decisión y los bosques aleatorios son capaces de abordar problemas tanto de clasificación como de regresión. Por otro lado, las redes neuronales, modelos más complejos inspirados en la estructura del cerebro humano, también son utilizadas en aprendizaje supervisado. En términos de proceso, el conjunto de datos etiquetados se divide en conjuntos de entrenamiento y prueba. El modelo se entrena utilizando el conjunto de entrenamiento y luego se evalúa en el conjunto de prueba para medir su rendimiento en datos no vistos. En la Figura 3.5 podemos ver las fases por las cuales pasa el modelo de aprendizaje supervisado. Finalmente, la evaluación del modelo se realiza mediante métricas de rendimiento como precisión, recall, F1-score (en problemas de clasificación) o error cuadrático medio (en problemas de regresión), para determinar el nivel de rendimiento del modelo.

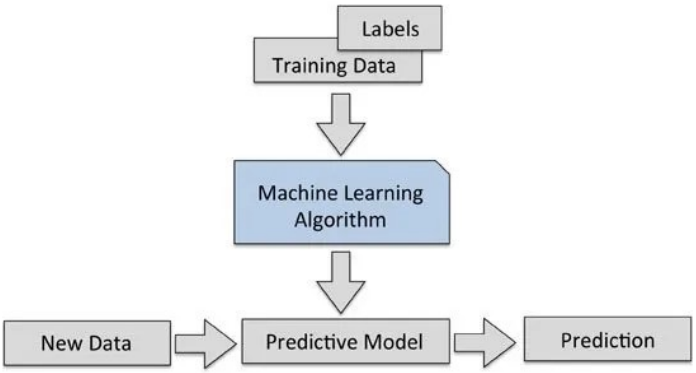


Figura 3.4: Proceso de aprendizaje supervisado. El objetivo principal del aprendizaje supervisado es aprender una función que asigne entradas a salidas, de modo que el modelo pueda hacer predicciones precisas sobre datos sin etiquetar. En este proceso, la entrada se denomina característica y la salida deseada es la etiqueta [31].

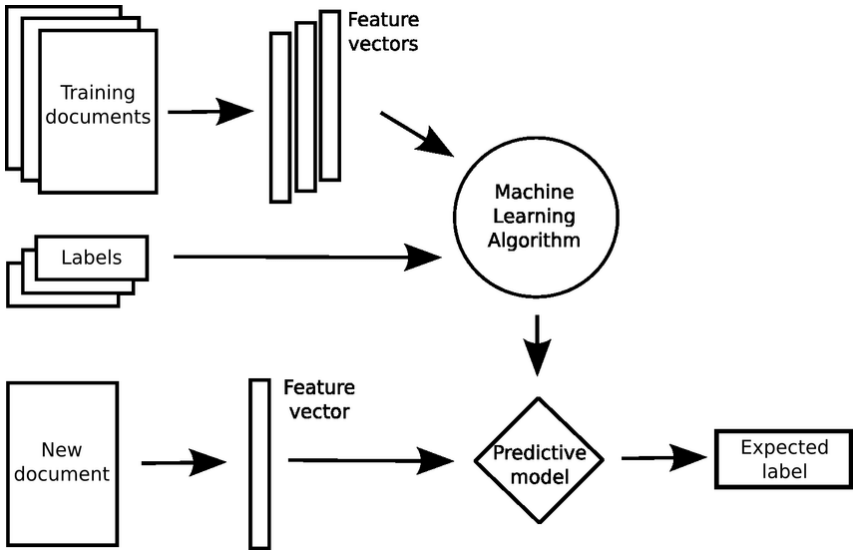


Figura 3.5: Modelo de aprendizaje supervisado. El conjunto de datos etiquetados se divide en conjuntos de entrenamiento y prueba. El modelo se entrena utilizando el conjunto de entrenamiento y luego se evalúa en el conjunto de prueba para medir su rendimiento en datos no vistos [31].

A continuación, se explican algunos algoritmos de *ML* basados en el aprendizaje supervisado como son RF, el LDA y el k-NN. Estos algoritmos se utilizan para realizar tareas específicas, como clasificación o regresión, basándose en ejemplos etiquetados previamente.

3.2.1.1. Random Forest

Es un algoritmo de aprendizaje supervisado que se utiliza tanto en problemas de clasificación como en problemas de regresión [32], [33]. Es una técnica de conjunto *(ensemble learning)* que combina múltiples árboles de decisión individuales para crear un modelo más robusto y preciso.

RF es un algoritmo de aprendizaje automático que destaca por varios aspectos clave. En primer lugar, se basa en la construcción de múltiples árboles de decisión, donde cada árbol se entrena de manera independiente utilizando un subconjunto aleatorio de datos de entrenamiento y características. Este enfoque de construcción de árboles permite que cada uno capture diferentes aspectos del conjunto de datos, aumentando así la diversidad del modelo. En la Figura 3.6 se puede ver la estructura de entrenamiento y test de un modelo de RF. Durante la construcción de cada árbol, se utiliza un muestreo aleatorio con reemplazo en el conjunto de datos de entrenamiento, lo que se conoce como bootstrap sampling. Además, en cada nodo de división, se realiza un muestreo aleatorio de características. Estas dos capas de aleatoriedad ayudan a decorrelacionar los árboles y a mejorar la generalización del modelo, reduciendo así el riesgo de sobreajuste.

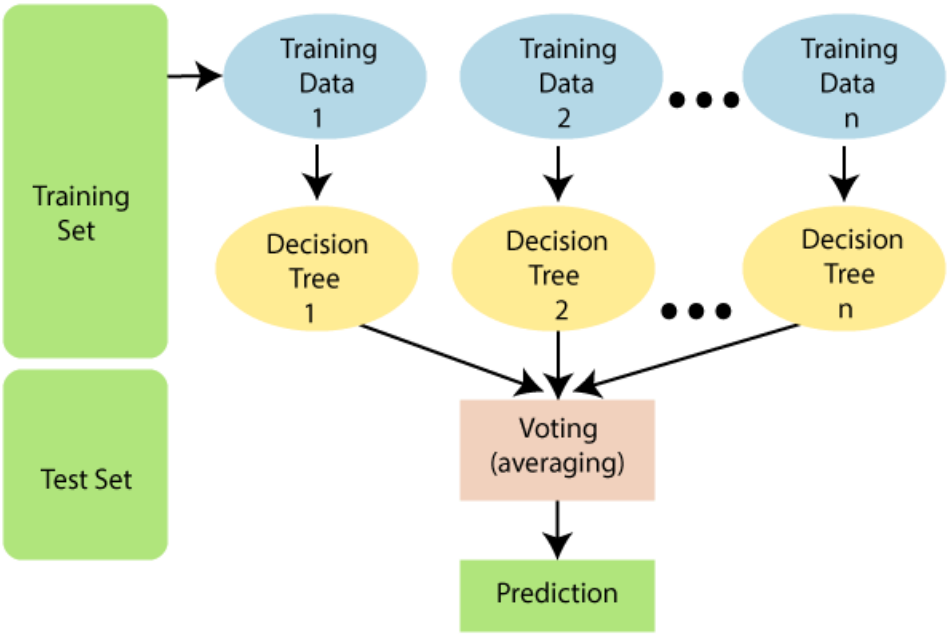


Figura 3.6: Estructura RF donde dados los datos de entrenamiento al pasar por el procesamiento por los árboles de decisión se llega a la predicción [32].

En cuanto a la predicción, en el caso de clasificación, cada árbol emite una predicción y la clase final se determina por votación. Mientras que en el caso de regresión, las predicciones de cada árbol se promedian para obtener la predicción final del bosque. Esta técnica de votación o promediación contribuye a mejorar la precisión del modelo final.

Una de las ventajas más destacadas de RF es su robustez y resistencia al sobreajuste, gracias a la introducción de aleatoriedad durante la construcción de los árboles, lo que ayuda a reducir la variabilidad del modelo. Además, proporciona una medida de la importancia de cada característica en la predicción, lo que puede ser útil para la selección de características y para comprender mejor el modelo.

RF es altamente versátil y puede aplicarse a una variedad de problemas, incluyendo clasificación y regresión. Es fácil de usar y generalmente no requiere una afinación extensa de hiperparámetros. Además, dado que la construcción de árboles es un proceso independiente, el algoritmo puede ser fácilmente paralelizado, lo que lo hace eficiente en términos de tiempo de entrenamiento en sistemas computacionales con recursos adecuados.

3.2.1.2. Análisis Discriminante Lineal (LDA)

Es una técnica de aprendizaje automático supervisado utilizada para la clasificación y reducción de dimensionalidad [34]. El objetivo principal de LDA es encontrar la combinación lineal de variables predictoras que mejor discrimine entre dos o más clases. Esta combinación lineal se elige para maximizar la separación entre las medias de las clases y minimizar la dispersión dentro de cada clase.

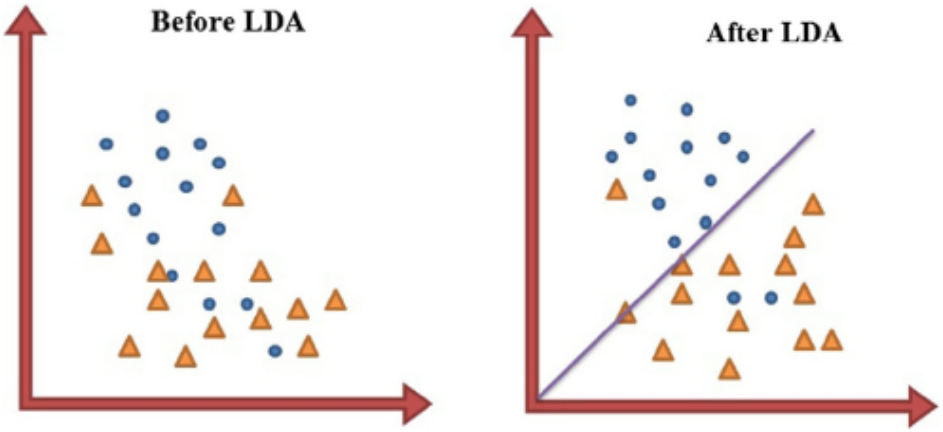


Figura 3.7: Ejemplo básico de la técnica LDA. Se puede observar que hay dos tipos de clase, que son los círculos azules y los triángulos naranjas. Asimismo, tras aplicar el LDA, se realiza la separación de las clases consiguiendo reducir la dimensionalidad conservando la mayor cantidad posible de información [35].

Para lograr esto, LDA calcula la media y la matriz de dispersión de las características para cada clase. Luego, utiliza estas estadísticas para calcular la matriz de dispersión total

y los coeficientes del discriminante lineal. Estos coeficientes definen el hiperplano óptimo en el espacio de características que maximiza la separación entre clases. Una vez que se han determinado estos coeficientes, se pueden utilizar para clasificar nuevas instancias según la clase a la que pertenezcan en función de su posición relativa con respecto al hiperplano discriminante.

Además de su uso como técnica de clasificación, LDA también se utiliza comúnmente para la reducción de dimensionalidad. En este contexto, el objetivo es proyectar los datos originales en un espacio de características de menor dimensión mientras se conserva la mayor cantidad posible de información discriminativa.

Los pasos seguidos en la realización del LDA son los siguientes:

1. Recopilación de Datos: Se requiere un conjunto de datos etiquetado con información de clase para realizar el análisis de discriminante lineal.
2. Cálculo de Medias y Covarianzas: Se calculan las medias y las matrices de covarianza para cada clase en el conjunto de datos.
3. Cálculo de la Matriz de Dispersiones: Se calcula una matriz de dispersiones que mide la variabilidad entre las clases y dentro de las clases.
4. Cálculo de Vectores y Valores Propios: Se calculan los vectores y valores propios de la matriz resultante de la inversa de la matriz de dispersiones multiplicada por la matriz de medias.
5. Selección de Componentes Discriminantes: Se seleccionan las componentes discriminantes basadas en los vectores propios más grandes.
6. Proyección y Clasificación: Los datos se proyectan en el espacio definido por las componentes discriminantes y se utilizan para clasificar nuevas observaciones.

LDA se utiliza en campos como reconocimiento facial, clasificación de documentos, y en general, en cualquier situación donde se necesite separar eficientemente diferentes clases basándose en características observadas.

3.2.1.3. k-Nearest Neighbors

Es un algoritmo de aprendizaje supervisado utilizado para clasificación y regresión [36]. Tiene como objetivo clasificar o predecir un punto de datos basándose en los puntos de datos vecinos en el espacio de características. En la clasificación, este algoritmo asigna una etiqueta a un punto de datos basándose en la mayoría de las etiquetas de los k puntos más cercanos. En la Figura 3.8 se puede ver un ejemplo de k-NN. Por otro lado, en la regresión,

k-NN predice el valor numérico promedio de los k puntos más cercanos. La elección de este parámetro afecta la suavidad de la decisión y la sensibilidad a los detalles locales. La función de distancia es fundamental en k-NN, ya que determina la similitud entre puntos en el espacio de características. Comúnmente se utiliza la distancia euclidiana u otras medidas de distancia según el contexto del problema.

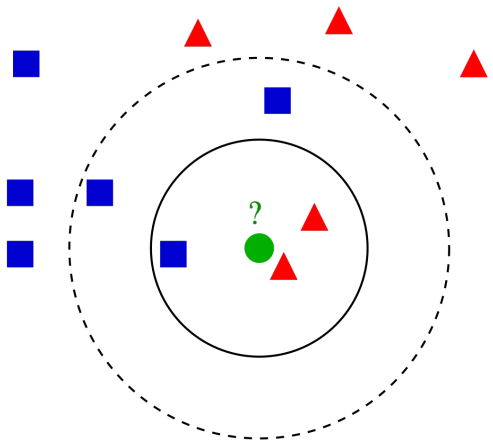


Figura 3.8: Ejemplo de clasificación k-NN. La muestra de prueba (punto verde) debe clasificarse en cuadrados azules o triángulos rojos. Si k = 3 (círculo de línea continua) se asigna a los triángulos rojos porque hay 2 triángulos y sólo 1 cuadrado dentro del círculo interior. Si k = 5 (círculo de línea discontinua), se asigna a los cuadrados azules (3 cuadrados frente a 2 triángulos dentro del círculo exterior) [36].

En cuanto a la toma de decisiones, en la clasificación, se realiza una votación mayoritaria entre los k vecinos más cercanos para determinar la clase del nuevo punto, mientras que en la regresión, se calcula el promedio de los valores numéricos de los k vecinos más cercanos.

Es importante tener en cuenta la sensibilidad a la escala en k-NN. Dado que este algoritmo puede ser influenciado por la escala de las características, a menudo es beneficioso normalizar los datos antes de aplicar el algoritmo para garantizar resultados más precisos y estables.

Los pasos a seguir en k-NN son los siguientes:

1. Recopilación de Datos: Se necesita un conjunto de datos etiquetado con características y etiquetas para entrenar el modelo.

2. Elegir k: Se elige el valor de k, que determina cuántos vecinos se deben tener en cuenta al realizar una predicción.

3. Función de Distancia: Se selecciona una función de distancia para medir la similitud entre puntos. La distancia euclidiana es común, pero se pueden utilizar otras medidas según el problema.

4. Predicción: Para clasificar o predecir un nuevo punto, se identifican los k vecinos más cercanos y se toma una decisión basada en la votación o el promedio

El algortimo k-NN se utiliza en diversas áreas, como reconocimiento de patrones, clasificación de imágenes, recomendación de productos, diagnóstico médico, etcétera. No obstante, puede ser computacionalmente costoso, especialmente en conjuntos de datos grandes, ya que requiere calcular la distancia entre el nuevo punto y todos los puntos en el conjunto de datos de entrenamiento.

3.2.2. Aprendizaje no supervisado

Una vez explicado el aprendizaje supervisado y los algoritmos que se basan en él, se procede a la explicación de lo contrario a este, el aprendizaje no supervisado. A diferencia del aprendizaje supervisado, en el no supervisado el modelo no se entrena utilizando un conjunto de datos etiquetados. El objetivo principal es descubrir la estructura inherente de los datos, identificando patrones, relaciones y agrupaciones sin la ayuda de etiquetas preexistentes [37]. En la Figura 3.9 podemos ver un esquema de este análisis:

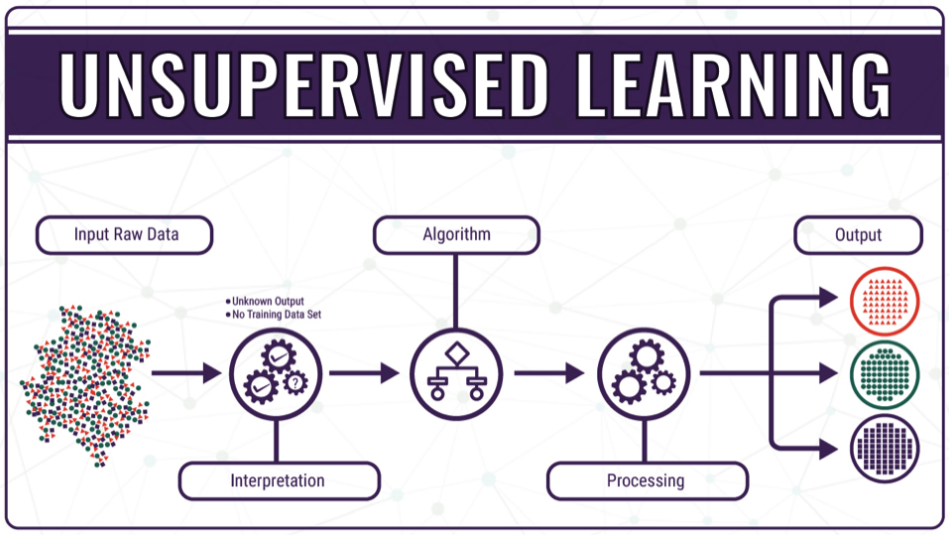


Figura 3.9: Esquema explicativo del proceso del aprendizaje supervisado [37].

El aprendizaje no supervisado engloba varios conceptos fundamentales que son clave para explorar y comprender conjuntos de datos sin la guía explícita de etiquetas o categorías predefinidas. Uno de estos conceptos es el agrupamiento *(clustering)*, que busca dividir un conjunto de datos en grupos o *clusters* de elementos que comparten similitudes entre sí. Algunos algoritmos comunes utilizados para este fin incluyen el k-means, el agrupamiento jerárquico y el *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN).

Otro aspecto importante del aprendizaje no supervisado es la reducción de dimensionalidad, que implica la disminución de la cantidad de variables o características en un conjunto de datos. Esto resulta útil para visualizar datos de alta dimensión o para eliminar redundancia. Algoritmos bien conocidos para la reducción de dimensionalidad son el análisis de componentes principales (PCA) y el t-SNE.

Finalmente, a través del aprendizaje no supervisado también es posible realizar la detección de anomalías. Este proceso implica identificar patrones inusuales o anomalías en los datos que pueden indicar problemas o eventos inesperados. Este enfoque es particularmente valioso en áreas como la seguridad cibernética, la detección de fraudes y el mantenimiento predictivo.

Un algoritmo basado en este análisis, el cuál es mencionado anteriormente, es el *agrupamiento jerárquico*. Este es un método de análisis de datos que busca construir una jerarquía de clusters [38]. Este enfoque puede ser aglomerativo o divisivo y se utiliza comúnmente para explorar la estructura de los datos cuando no se conoce previamente el número de clusters.

El método aglomerativo, comienza tratando cada punto de datos como un cluster individual y luego combina gradualmente los clusters más similares hasta que todos los puntos estén en un solo cluster.

Por otro lado, el enfoque divisivo sigue una ruta inversa. Comienza con un cluster que contiene todos los puntos de datos y luego divide gradualmente este cluster en clusters más pequeños hasta que cada punto esté en su propio cluster.

En el agrupamiento jerárquico, la distancia y el enlace son dos conceptos clave que determinan cómo se calcula la similitud entre los clusters y cómo se fusionan durante el proceso de agrupamiento. La distancia se refiere a la medida de similitud o disimilitud entre dos puntos de datos o entre dos clusters. Puede haber varias formas de calcular la distancia, como la distancia euclidiana, la distancia de Manhattan, la correlación, la distancia de Mahalanobis, entre otras. En cuanto al enlace, se refiere a cómo se decide fusionar los clusters durante el proceso de clustering jerárquico en función de la distancia entre ellos. Hay varios métodos de enlace, incluyendo el enlace único *(single linkage)*, enlace completo *(complete linkage)*, y el enlace promedio *(average linkage)*, entre otros.

- El enlace único fusiona los clusters más cercanos entre sí, es decir, los puntos más cercanos de cada cluster.
- El enlace completo, por otro lado, fusiona los clusters basándose en la distancia más lejana entre los puntos de los clusters.
- El enlace promedio calcula la distancia promedio entre todos los puntos de los clusters antes de fusionarlos.

Como aplicaciones prácticas, se destaca que se puede utilizar en diversas áreas, como biología (clasificación de especies), marketing (segmentación de clientes), y análisis de texto (agrupación de documentos), entre otros.

Como ventaja, destaca que no se requiere especificar el número de clusters de antemano ya que proporciona una jerarquía de clusters y como desventaja, que puede ser computacionalmente costoso, especialmente para grandes conjuntos de datos.

Finalmente, una vez comentado ambos tipos de algoritmos de aprendizaje de Machine Laerning con sus respectivos métodos, este TFG tendrá un enfoque centrado en el aprendizaje supervisado, ya que a través de la utilización de datos etiquetados se puede desarollar modelos predictivos más precisos y eficientes. Aparte, también permite una mejor comprensión de los datos subyascentes y las relaciones que existen entre ellos.

3.3. Parámetros de señales de voz

En esta Sección se explican las características propias de las señales de voz y las emociones propuestas para este TFG, las cuales van a ser descritas en profundidad a continuación.

Las características *Duration*, *PeakTime*, *Amp(dB)* y *PeakAmp* se presentan en la Figura 3.10.

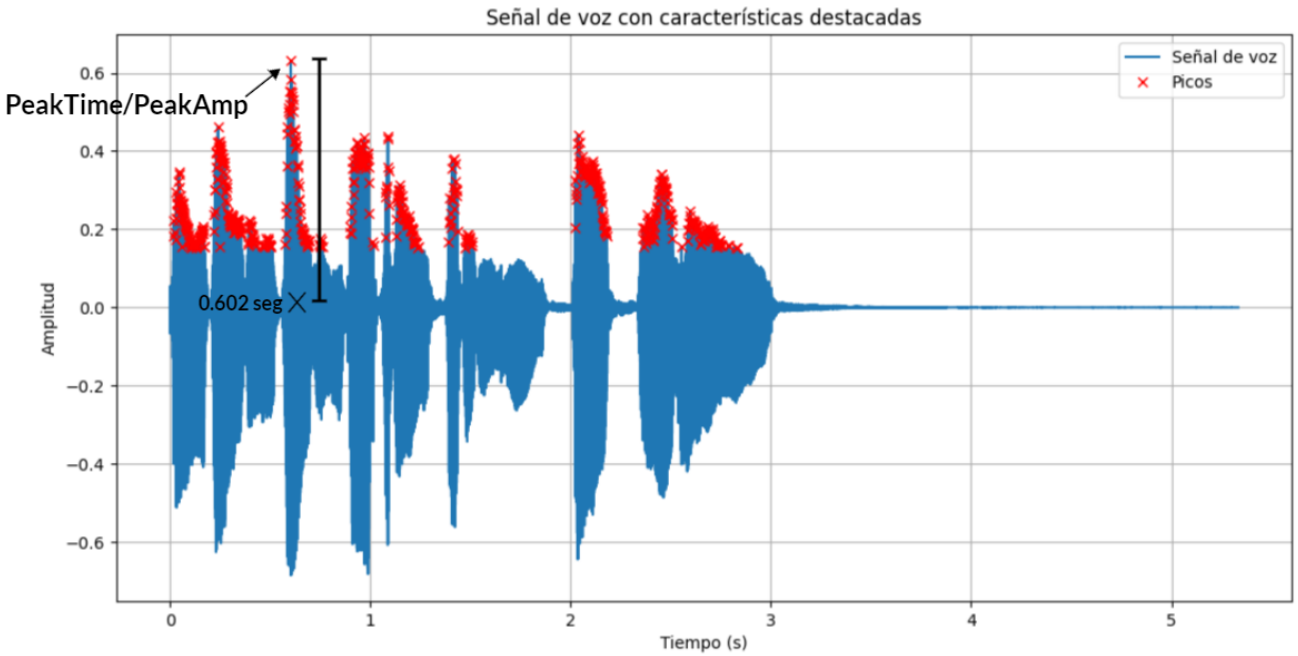


Figura 3.10: Ejemplo de Figura de una señal de voz generada en Python.

- La **duración** se refiere al tiempo total que abarca la grabación o el segmento de audio [39]. Esta puede ser expresada en diferentes unidades de tiempo, siendo milise-

gundos la unidad que se utiliza para esta característica en la base de datos utilizada. Por ejemplo, en el reconocimiento de habla, la duración puede afectar la precisión del reconocimiento y la velocidad de procesamiento, mientras que en el análisis de emociones en el habla, la duración podría ser un factor relevante para comprender cómo varían las características emocionales a lo largo del tiempo. En la Figura 3.10, la duración se representa en el eje X, donde se puede observar la longitud de la señal, siendo esta de 5.33 segundos para ese ejemplo.

- El **Tiempo de Pico** indica el instante en el que la señal de voz alcanza su valor más alto en términos de intensidad de sonido [40]. El *PeakTime* puede ser útil en diversas aplicaciones de procesamiento de señales de voz, por ejemplo en la segmentación de señales de voz, la detección de eventos importantes, etcétera. En la Figura 3.10 se puede observar en términos de tiempo que en el 0.602 segundos (marcado con una "X"negra) se produce un *PeakTime*.

- La **amplitud** es una medida que describe la intensidad o volumen de la señal de voz [40]. En la Figura 3.10 se representa en el eje Y.

- Cuanto mayor sea el valor de la amplitud, más fuerte será la sensación de sonido que percibimos. También cuando se refiere al valor máximo absoluto de la amplitud en una señal de audio durante un período de tiempo específico se le conoce como **Amplitud de Pico** o *PeakAmp*. En la Figura 3.10 se puede observar en términos de dB, que en el 0.602 segundos (marcado con una X negra) se produce un *PeakAmp*.

- Las características de **frecuencia fundamental mínima**, **máxima** y **media** son medidas relacionadas con la frecuencia fundamental (F0) de una señal de audio, que es la frecuencia más baja perceptible en un sonido y está asociada con la periodicidad de la vibración de las cuerdas vocales en el habla humana [41]. Estas características son importantes en el análisis de la prosodia y la entonación en el habla. Como indica su nombre una indica la frecuencia más baja y mas alta detectada en una señal de audio y el promedio de todas las frecuencias fundamentales. Se mide en Hercios (Hz).

- **La desviación estándar de las frecuencias fundamentales** indica cuánto varía la frecuencia fundamental a lo largo de la señal de audio [41]. Que sea alta sugiere una mayor variabilidad en la frecuencia fundamental, lo que podría indicar cambios en el tono de la voz a lo largo del tiempo, como en la entonación o el énfasis. Por otro lado, que sea baja indica una menor variabilidad en la frecuencia fundamental, lo que sugiere una voz más constante en términos de tono. Se mide en Hercios (Hz).

- **La fluctuación del retardo** o *Jitter* es una ligera desviación en la exactitud de la señal de reloj *(workclock)* [40], la cual puede llegar a afectar a la amplitud, la

frecuencia o la fase de una señal, originando también un posible ruido o cambio abrupto de la señal como se puede observar en la Figura 3.11.

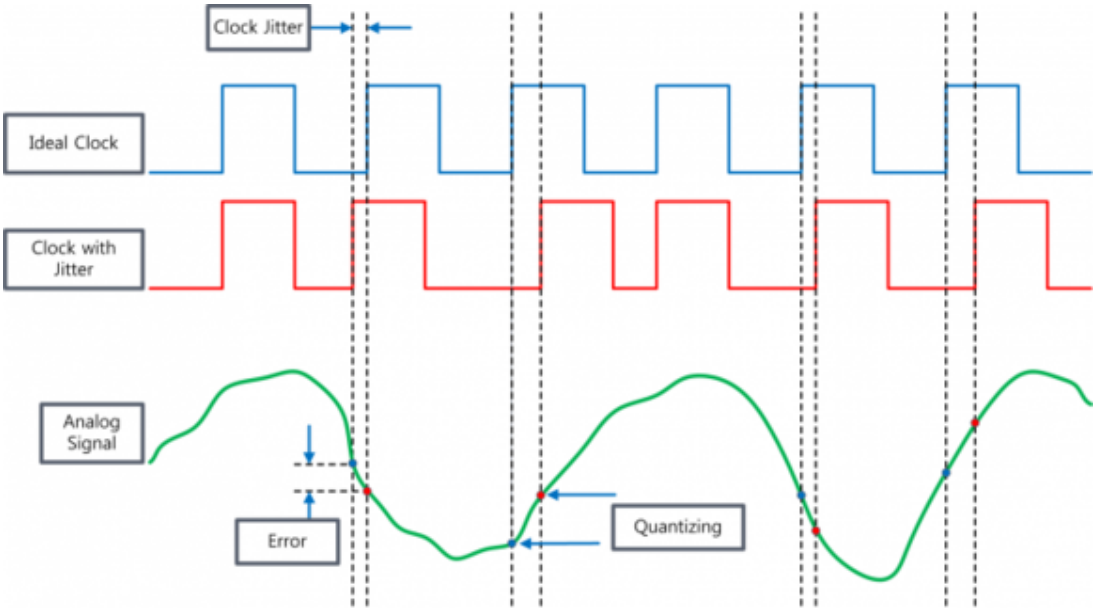


Figura 3.11: Problemas que aparecen cuando se está realizando el muestreo con un reloj de muestreo con fluctuaciones. Debido a las pequeñas diferencias entre el reloj ideal y el reloj con fluctuaciones, los puntos de la señal analógica están siendo muestreados incorrectamente [40].

- El **Shimmer** [42] es una medida de la variación cíclica en la amplitud de la señal de voz, específicamente en relación con la regularidad y estabilidad de la vibración de las cuerdas vocales. Se puede observar mejor esta representación en la Figura 3.12. El *Shimmer* (ShdB) se calcula utilizando la siguiente fórmula:

$$\text{ShdB} = \frac{1}{N-1} \sum_{i=1}^{N-1} 20 \log_{10} \left(\frac{A_i}{A_{i+1}} \right) \tag{3.1}$$

  Donde A_i es la amplitud del i-ésimo ciclo y N es el número total de ciclos en la señal. Se mide en decibelios(dB).

- **El valor máximo y medio y la desviación estándar de la relación señal ruido-armónico**, como indican están relacionadas con la medición de la relación de ruido a armónicos en señales de audio, lo que proporciona información sobre la calidad de la señal y la presencia de ruido en ella [42]. Estas características se miden en decibelios (dB). En la Figura 3.13 se puede observar el HNR que predomina en la señal, asi como el MinHNR que seria -0.56 dB y el MaxHNR es 0.39 dB.

Y en lo referido a las emociones [7], están basadas en el estudio cuyo objetivo es investigar, si el reconocimiento de la emoción de las expresiones vocales difiere en función del

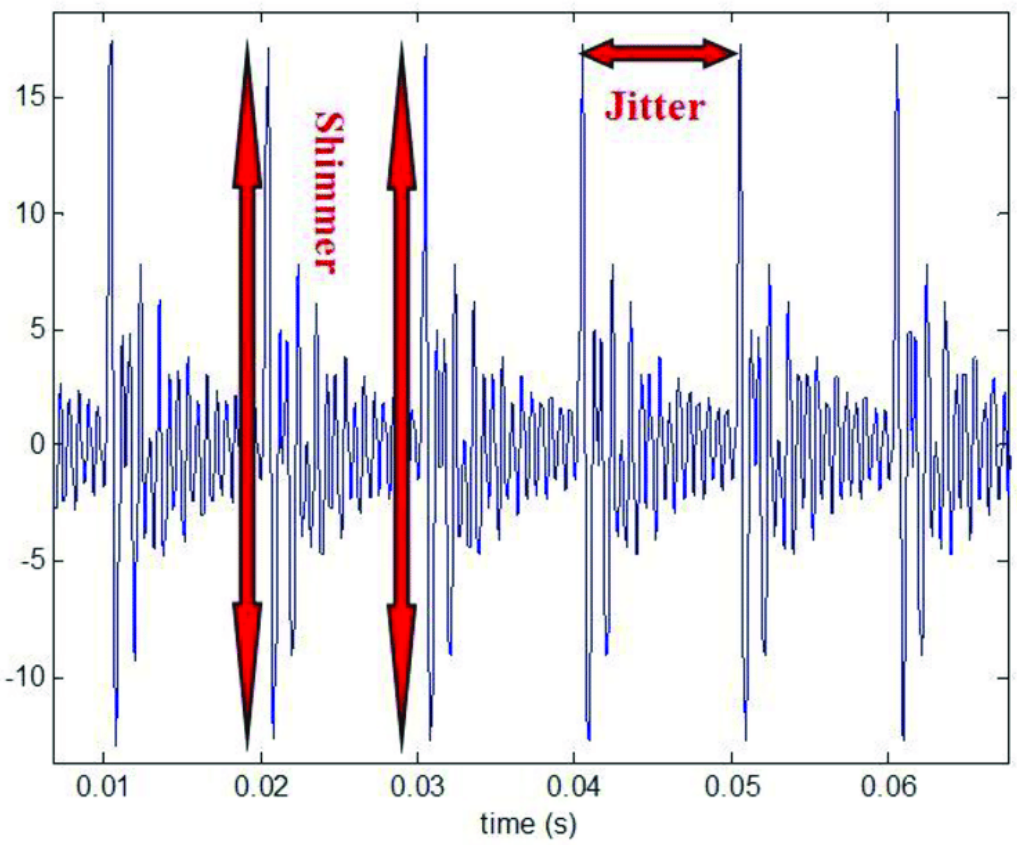


Figura 3.12: Medidas de perturbación Jitter y Shimmer en una señal de voz [43].

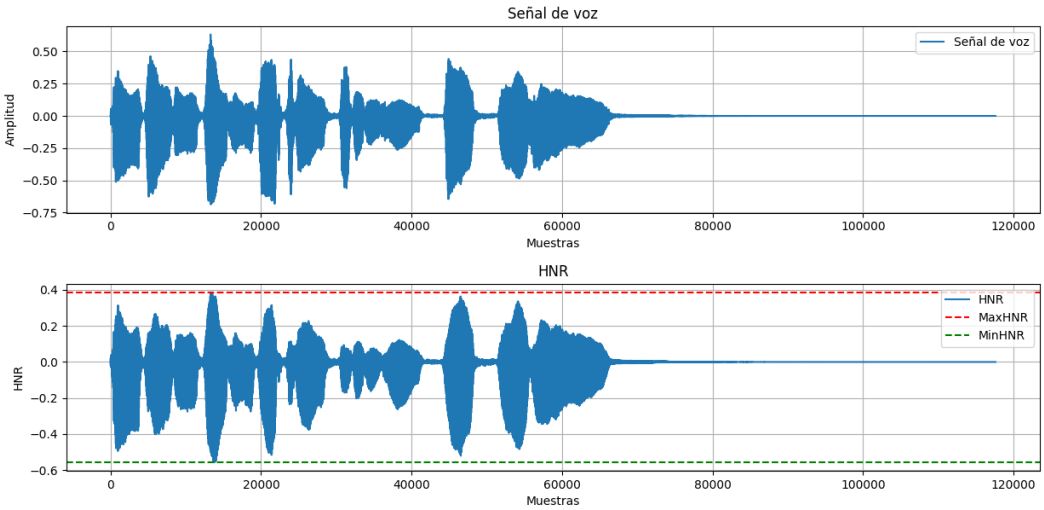


Figura 3.13: Figura generada en Python donde se puede apreciar una señal de voz y el HNR.

género de los decodificadores y codificadores y proporcionar estimaciones de parámetros sobre la magnitud y dirección de estos efectos. Esto lo realizan mediante el análisis de un amplio conjunto de estímulos incrustados en el habla (es decir, palabras, pseudopalabras, frases, pseudosentencias) y vocalizaciones no verbales (es decir, sonidos cortos y distintivos que no forman parte del habla estructurada, pero que transmiten claramente estados emocionales, también llamadas ráfagas de afecto).

Capítulo 4

Metodología y Desarrollo

Este capítulo describe la metodología seguida y el desarrollo del proyecto. Las Figuras 4.1 y 4.2 resumen el flujo del trabajo realizado, que se describe con detalle en las siguientes secciones. En la Sección 4.1 se describe la base de datos utilizada. Por su parte, la Sección 4.2 explica ciertos aspectos u acciones que tiene que pasar la base de datos antes de la realización de los experimentos, los cuales son explicados en la Sección 4.3 aplicados a cada algoritmo. Y por último, en la Sección 4.4 se definen las métricas obtenidas a partir de estos experimentos realizados.

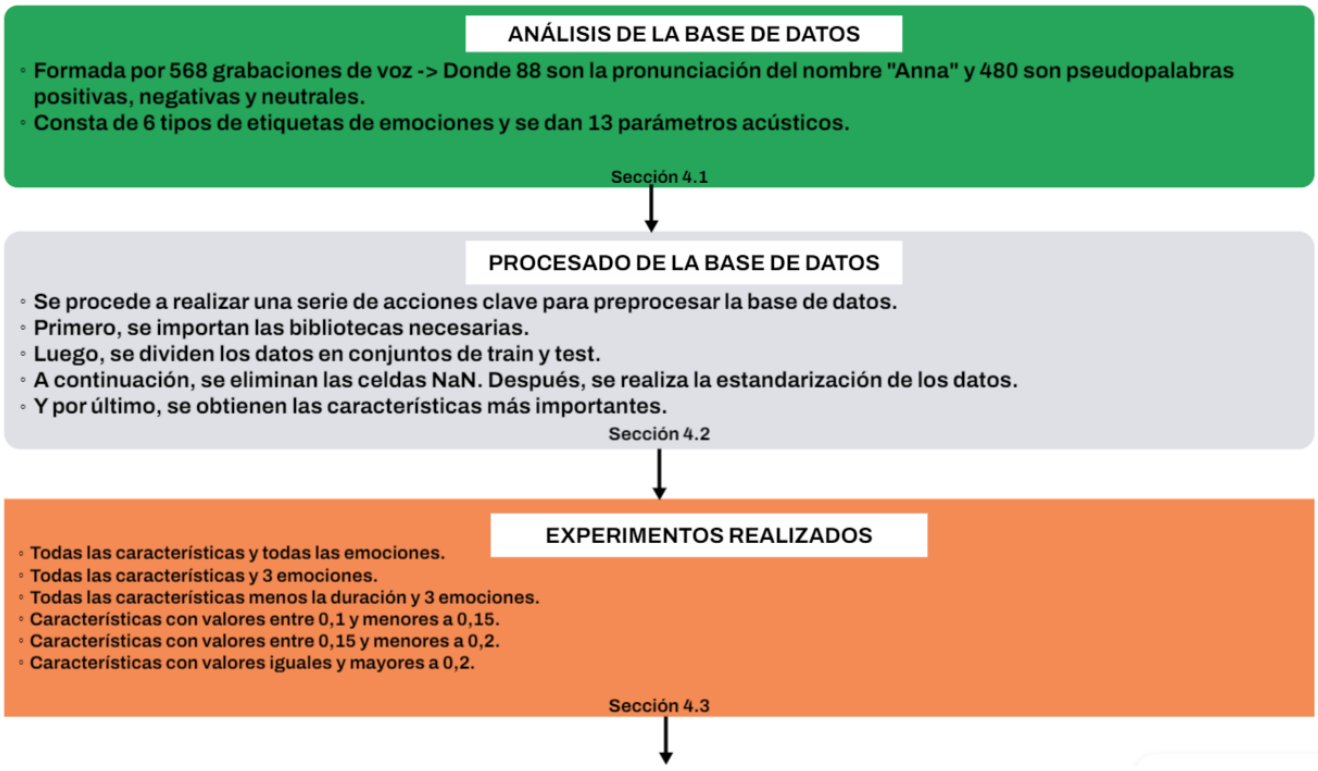


Figura 4.1: Representación del flujo de trabajo seguido (Parte 1).

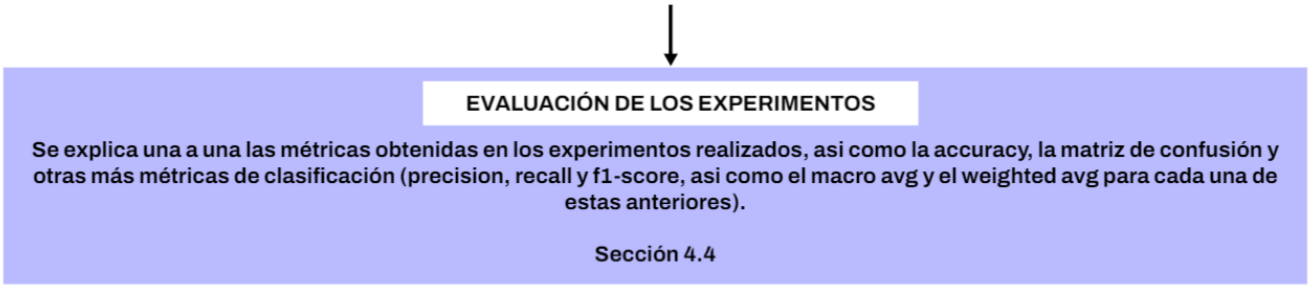


Figura 4.2: Representación del flujo de trabajo seguido (Parte 2).

4.1. Descripción de la base de datos

Este TFG utiliza la base de datos publicada en el articulo *Emotion Recognition and Confidence Ratings predicted by Vocal Stimulus Type and Prosodic Parameters* (ERCRVSTPP) de Adi Lausen y Kurt Hammerschmidt, la cual está disponible en el repositorio Open Science FrameWork (OSF) con el nombre de *AcousticParametersGroupWords* [18]. La base de datos está creada a partir de los corpus de voz (colecciones de grabaciones de voz) *Anna* [19] y *Magdeburg Prosody Corpus* [20], los cuales se explican más adelante.

La base de datos empleada está formada por 568 grabaciones de voz, donde incluye una serie de sonidos formados por hombres y mujeres, pseudo-palabras y sustantivos semánticos positivos, negativos y neutros. Los cuales constan de 6 tipos de etiquetas de emociones, las cuáles son *fear* (miedo), *disgust* (desagrado), *happy* (feliz), *sad* (triste), *neutral* (neutral) y *angry* (enfado) y se dan 13 parámetros acústicos: la duración (*Duration*), el Tiempo de Pico es el pico máximo de amplitud en la señal de voz (*PeakTime*), la amplitud medida en dB (*Amp*), la amplitud máxima alcanzada por la señal de voz (*PeakAmp*), la frecuencia fundamental mínima (*MinF0*), máxima (*MaxF0*) y media (*MeanF0*), la desviación estándar de las frecuencias fundamentales (*StDevF0*), la fluctuación del retardo (*Jitter*), la perturbación de la amplitud (*Shimmer*), el valor máximo (*MaxHNR*) y medio (*MeanHNR*) de la relación señal ruido-armónico y la desviación estándar de la relación señal-ruido armónico (*StDevHNR*). En las Figuras 4.3 y 4.4 se puede observar más detalladamente un fragmento de la base de datos.

SoundName	Database_...	Sex_Voice	Emotion_L...	Begin	End	Duration	PeakTime	Amp(dB)	PeakAmp
andre01Ann...	Anna_Andre	male	neutral	0	0.7799	0.7799	0.318	58.4	75.8
andre03Ann...	Anna_Andre	male	angry	0	0.8513	0.8513	0.466	51.2	76.3
andre04Ann...	Anna_Andre	male	fear	0	0.5971	0.5971	0.325	59.3	75.1

Figura 4.3: Conjuntos de datos presentes en la base de datos empleada. Se muestra el nombre del fichero de audio, la base de datos original a la que pertenece, el sexo del sujeto que reproduce la voz y su emoción. También se muestran parámetros de la grabación en las columnas derechas y una parte de las características de la voz [18].

Cabe destacar que esta base de datos utilizada forma parte de una base de datos

MinF0	MaxF0	MeanF0	StDevF0	Jitter	Shimmer	MaxHNR	MeanHNR	StDevHNR
75.3	99.6	85	6.15	0.005902351	1.534172192	26.27	4.8	2.91
190.9	254.3	234.3	18.7	0.000501592	1.183032832	24.72	1.35	5.72
109.1	184.2	153.3	24.68	0.002341748	1.287565895	18.89	6.65	5.04

Figura 4.4: Conjuntos de datos presentes en la base de datos empleada. Se muestran sólo el resto de las características de la voz [18].

original que consta de cinco corpus de voz que contienen frases y palabras [18]. Consta de 1038 archivos de audio [7], los cuales se han captado mediante grabaciones de historias con el dispositivo móvil por actores entrenados utilizando su voz para representar una emoción. En lo referido al análisis acústico de estos, para extraer las características mediante el software fonético *Praat* en el ordenador y luego, el volumen del sonido de los ensayos de práctica con un sonómetro profesional. En la Figura 4.5 se puede ver más detalladamente esto que se indica.

Speech corpora	Description of content	Initial content	Number of selected files
Anna (Hammerschmidt and Juergens, 2007)	Name "*Anna*" uttered for 8 emotions [(*anger, affection, contempt, despair, fear, happiness, sensual satisfaction, triumph*) + *neutral* (baseline expression)] by 22 *German drama students* [10 males (M); 12 females (F)] (same for all emotions).	198 audio files	88 [(emotion category of interest + baseline expression) ×22 speakers]
Montreal Affective Voices (Belin et al., 2008)	Portrayals of *non-verbal emotional sounds/affect bursts* (e.g., laughing, crying) for 8 emotions [(*anger, disgust, fear, happiness, pain, pleasure, sadness, surprise*) + baseline expression (*neutral*)] by 10 *francophone actors* (5 M; 5 F) (same for all emotions).	90 audio files	70 [(emotion category of interest + baseline expression) ×10 Speakers]
Berlin Database of Emotional Speech (Burkhardt et al., 2005)	Portrayals of 6 emotions [(*anger, boredom, disgust, fear, happiness, sadness*) + baseline expression (*neutral*)] by 10 *German untrained actors* (5 M; 5 F). The database consists of 10 *semantic neutral sentences* (same for all emotions).	816 audio files	120 [(emotion category of interest + baseline expression) ×2 Speakers × 10 sentences]
Magdeburg Prosody Corpus (Wendt and Scheich, 2002)	Portrayals of 5 emotions [(*anger, disgust, fear, happiness, sadness*) + baseline expression (*neutral*)] by 2 *German actors* (1 M; 1 F). The corpus consists of 3318 *nouns* classified according to their positive-, negative- and neutral semantic content and of 222 *pseudo-words* (same for all emotions).	3318 audio files (nouns)+222 audio files (pseudo-words)	480 [(all emotions + baseline expression) ×2 Speakers × 10 nouns per semantic category (i.e., positive, negative, neutral)/10 Pseudo-words]
Paulmann Prosodic Stimuli (Paulmann and Kotz, 2008; Paulmann et al., 2008)	Portrayals of 6 emotions (*anger, disgust, fear, happiness, sadness, surprise*) + baseline expression (*neutral*) by 2 *German actors* (1 M; 1 F). The stimulus set consists of 210 *lexical sentences* and 210 *pseudo-sentences* (different for each emotion).	420 audio files	280 [(10 lexical sentences & 10 pseudo-sentences for each emotion + baseline expression) ×2 speakers]

Figura 4.5: Descripción de la base de datos original formada por los cincos corpus de voz. Se puede observar detalladamente cada corpus de voz, así como su descripción, el total de archivos de audio que contiene y los seleccionados [18].

El primer corpus de voz [19] trata sobre la pronunciación del nombre *Anna* para ocho emociones. Las emociones son ira, afecto, desprecio, desesperación, miedo, felicidad, satisfacción sensual, triunfo más una expresión neutra. El sonido se reprodujo por 22 estudiantes alemanes, donde 10 eran hombres y 12 eran mujeres. Del total de las grabaciones de audio, se escogen 88 archivos para la base de datos.

El segundo consiste en representaciones de sonidos/explosiones de efectos emocionales no verbales (por ejemplo, risa, llanto) para ocho emociones [44]. Las emociones son ira,

asco, miedo, felicidad, dolor, placer, tristeza, sorpresa más una expresión de referencia neutra. El sonido se reprodujo por 10 personas francesas donde cinco eran hombres y las otras cinco eran mujeres, donde se escogen 70 archivos.

El tercero es sobre la representación de seis emociones. Las emociones son ira, aburrimiento, asco, miedo, felicidad, tristeza más una expresión de referencia neutra [45]. El sonido se reprodujo por 10 actores alemanes donde 5 eran hombres y los otro 5 mujeres. La base de datos consta de 10 frases semánticamente neutras (iguales para todas las emociones). Se escogieron 120 archivos para la base de datos.

El cuarto trata de las representaciones de cinco emociones [20]. Las emociones son ira, asco, miedo, felicidad, tristeza más una expresión de referencia neutra. El sonido se reprodujo mediante dos actores alemanes, una mujer y un hombre. El corpus consta de 3318 sustantivos clasificados según su contenido semántico positivo, negativo y neutro y de 222 pseudopalabras (las mismas para todas las emociones). Se escogieron para este 480 archivos.

Y por último, el quinto es sobre la representación de seis emociones [46]. Las emociones son ira, asco, miedo, felicidad, tristeza, sorpresa más una expresión de referencia neutra por dos actores alemanes, hombre y mujer. El conjunto de estímulos consta de 210 frases léxicas y 210 pseudofrases (diferentes para cada emoción). Para este último se escogen 280 archivos.

Por lo tanto, el motivo de que se haya hecho más enfoque a las palabras y pseudopalabras, es debido a que el archivo *AcousticParametersGroupWords* se amolda más a las capacidades disponibles, ya que la base de datos es mas compacta y balanceada, así como accesible, ya que el archivo perteneciente a las frases era demasiado extenso como para trabajarlo con *Python*. Por lo tanto, la base de datos seleccionada para su procesamiento en este TFG, está formada por el primer corpus de voz de la tabla perteneciente a *Anna* [19] y el cuarto corpus de voz perteneciente *Magdeburg Prosody Corpus* [20]. Cuenta con 568 archivos de audio, 6 tipos de etiquetas de emociones, las cuáles son *fear*, *disgust*, *happy*, *sad*, *neutral* y *angry* y se dan 13 parámetros acústicos: *Duration*, *PeakTime*, *Amp*, *PeakAmp*, *MinF0*, *MaxF0*, *MeanF0*, *StDevF0*, *Jitter*, *Shimmer*, *MaxHNR*, *MeanHNR* y *StDevHNR*.

Una manera más representativa de entender la distribución de los datos en la base de datos empleada, se puede observar en la Figura 4.6. Además, de cómo queda la base de datos utilizada en lo referido a las emociones en la Figura 4.7, donde se aprecia el número de muestras que se tiene para cada emoción.

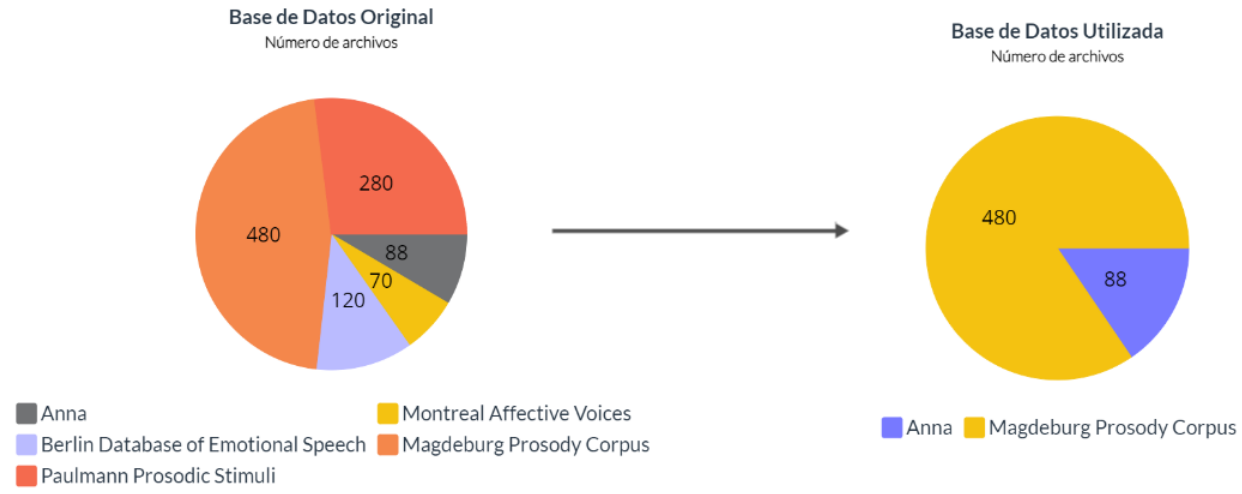


Figura 4.6: Esquema explicativo de la transición de la base de datos original a la utilizada.

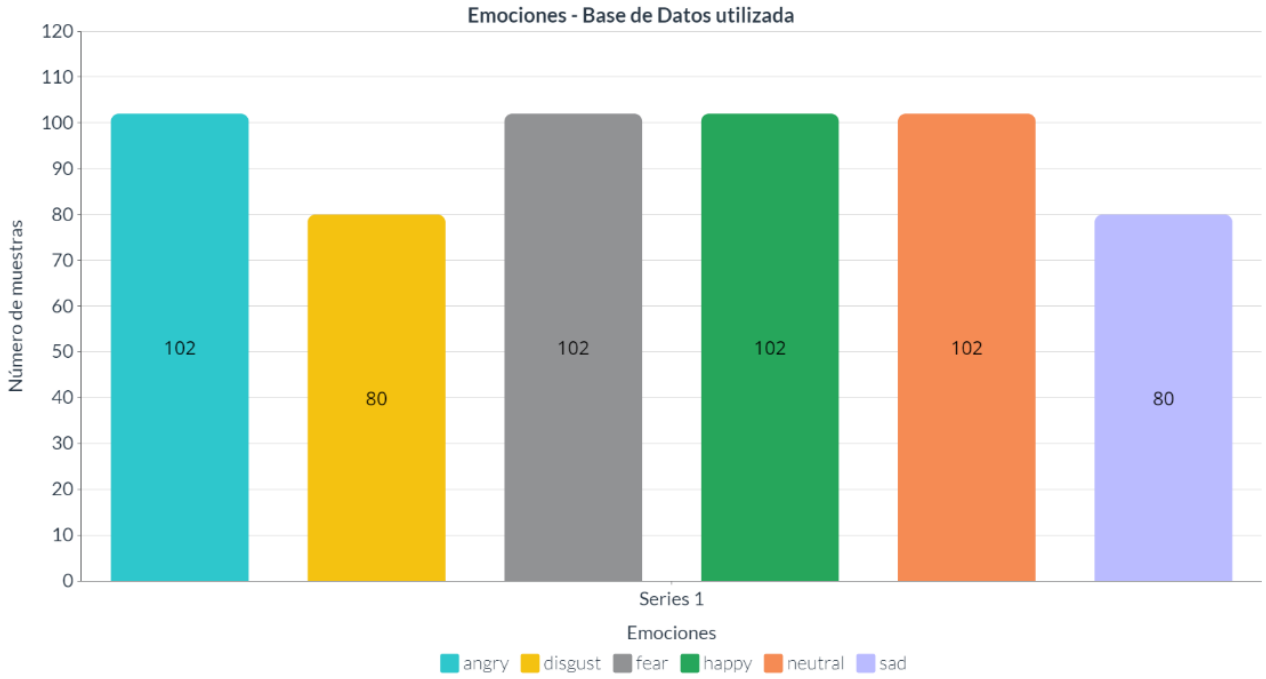


Figura 4.7: Número de muestras que se tiene para cada emoción en la base de datos utilizada, donde se puede observar que se trata de una base de datos balanceada. Se trabaja para este TFG con seis emociones, las cuales son *fear*, *disgust*, *happy*, *sad*, *neutral* y *angry*.

4.2. Procesado de la Base de Datos

Para la elaboración de los experimentos, se ha realizado previamente un procesado de la Base de Datos para poder llevarlos a cabo. Una de las acciones que forman parte de este procesado es la eliminación de los datos NaN (Not a Number), es decir, datos que no han sido almacenados en la Base de Datos. La eliminación de los datos NaN se ha procedido a mediante la función *dropna()*. En este trabajo, el número no es muy elevado ya que mediante la función *dropna()* se identifican solo dos casos para las emociones de miedo e ira, quedándose así en 566 muestras de voz.

Otra acciones a destacar es la estandarización de los datos mediante la función *StandardScaler()*. Es un proceso de transformación de los datos de tal manera que cada característica tenga una media de 0 y una desviación estándar de 1. Más detalladamente, para cada característica en el conjunto de datos *StandardScaler()*, se calcula la media y la desviación estándar. Luego, cada valor de la característica se transforma usando la fórmula:

$$z = \frac{x - \mu}{\sigma} \tag{4.1}$$

Donde x es el valor original de la característica, μ es la media de la característica, σ es la desviación estándar de la característica y z es el valor estandarizado.

Esto se ha realizado en el caso de que estos datos pueden contener características con diferentes escalas (amplitudes, frecuencias, duraciones), ya que si no se estandarizan, las características con mayores magnitudes pueden afectar y destacar más que otras en el proceso de aprendizaje del modelo.

Asimismo, también se han llevado a cabo otra serie de aspectos que sirven como fundamentos base, como por ejemplo, la importación de las bibliotecas necesarias para cada algoritmo. Para el RF se usa el *RandomForestClassifier*. Donde para este método se deja por defecto los parámetros de $n_{estimators} = 100$ que es el número de árboles en el bosque y el $random_{state} = 42$ que hace que el generador de números aleatorios de *scikit-learn* siempre genere la misma secuencia de números aleatorios cada vez que se ejecute el código. Por otro lado, para el LDA, se usa el *LinearDiscriminantAnalysis*. Y para el k-NN, se utiliza el *KNeighborsClassifier* teniendo en cuenta el valor del parámetro k, que es el número de vecinos a considerar. En nuestra propuesta, en este último hemos analizado los valores de 3, 5 y 10 para k.

En lo referido a las características, tiene importancia el realizar un análisis de estas, ya que al predominar muchas características, puede dar lugar a *overfitting* donde podrían capturarse ruido o anomalías en los datos. Es por esto, que la obtención de las características más importantes tiene gran importancia. Se realiza mediante el método *Mutual*

Info Classif, donde se establecerán tres umbrales. El primero pertenece a los valores de las características más importantes entre $0,1$ y menores a $0,15$. El segundo para los valores de las características más importantes entre $0,15$ y menores a $0,2$. Y el tercero corresponde con los valores de las características más importantes iguales o mayores a $0,2$. Esto se realiza para así obtener resultados más precisos e interpretables para la detección de emociones.

Por último, cabe introducir que habría que realizar la división de los datos en conjuntos de train y test para analizar el funcionamiento de los modelos en datos que previamente no han sido analizados. Y por último, que la obtención de los resultados sean en base a un número de diez iteraciones sobre los datos test, en los cuales se ha realizado la media para el resultados, para así conseguir resultados más fiables.

4.3. Descripción de los experimentos realizados

La Sección 4.2 anterior se refiere a la preparación de la base de datos, sin embargo, los datos que se usan para train y test van a depender de los datos que se analizan en cada uno de los experimentos. Se han diseñado una serie de experimentos para evaluar posibles clasificaciones de emociones considerando un corto o amplio espectro de emociones y con selección de diferentes conjuntos de características.

Cabe destacar que los experimentos se han realizado para observar si al reducir el número de emociones, tener en cuenta o no ciertas características o la obtención de las más importantes, difiere de los valores obtenidos del primer experimento en función de cada algoritmo aplicado y si puede llegar a beneficiar o al caso contrario, perjudicar al valor de la clasificación obtenida en base a las métricas detalladas en la siguiente Sección 4.4.

4.3.1. Experimento 1

El primer experimento tiene como objetivo analizar la capacidad de clasificar un amplio espectro de emociones usando todas las características teóricas de la voz.

Para este experimento se cuenta con 566 muestras de voz, contando con 6 tipos de etiquetas de emociones, las cuales son *fear*, *disgust*, *happy*, *sad*, *neutral* y *angry*. Asimismo, tienen 13 parámetros acústicos: *Duration*, *PeakTime*, *Amp*, *PeakAmp*, *MinF0*, *MaxF0* y *MeanF0*, *StDevF0*, *Jitter*, *Shimmer*, *MaxHNR*, *MeanHNR* y *StDevHNR*.

4.3.2. Experimento 2

Para este segundo experimento el objetivo es ahora analizar la capacidad de clasificar un filtrado espectro de emociones usando todas las características teóricas de la voz, ya que se he procedido a reducir el número de emociones.

Se dispone ahora de un total 283 muestras de voz, ya que al reducir el número de emociones también se reduce el número de muestras de voz. Se cuenta con 3 tipos de etiquetas de emociones, las cuales son *happy*, *sad* y *angry*. Y se tienen los 13 parámetros acústicos: *Duration*, *PeakTime*, *Amp*, *PeakAmp*, *MinF0*, *MaxF0* y *MeanF0*, *StDevF0*, *Jitter*, *Shimmer*, *MaxHNR*, *MeanHNR* y *StDevHNR*.

4.3.3. Experimento 3

Este tercer experimento tiene como objetivo analizar la capacidad de clasificar un filtrado espectro de emociones usando varias características teóricas de la voz, ya que se he procedido a reducir el número de emociones y a la omisión de la característica *Duration*. En lo referido a la característica, se omite *Duration* con el fin de llegar a conocer si esa característica realmente influye o no, o si tiene el mismo peso que las demás características al no considerarse de relativa importancia.

Por lo tanto, se tiene un total de 283 muestras de voz, contando con 3 tipos de etiquetas de emociones, las cuales son *happy*, *sad* y *angry*. Y se tienen ahora 12 parámetros acústicos al haber quitado una característica: *PeakTime*, *Amp*, *PeakAmp*, *MinF0*, *MaxF0* y *MeanF0*, *StDevF0*, *Jitter*, *Shimmer*, *MaxHNR*, *MeanHNR* y *StDevHNR*.

4.3.4. Experimento 4

El cuarto experimento tiene como objetivo el poder analizar la capacidad de clasificar un filtrado espectro de emociones usando las características teóricas más importantes. Se reduce el número de emociones a 3, se omite la característica *Duration* y al haber obtenido las características más importantes, estas se han dividido en tres intervalos, para comprobar si desde menor hasta mayor valor de importancia de las características, la clasificación correcta es más alta o baja. A este experimento le corresponde el primer intervalo, que son las características más importantes con valores entre $0,1$ y menores a $0,15$.

Consta de 283 muestras de voz, contando con 3 tipos de etiquetas de emociones, las cuales son *happy*, *sad* y *angry*. Y se tienen ahora 7 parámetros acústicos de las más importantes en función del intervalo mencionado anteriormente: *PeakTime*, *PeakAmp*, *MinF0*, *MeanF0*, *Jitter*, *Shimmer* y *StDevHNR*.

4.3.5. Experimento 5

El quinto experimento sigue la misma línea que el cuarto experimento, difiriendo en que para este le corresponde ahora el segundo intervalo de las características más importantes, el cual trata de las características más importantes con valores entre $0,15$ y menores a $0,2$.

Está formado por 283 muestras de voz, contando con 3 tipos de etiquetas de emociones, las cuales son *happy*, *sad* y *angry*. Y se tienen ahora 6 parámetros acústicos correspondientes al segundo intervalo mencionado: *MinF0*, *MaxF0*, *MeanF0*, *StDevF0*, *Jitter* y *MeanHNR*.

4.3.6. Experimento 6

El sexto y último experimento cuenta con los mismos aspectos que el cuarto y quinto experimento, pero correspondiéndole esta vez el tercer intervalo de las características más importantes, el cual se basa en tener en cuenta las características con valores iguales y mayores a 0,2.

Está formado por 283 muestras de voz, contando con 3 tipos de etiquetas de emociones, las cuales son *happy*, *sad* y *angry*. Y se tienen ahora 3 parámetros acústicos correspondientes al tercer intervalo comentado: *MaxF0*, *StDevF0* y *MeanHNR*.

El código correspondiente a los experimentos realizados, se encuentran disponible en abierto en el siguiente enlace: Link.

4.4. Métricas de evaluación

Este TFG se basa en realizar clasificaciones, las cuales se evalúan por comparativa con las etiquetas reales de la base de datos en base a una serie de métricas, que se detallan a continuación en esta Sección para validar la propuesta de este TFG [47]. Son las siguientes:

- *Accuracy*: Es el número de predicciones correctas realizadas por el modelo dividido por el número total de predicciones. Su fórmula es la siguiente:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

  Donde TP son los casos positivos que el modelo ha clasificado correctamente como positivos. TN son los casos negativos que el modelo ha clasificado incorrectamente

como positivos. FP son los casos negativos que el modelo ha clasificado correctamente como negativos. Y FN son los casos positivos que el modelo ha clasificado incorrectamente como negativos.

- *Recall* (Exhaustividad): Es el número de predicciones positivas correctas dividido por el número total de casos con la clase positiva (aciertos).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

- *Precision*: Es la exactitud de las predicciones positivas realizadas. Su fórmula es:

$$\text{Precisión } (\textit{Precision}) = \frac{TP}{TP + FP} \tag{4.4}$$

- *F1-score* (Puntuación F1): Combina la *Precision* y la Exhaustividad calculando su media armónica. Su fórmula es:

$$F1 = 2 \times \frac{\textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}} \tag{4.5}$$

- *macro-avg*: También puede denominarse macro media o promedio no ponderado. Esta métrica calcula la métrica para cada clase de forma independiente y luego las promedia sobre el número total de clases. Su fórmula es:

$$\text{Macro Avg}(m) = \frac{1}{N} \sum_{i=1}^{N} m_i \tag{4.6}$$

  Donde N es el número total de clases y m_i es la métrica m calculada para la clase i.

- *weighted-avg*: El promedio ponderado también se denomina media ponderada. Esta métrica puede calcularse tomando el número total de ocurrencias de cada clase y multiplicándolo por la métrica de esa clase. Su fórmula es:

$$\text{Weighted Avg}(m) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \right) \times \frac{TP_i + FN_i}{TP_i + TN_i + FP_i + FN_i} \tag{4.7}$$

  Donde N es el número total de clases, TP_i es el número de verdaderos positivos para la clase i, TN_i es el número de verdaderos negativos para la clase i, FP_i es el número de falsos positivos para la clase i y FN_i es el número de falsos negativos para la clase i.

Otra forma de representar la fórmula sería:

$$\text{Weighted Average} = \frac{\sum_{i=1}^{N}(\text{Métrica}_i \times \text{Soporte}_i)}{\sum_{i=1}^{N} \text{Soporte}_i} \tag{4.8}$$

$\sum_{i=1}^{N}(\text{Métrica}_i \times \text{Soporte}_i)$ representa la suma de las métricas ponderadas por el soporte de cada clase. Y $\sum_{i=1}^{N} \text{Soporte}_i$ representa la suma de los soportes de todas las clases.

- *Matriz de Confusión*: La matriz de confusión es una medida muy utilizada para tratar problemas de clasificación. Puede aplicarse tanto a la clasificación binaria como a los problemas de clasificación multiclase. En la Figura 4.8 se muestra un ejemplo de matriz de confusión para clasificación binaria.

  Las matrices de confusión representan los recuentos de los valores predichos y reales.

Empty Cell	Empty Cell	**Predicted**	
Actual		Negative	Positive
	Negative	TN	FP
	Positive	FN	TP

Figura 4.8: Matriz de confusión para la clasificación binaria [47].

Capítulo 5

Resultados y Discusión

En este capítulo se muestran y discuten los resultados obtenidos de los experimentos anteriormente mencionados para cada algoritmo. En la Sección 5.1 se comentan los resultados obtenidos para el experimento 1, donde se mantienen todas las características y emociones. En la Sección 5.2 se presentan los resultados del experimento 2, que trata de mantener todas las características y filtrar por 3 emociones (*angry*, *happy* y *sad*). En la Sección 5.3 se exponen los resultados del experimento 3, al omitir la característica *Duration* de las demás características y filtrar por las 3 emociones indicadas. En la Sección 5.4 se describen los resultados del experimento 4, al realizarse la obtención de las características más relevantes clasificadas por intervalos los cuales se mencionan en la Sección 4.3 y eligiendo las que cumplen el primer intervalo y se mantiene el filtrado de las 3 emociones. Por otro lado, en la Sección 5.5 se analizan los resultados del experimento 5, misma temática que el experimento 4, solo que se cambia por el segundo intervalo de las características más importantes. Asimismo, para la Sección 5.6 se indican los resultados obtenidos para el experimento 6, que sigue la linea de los experimentos 4 y 5, solo que se elige el tercer intervalo de importancia. Y por último, en la Sección 5.7 se discuten los resultados obtenidos en los seis experimentos, haciendo una comparativa de ellos con los mejores casos para cada uno de los diferentes experimentos y algoritmos.

5.1. Conservación de todas las características y todas las emociones.

En el experimento 1, se han obtenido las métricas presentadas en la Tabla 5.1.De esta forma, se muestran las métricas medias para cada una de las clases.

Para este experimento en lo referido a las métricas, se observa en la Tabla 5.1 que las de mayor valor se obtienen con el método RF, seguidas por el k-NN y por último el LDA. La

Tabla 5.1: Resultados de las métricas obtenidas en el experimento 1 con un *Support* igual a 114 predicciones posibles para cada método. Las abreviaciones *Acc*, *Prec*, *Rec* y *F1* corresponden a las métricas: *Accuracy*, *Precision*, *Recall* y *F1-Score*.

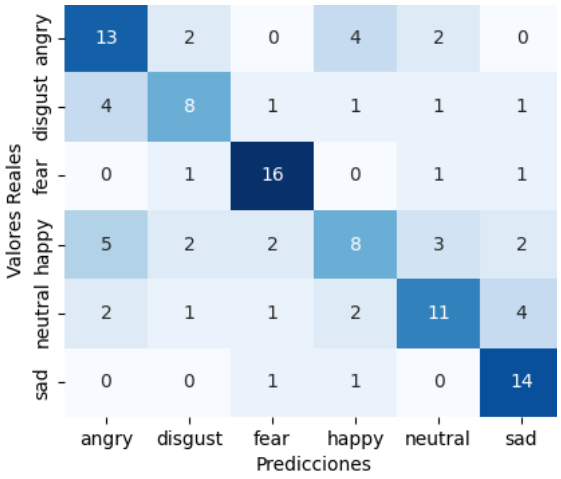
		Macro avg			*Weighted avg*		
Método	*Acc*	*Prec*	*Rec*	*F1*	*Prec*	*Rec*	*F1*
RF	0,606	0,607	0,612	0,601	0,605	0,606	0,597
LDA	0,508	0,503	0,504	0,491	0,500	0,508	0,492
kNN	0,543	0,597	0,532	0,532	0,587	0,543	0,535

métrica *Accuracy* para el RF es de 0,606. Por otro lado, el método k-NN obtiene un valor inferior de *Accuracy*, siendo de 0,543 y mencionando que para este experimento el valor de k que mayor precisión daba es de 10 correspondiente a el número de vecinos. Y por último, para el método LDA presenta una *Accuracy* de 0,508. Estos valores implican que el método que etiqueta mayor proporción de muestras como correctas es RF, superando la mitad de muestras del total. Con respecto al resto de métricas, la *Precision* destaca sobre las demás en el promedio no ponderado con un valor de 0,607 y en el ponderado con un valor de 0,605 el método RF. En lo referido a la métrica *Recall*, en el promedio no ponderado vuelve a destacar el valor obtenido para el método RF de 0,612 y en el ponderado también con un valor de 0,606. Y por último, para la métrica *F1-Score*, vuelve a sobresalir ante los demás el método RF con un valor de 0,601 en el promedio no ponderado y 0,597 en el promedio ponderado.

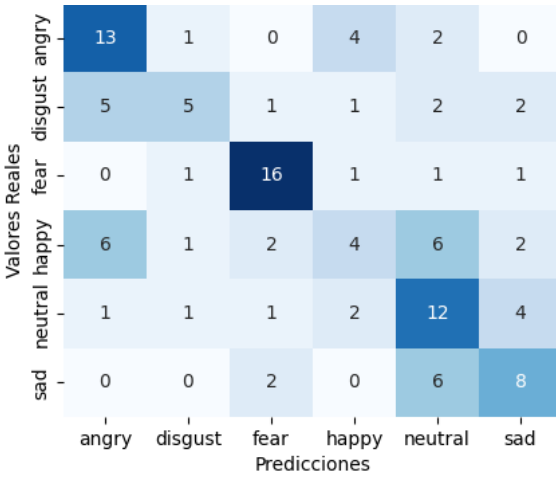
Asimismo, se muestran las matrices de confusión correspondiente a cada uno de los métodos aplicados en la Figura 5.1, donde la Figura 5.1a es la obtenida al aplicar el RF, para el LDA la Figura 5.1b y para el k-NN la Figura 5.1c. De esta forma, se muestran las clasificaciones correctas e incorrectas para cada una de las clases.

Al analizar las matrices de confusión, en la Figura 5.1 se puede observar que en las tres destacan las emociones de *angry* y *fear*, ya que cuentan con el mayor número de predicciones correctamente clasificadas. Para el método RF clasifica correctamente 13 muestras para la clase *angry* y 16 muestras para la clase *fear*. Ocurre lo mismo para el método LDA, donde clasifica correctamente 13 muestras para la clase *angry* y 16 muestras para la clase *fear*. Difiriendo se encuentra el método k-NN con un valor ligeramente inferior de 14 muestras clasificadas correctamente para la clase *fear* pero clasifica 13 muestras para la clase *angry*. Sin embargo los métodos LDA y k-NN coinciden para una tercera emoción con alto valor de predicción que es *neutral* con 12 y 16 muestras clasificadas correctamente, pero para el RF una tercera emoción a destacar es la de *sad* con 14 muestras clasificadas correctamente, no *neutral*. Estos resultados se reflejan en la *Precision* y el *Recall*, donde un menor número de falsos positivos contribuye a una alta Precisión, mientras que un número reducido de falsos negativos favorece un alto Recall.

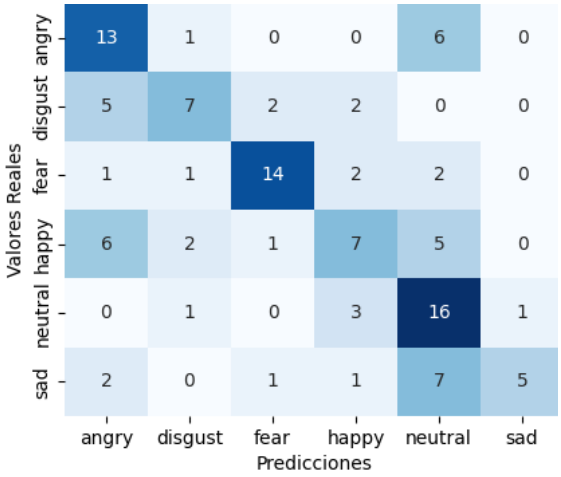
Por lo tanto, cuando se analiza un amplio rango de emociones usando una alta varia-



(a) Matriz de Confusión del RF para el experimento 1.



(b) Matriz de Confusión del LDA para el experimento 1.



(c) Matriz de Confusión del kNN para el experimento 1.

Figura 5.1: Matrices de Confusión para el experimento 1.

bilidad de características de la voz, el método con mayor rendimiento y mayores valores de predicción de emociones es el RF.

5.2. Conservación de todas las características y filtrado a 3 emociones.

Para el experimento 2, se han obtenido las métricas presentadas en la Tabla 5.2. De esta forma, se muestran las métricas medias para cada una de las clases.

En este experimento se ha reducido el número de emociones a 3 por lo que el número de muestras de test medio se reduce a 57 y, en lo referido a las métricas, se observa que ha afectado de manera positiva ya que han aumentado los valores para los tres métodos. En la Tabla 5.2 se aprecia que este experimento las métricas de mayor valor son para

Tabla 5.2: Resultados de las métricas obtenidas en el experimento 2 con un *Support* igual a 57 predicciones posibles para cada método. Las abreviaciones *Acc*, *Prec*, *Rec* y *F1* corresponden a las métricas: *Accuracy*, *Precision*, *Recall* y *F1-Score*.

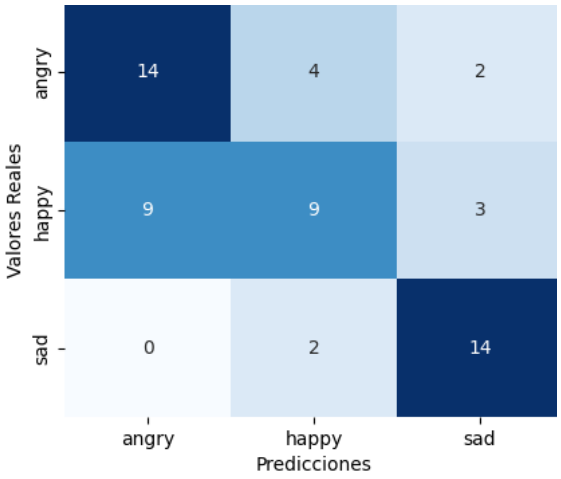
		Macro avg			*Weighted avg*		
Método	*Acc*	*Prec*	*Rec*	*F1*	*Prec*	*Rec*	*F1*
RF	0,647	0,648	0,664	0,647	0,643	0,647	0,635
LDA	0,666	0,705	0,669	0,679	0,688	0,666	0,670
kNN	0,596	0,631	0,596	0,602	0,616	0,596	0,595

el método LDA, seguidas por el RF y por último el k-NN. La *Accuracy* para el método LDA es de 0,666. Para el método RF obtiene un valor ligeramente inferior de *Accuracy* de 0,647. Y por último, para el método k-NN presenta una *Accuracy* de 0,596 con valor k igual 5 número de vecinos. Estos valores implican que el método que etiqueta mayor proporción de muestras como correctas es LDA, superando la mitad de muestras del total. Para el resto de métricas, para la *Precision* destaca sobre las demás tanto en el promedio no ponderado con un valor de 0,705 y en el ponderado con un valor de 0,688 el método LDA. Luego en lo referido a la métrica *Recall*, en el promedio no ponderado vuelve a destacar el valor obtenido para el método LDA de 0,669 y en el ponderado también con un valor de 0,666. Y por último, para la métrica *F1-Score*, vuelve a sobresalir ante los demás el método LDA con un valor de 0,679 en el promedio no ponderado y 0,670 en el promedio ponderado.

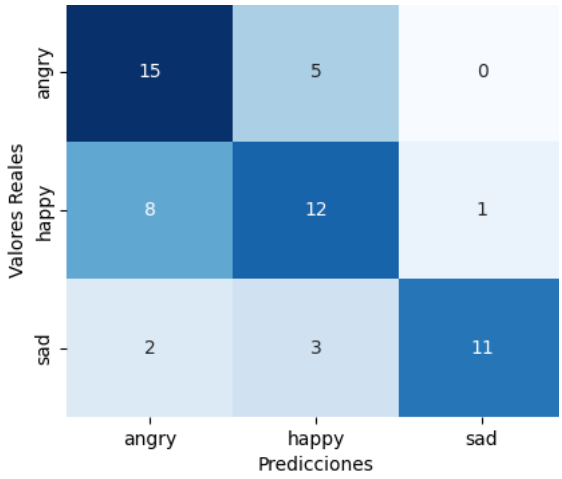
Asimismo, se muestran las matrices de confusión correspondiente a cada uno de los métodos aplicados en la Figura 5.2, donde la Figura 5.2a es la obtenida al aplicar el RF, para el LDA la Figura 5.2b y para el k-NN la Figura 5.2c. De esta forma, se muestran las clasificaciones correctas e incorrectas para cada una de las clases.

En las matrices de confusión, analizando la Figura 5.2 se puede observar que en las tres matrices destaca la clase emoción de *angry* con un mayor número de predicciones correctamente clasificadas. El método LDA clasifica correctamente 15 muestras para la clase *angry*. Difiriendo se encuentra el método RF con un valor ligeramente inferior de 14 muestras clasificadas correctamente para la clase *angry*. Sin embargo, ocurre lo mismo para el método k-NN que el de LDA, donde clasifica correctamente 15 muestras para la clase *angry*. Luego para RF destaca también la emoción *sad* con 14 muestras clasificadas correctamente y para LDA, la emoción de *happy* con 12 muestras clasificadas correctamente. Estos resultados se reflejan en la *Precision* y el *Recall*, donde un menor número de falsos positivos contribuye a una alta Precisión, mientras que un número reducido de falsos negativos favorece un alto Recall.

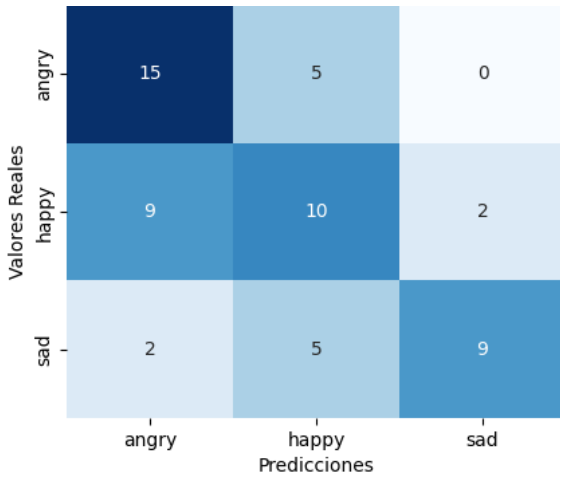
Por consiguiente, al haber reducido la gama de emociones y examinándola utilizando una gran diversidad de características vocales, para esta sección el método que destaca con mayor rendimiento y mayores valores de predicción de emociones es el LDA.



(a) Matriz de Confusión del RF para el experimento 2.



(b) Matriz de Confusión del LDA para el experimento 2.



(c) Matriz de Confusión del kNN para el experimento 2.

Figura 5.2: Matrices de Confusión para el experimento 2.

5.3. Omisión de la característica *Duration* de las demás y filtrado a 3 emociones.

En el experimento 3, se han obtenido las métricas presentadas en la Tabla 5.3. De esta forma, se muestran las métricas medias para cada una de las clases.

Tanto el anterior experimento como éste, como se ha mencionado antes, tienen los mismos aspectos solo que se omite la característica duración, por lo que se pasa a tener en cuenta solo 12. Al analizar los valores obtenidos, tanto en las métricas, como en las matrices de confusión, se observa que los valores no tiene una diferencia notable al quitar esta característica con los valores del experimento 2. Lo que sí difiere es que para este experimento vuelve a ser el método con mayor valor de *Acurracy* y del resto de métricas el RF. El RF cuenta con una *Acurracy* de 0,659. Por su lado el LDA cuenta con un valor

Tabla 5.3: Resultados de las métricas obtenidas en el experimento 3 con un *Support* igual a 57 predicciones posibles para cada método. Las abreviaciones *Acc*, *Prec*, *Rec* y *F1* corresponden a las métricas: *Accuracy*, *Precision*, *Recall* y *F1-Score*.

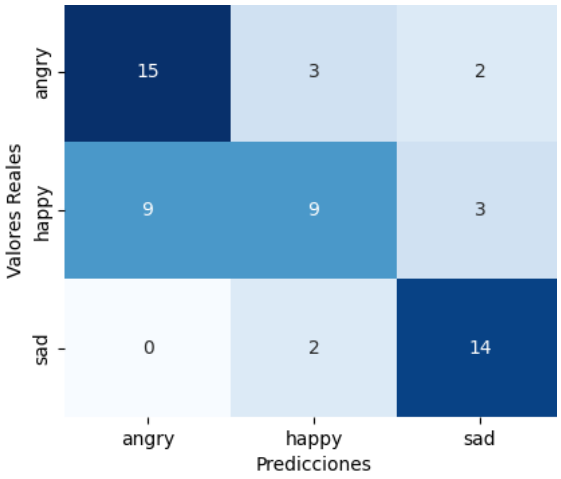
		Macro avg			*Weighted avg*		
Método	*Acc*	*Prec*	*Rec*	*F1*	*Prec*	*Rec*	*F1*
RF	0,659	0,663	0,677	0,657	0,659	0,659	0,646
LDA	0,649	0,690	0,652	0,665	0,671	0,649	0,654
kNN	0,578	0,589	0,575	0,574	0,583	0,578	0,573

de *Accuracy* de 0,649. Y el k-NN de 0,578 con valor de 3 número de vecinos para *k*. Estos valores implican que el método que etiqueta mayor proporción de muestras como correctas es RF, superando la mitad de muestras del total. Para el resto de métricas, la *Precision* destaca sobre las demás en el promedio no ponderado el método LDA con un valor de 0,690 y en el ponderado también, con un valor de 0,671. En lo referido a la métrica *Recall*, en el promedio no ponderado destaca el valor obtenido para el método RF de 0,677 y en el ponderado también con un valor de 0,659. Y por último, para la métrica *F1-Score*, vuelve a sobresalir ante los demás el método LDA con un valor de 0,655 en el promedio no ponderado y 0,654 en el promedio ponderado. Por tanto, se puede observar que es el primer experimento donde ocurre que a pesar de que el método RF es el que tiene una *Accuracy* mayor, no es el que más predomina en el resto de las métricas, ya que para estas, destaca más el método LDA.

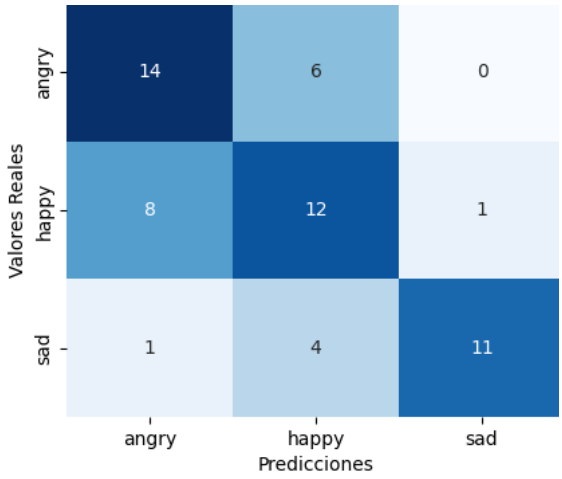
Asimismo, se muestran las matrices de confusión correspondiente a cada uno de los métodos aplicados en la Figura 5.3, donde la Figura 5.3a es la obtenida al aplicar el RF, para el LDA la Figura 5.3b y para el k-NN la Figura 5.3c. De esta forma, se muestran las métricas medias y las clasificaciones correctas e incorrectas para cada una de las clases.

En las matrices de confusión, analizando la Figura 5.3, se puede observar que en las tres vuelve a destacar la emoción de *angry* con un mayor número de predicciones correctamente clasificadas. El método RF clasifica correctamente 15 muestras para la clase *angry*. Difiriendo se encuentra el método LDA con un valor ligeramente inferior de 14 muestras clasificadas correctamente para la clase *angry*. Sin embargo, ocurre lo mismo para el método k-NN que para el RF, donde clasifica correctamente 15 muestras para la clase *angry*. Para RF destaca también la emoción *sad* con 14 muestras clasificadas correctamente y para LDA la emoción de *happy* con 12 muestras clasificadas correctamente. Estos resultados se reflejan en la *Precision* y el *Recall*, donde un menor número de falsos positivos contribuye a una alta Precisión, mientras que un número reducido de falsos negativos favorece un alto Recall.

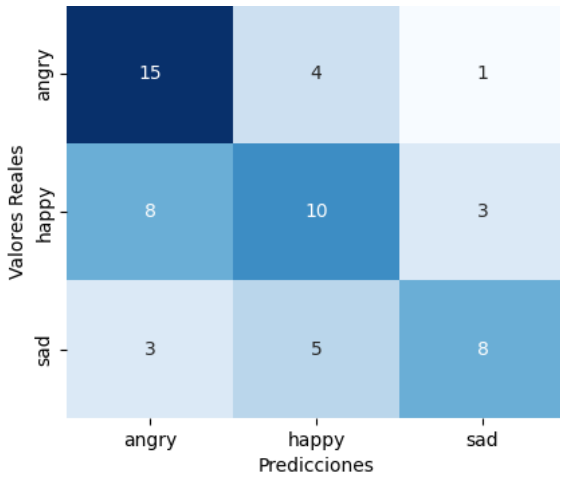
De este modo, al haber reducido la gama de emociones y evaluarla en función de los atributos vocales filtrando la característica *Duration*, para esta sección el método que destaca con mayor rendimiento y mayores valores de predicción de emociones es el RF a



(a) Matriz de Confusión del RF para el experimento 3.



(b) Matriz de Confusión del LDA para el experimento 3.



(c) Matriz de Confusión del kNN para el experimento 3.

Figura 5.3: Matrices de Confusión para el experimento 3.

pesar de lo comentado del LDA, al tener la *Accuracy* un mayor peso entre las métricas.

5.4. Características con valores de importancia entre 0,1 y menores a 0,15 manteniendo el filtrado a 3 emociones.

En los siguientes experimentos, 4, 5 y 6, se realiza la obtención de las características más importantes, donde se pueden observar en la Figura 5.4 los valores de las características más importantes de la base de datos utilizada.

En el experimento 4 se utilizan como entradas aquellas que tienen un valor de importancia en el intervalo con los valores entre 0,1 y menores a 0,15. Se han obtenido las

métricas presentadas en la Tabla 5.4. De esta forma, se muestran las métricas medias para cada una de las clases.

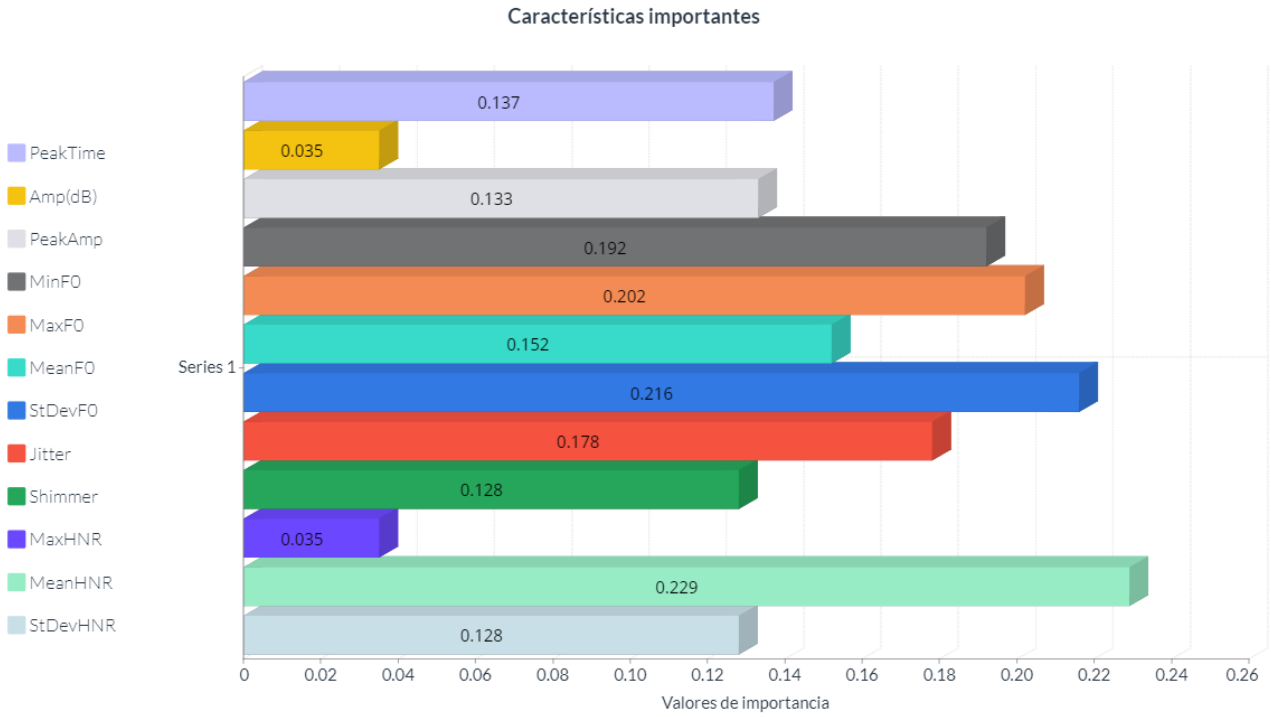
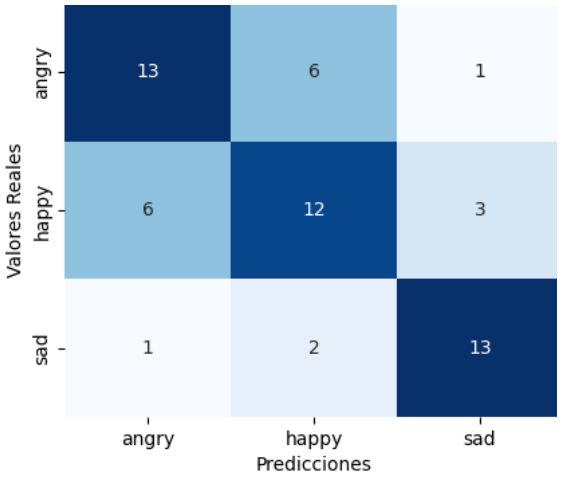


Figura 5.4: Resultado de aplicar el método *Mutual Info Classif* a las características de la Base de Datos utilizada.

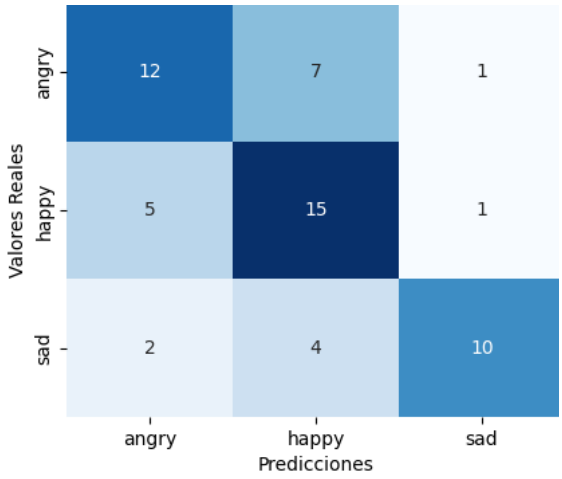
Tabla 5.4: Resultados de las métricas obtenidas en el experimento 4 con un *Support* igual a 57 predicciones posibles para cada método. Las abreviaciones *Acc*, *Prec*, *Rec* y *F1* corresponden a las métricas: *Accuracy*, *Precision*, *Recall* y *F1-Score*.

		Macro avg			*Weighted avg*		
Método	*Acc*	*Prec*	*Rec*	*F1*	*Prec*	*Rec*	*F1*
RF	0,666	0,669	0,680	0,672	0,662	0,666	0,662
LDA	0,649	0,680	0,646	0,655	0,668	0,649	0,651
kNN	0,649	0,697	0,643	0,655	0,681	0,649	0,652

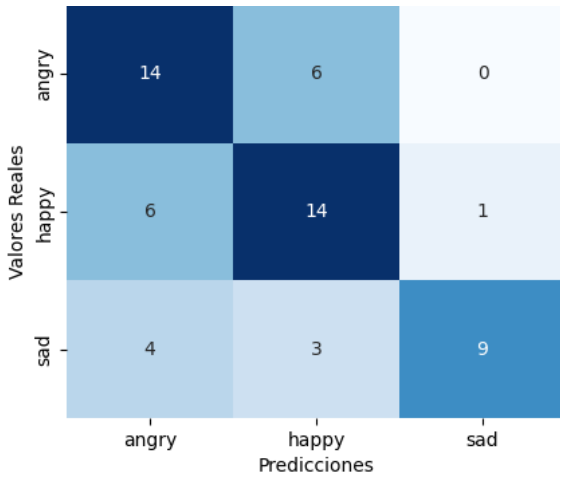
Analizando la Tabla 5.4, es destacable que en este experimento los métodos LDA y k-NN tienen el mismo valor para la *Acurracy* siendo 0,649 y con k para k-NN con valor de 10. Sin embargo destaca con una mayor el RF con 0,666. Estos valores implican que el método que etiqueta mayor proporción de muestras como correctas es RF, superando la mitad de muestras del total. También se puede observar que para el resto de métricas entre LDA y k-NN, por ejemplo, los promedios ponderados y no ponderados de *recall* y *f1-score* tiene ambas valores muy diferentes, sin embargo para *precision* son valores similares. Para la métrica *Precision* destaca sobre las demás el método k-NN en el promedio no ponderado con un valor de 0,697 y en el ponderado con un valor de 0,681. En lo referido a la métrica *Recall*, en el promedio no ponderado destaca el valor obtenido para el método RF de 0,680



(a) Matriz de Confusión del RF para el experimento 4.



(b) Matriz de Confusión del LDA para el experimento 4.



(c) Matriz de Confusión del kNN para el experimento 4.

Figura 5.5: Matrices de Confusión para el experimento 4.

y en el ponderado también con un valor de 0,666. Y por último, para la métrica *F1-Score*, vuelve a sobresalir ante los demás el método RF con un valor de 0,672 en el promedio no ponderado y 0,662 en el promedio ponderado.

Asimismo, se muestran las matrices de confusión correspondiente a cada uno de los métodos aplicados en la Figura 5.5, donde la Figura 5.5a es la obtenida al aplicar el RF, para el LDA la Figura 5.5b y para el k-NN la Figura 5.5c. De esta forma, se muestran las clasificaciones correctas e incorrectas para cada una de las clases.

Por otro lado, al analizar las matrices de confusión, en la Figura 5.1 se puede observar que esta vez en las tres destaca la emoción de *happy*, ya que cuenta con el mayor número de predicciones correctamente clasificadas. El método RF clasifica correctamente 13 muestras para la clase *angry* y la clase *sad*, y clasifica 12 muestras correctamente para la clase *happy*. El método LDA, clasifica correctamente 12 muestras para la clase *angry*, 15 muestras para la clase *happy* y 10 muestras para la clase *sad* clasificadas correctamente. Se observa en

la Figura 5.5c que el método k-NN cuenta con los valores máximos de predicción siendo de 14 muestras clasificadas correctamente para las clases emociones de *angry* y *happy*. Estos resultados se reflejan en la *Precision* y el *Recall*, donde un menor número de falsos positivos contribuye a una alta *Precision*, mientras que un número reducido de falsos negativos favorece un alto *Recall*.

En consecuencia, al reducir la serie de emociones y usar el intervalo correspondiente de cualidades de la voz más importantes, vuelve a ser el método más eficaz el RF para este experimento al tener la mayor *Accuracy*.

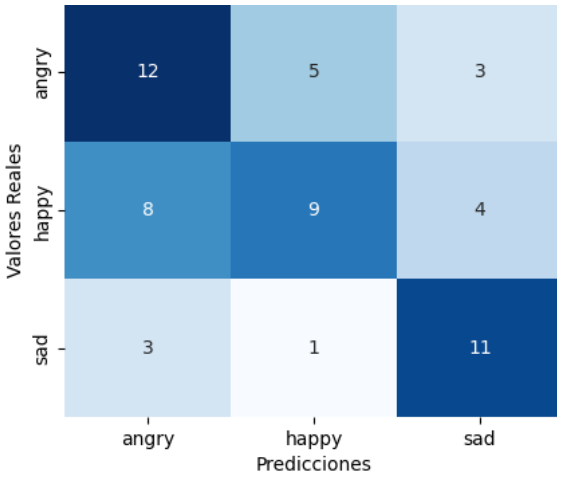
5.5. Características con valores de importancia entre 0,15 y menores a 0,2 manteniendo el filtrado a 3 emociones.

En el experimento 5, en el cual se han obtenido las características más importantes con valores de importancia entre 0,15 y 0,2. El valor de importancia de cada una de las características se muestra en la Figura 5.4. En este experimento se han obtenido las métricas presentadas en la Tabla 5.5. De esta forma, se muestran las métricas medias para cada una de las clases.

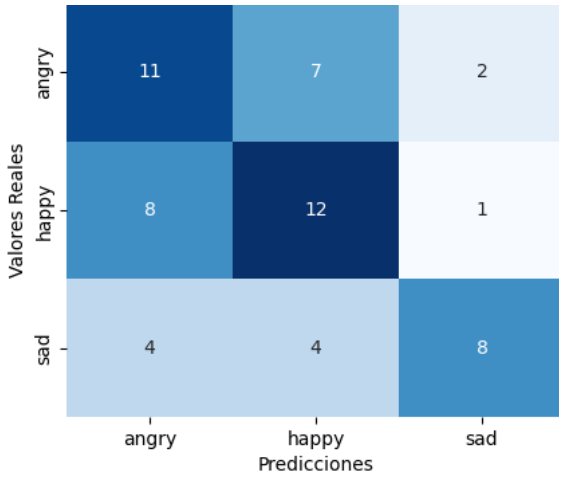
Tabla 5.5: Resultados de las métricas obtenidas en el experimento 5 con un *Support* igual a 57 predicciones posibles para cada método. Las abreviaciones *Acc*, *Prec*, *Rec* y *F1* corresponden a las métricas: *Accuracy*, *Precision*, *Recall* y *F1-Score*.

		Macro avg			*Weighted avg*		
Método	*Acc*	*Prec*	*Rec*	*F1*	*Prec*	*Rec*	*F1*
RF	0,571	0,578	0,583	0,573	0,576	0,571	0,566
LDA	0,543	0,575	0,540	0,549	0,564	0,543	0,546
kNN	0,578	0,595	0,578	0,580	0,597	0,578	0,580

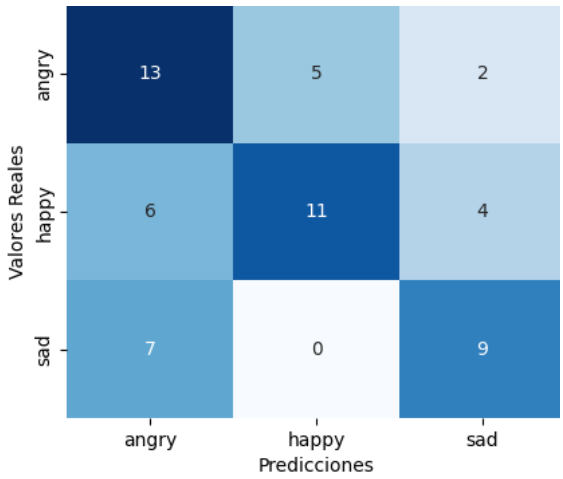
Se puede observar en la Tabla 5.5 que al aplicar el segundo intervalo a este experimento, los valores de las métricas se reducen bastante en comparación con los del experimento 4 del primer intervalo, ya que ahora la mayor *Acurracy* es de 0,578 perteneciente al método k-NN con valor k igual a 5. El siguiente mejor método es el RF con una *Accuracy* de 0,571 y, el menos exacto es LDA con la *Accuracy* de 0,543. Estos valores implican que el método que etiqueta mayor proporción de muestras como correctas es k-NN, superando la mitad de muestras del total. Para el resto de métricas, la *Precision* destaca sobre las demás tanto en el promedio no ponderado con un valor de 0,595 como en el ponderado con un valor de 0,597 el método k-NN. En lo referido a la métrica *Recall*, en el promedio no ponderado destaca el valor obtenido para el método RF de 0,583 pero en el ponderado



(a) Matriz de Confusión del RF para el experimento 5.



(b) Matriz de Confusión del LDA para el experimento 5.



(c) Matriz de Confusión del kNN para el experimento 5.

Figura 5.6: Matrices de Confusión para el experimento 5.

el k-NN con un valor de 0,578. Y por último, la métrica *F1-Score*, esta vez destaca en los dos promedios el k-NN con un valor de 0,580 en el promedio no ponderado y lo mismo para el promedio ponderado.

Asimismo, se muestran las matrices de confusión correspondiente a cada uno de los métodos aplicados en la Figura 5.6, donde la Figura 5.6a es la obtenida al aplicar el RF, para el LDA la Figura 5.6b y para el k-NN la Figura 5.6c. De esta forma, se muestran las métricas medias y las clasificaciones correctas e incorrectas para cada una de las clases.

Analizando la Figura 5.6a, se observa que el método RF clasifica correctamente 12 muestras de la emoción *angry* y 11 muestras para la emoción *sad*. Después, en la Figura 5.6b, el LDA clasifica correctamente 11 muestras la emoción *angry* y 12 muestras la emoción *happy*. Y por otra parte, en la Figura 5.6c, el método k-NN clasifica correctamente 13 muestras para la clase *angry* y 11 muestras para la clase *happy*. Vuelve a predominar la clase *angry* para este experimento.

Por lo tanto, al reducir el extenso espectro de emociones empleando el intervalo correspondiente de rasgos de la voz más importantes, para este experimento el método más factible es el k-NN.

5.6. Características con valores de importancia iguales y mayores a 0,2 manteniendo el filtrado a 3 emociones.

Para el experimento 6, se obtienen las características más importantes como aquellas con valores de importancia iguales y mayores a 0,2. El valor de importancia de cada característica se muestra en la Figura 5.4.

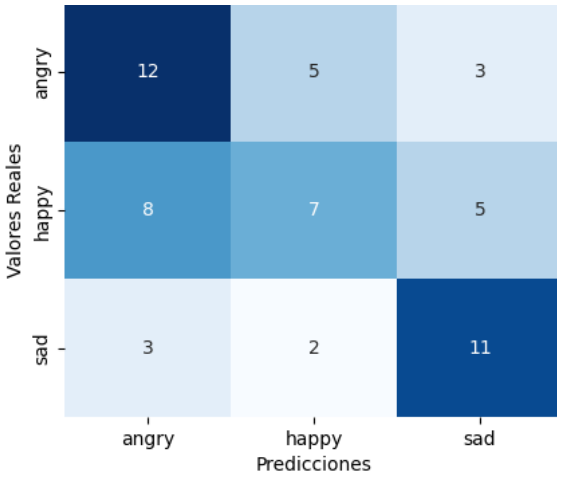
En este experimento se han obtenido las métricas presentadas en la Tabla 5.6. De esta forma, se muestran las métricas medias para cada una de las clases.

Tabla 5.6: Resultados de las métricas obtenidas en el experimento 6 con un *Support* igual a 57 predicciones posibles para cada método. Las abreviaciones *Acc*, *Prec*, *Rec* y *F1* corresponden a las métricas: *Accuracy*, *Precision*, *Recall* y *F1-Score*.

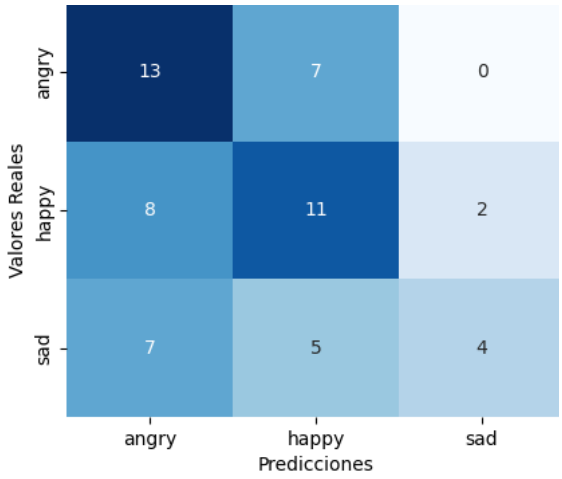
		Macro avg			*Weighted avg*		
Método	*Acc*	*Prec*	*Rec*	*F1*	*Prec*	*Rec*	*F1*
RF	0,524	0,524	0,538	0,522	0,521	0,524	0,514
LDA	0,491	0,536	0,474	0,468	0,526	0,491	0,476
kNN	0,526	0,532	0,525	0,526	0,529	0,526	0,525

Como se ha discutido en el experimento 5, se vuelve a apreciar una *Accuracy* y el resto de métricas mas bajas, donde es el k-NN el que tiene una mayor con valor de 0,526, usando el valor de 10 como número de vecinos para el parámetro k. Este valor no difiere por mucho el valor de 0,524 perteneciente al RF. Y por último el LDA con una *Accuracy* de 0,491. Estos valores implican que el método que etiqueta mayor proporción de muestras como correctas es k-NN, superando la mitad de muestras del total. En el resto de métricas, para la *Precision* destaca en el promedio no ponderado con un valor de 0,536 el método LDA y en el ponderado con un valor de 0,529 el método k-NN. Luego en lo referido a la métrica *Recall*, en el promedio no ponderado sobresale el valor obtenido para el método RF de 0,538 y en el ponderado con n valor de 0,526 el k-NN. Y por último, para la métrica *F1-Score*, sobresale ante los demás el método k-NN con un valor de 0,526 en el promedio no ponderado y 0,525 en el promedio ponderado.

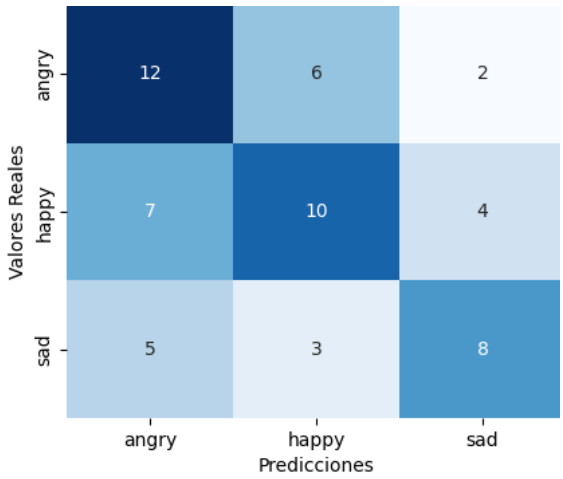
Asimismo, se muestran las matrices de confusión correspondiente a cada uno de los métodos aplicados en la Figura 5.7, donde la Figura 5.7a es la obtenida al aplicar el RF, para el LDA la Figura 5.7b y para el k-NN la Figura 5.7c. De esta forma, se muestran las clasificaciones correctas e incorrectas para cada una de las clases.



(a) Matriz de Confusión del RF para el experimento 6.



(b) Matriz de Confusión del LDA para el experimento 6.



(c) Matriz de Confusión del kNN para el experimento 6.

Figura 5.7: Matrices de Confusión para el experimento 6.

En la Figura 5.7a, se observa que el método RF clasifica correctamente 12 muestras la emoción *angry* y 11 muestras la emoción *sad*. En la Figura 5.7b, el LDA clasifica correctamente 13 muestras la emoción *angry* y 11 muestras la emoción *happy*. Y por otra parte, en la Figura 5.7c, el método k-NN clasifica correctamente 12 muestras para la clase *angry* y 10 muestras para la clase *happy*. Vuelve a predominar la clase *angry* para este experimento también con los valores de predicción más altos.

Por ende, al examinar la gama de emociones reducida utilizando el intervalo correspondiente de las características vocales más importantes, es el método k-NN el más adecuado para este experimento al poseer una mayor *Accuracy*.

5.7. Comparativa entre los experimentos realizados.

La Tabla 5.7 incluye de forma resumida los métodos mas eficientes por cada experimento para esta sección según los resultados obtenidos.

Tabla 5.7: Selección del mejor método para cada experimento según las métricas obtenidas con un *Support* igual a 114 predicciones posibles para el experimento 1 y de 57 para el resto de experimentos. Las abreviaciones *Acc*, *Prec*, *Rec* y *F1* corresponden a las métricas: *Accuracy*, *Precision*, *Recall* y *F1-Score*.

			Macro avg			*Weighted avg*		
Experimento	Método	*Acc*	*Prec*	*Rec*	*F1*	*Prec*	*Rec*	*F1*
1	RF	0,606	0,607	0,612	0,601	0,605	0,606	0,597
2	LDA	**0,666**	**0,705**	0,669	**0,679**	**0,688**	**0,666**	**0,670**
3	RF	0,659	0,663	0,677	0,657	0,659	0,659	0,646
4	RF	**0,666**	0,669	**0,680**	0,672	0,662	**0,666**	0,662
5	kNN	0,578	0,595	0,578	0,580	0,597	0,578	0,580
6	kNN	0,526	0,532	0,525	0,526	0,529	0,526	0,525

Una vez plasmados todos los resultados en la Tabla 5.7 y compararlos, se llega a la conclusión de que la *Accuracy* tiene un valor mayor cuando se reducen el número de emociones y cuando el valor de las características es menor. Sin embargo, ocurre que, tanto el método LDA explicado en el experimento 2, como el método RF en su caso para el experimento 4, tienen la misma *Accuracy* como se puede observar tanto en la Tabla 5.7 con un valor de 0,666. Fijándonos mas detalladamente en el resto de métricas obtenidas, para la métrica *Precision* destaca sobre las demás tanto en el promedio no ponderado con un valor de 0,705 y en el ponderado con un valor de 0,688 el método LDA para el experimento 2. En lo referido a la métrica *Recall*, en el promedio no ponderado destaca el valor obtenido en el experimento 4 para el método RF con un valor de 0,680 y en el ponderado hay un empate con valor de 0,666 tanto para el LDA del experimento 2 como para el RF del experimento 4. Y por último, para la métrica *F1-Score*, vuelve a sobresalir ante los demás el método LDA del experimento 2 con un valor de 0,679 en el promedio no ponderado y 0,670 en el promedio ponderado. Por lo que a la hora de elegir un método de los 6 experimentos realizados y adaptarlo a la propuesta planteada en este TFG, sería el LDA del experimento 2 el método más adecuado. Y aunque tenga el mismo valor de *Accuracy* que el método RF del experimento 4, siendo encima este el más factible para la mayoría de los experimentos, el LDA del experimento 2 destaca por presentar un mayor número de ocasiones el ser el método con mayor valores en las métricas de la Tabla 5.7 donde este factor tiene más peso a la hora de la elección. Por lo tanto, la reducción de características planteado en este TFG no es efectivo con respecto al uso del total de características de la voz. Además, la información de la duración es relevante para la detección de las características.

Sin embargo, en el experimento 2 se analizan únicamente 3 emociones, mientras que

en el experimento 1 se analizan en espectro completo de emociones incluido en este TFG. Por tanto, puede observarse que el RF es mejor cuando se analiza un mayor número de emociones y, cuando el número es menor, es el LDA el algoritmo más adecuado.

Asimismo, también enfatizar otro aspecto peculiar y es que en todos los experimentos, siempre destaca la clase de emoción *angry* como una de las emociones con mayores niveles de predicción y esto puede ser debido a que las personas tienden a hablar más alto y con un tono más agudo cuando están enojadas. Los algoritmos de procesamiento de señales de audio pueden detectar estos patrones con mayor facilidad.

Capítulo 6

Conclusiones y Líneas Futuras

En este capítulo se comentan las principales conclusiones extraídas del estudio, así como las limitaciones, líneas futuras e impacto social. En la Sección 6 se detallan los principales hallazgos del TFG. En la Sección 6 las limitaciones y futuras líneas de investigación. La Sección 6 resume las consecuencias sociales, medioambientales y económicas de este TFG. Por último, en la Sección 6.3.3 se describen las lecciones aprendidas en el desarrollo de este TFG.

6.1. Conclusiones

Durante la realización de este TFG se realiza un estudio de la clasificación de emociones en grabaciones de voz mediante técnicas de ML donde el objetivo es identificar y categorizar automáticamente las emociones expresadas en el habla humana, la cual se ha convertido en una herramienta esencial en diversos ámbitos, por ejemplo, en el ámbito de la salud, facilita el diagnóstico y seguimiento de trastornos emocionales. Este enfoque tiene el potencial para abrir nuevas posibilidades para comprender y mejorar las relaciones humanas a través del análisis de la voz.

En relación con los modelos utilizados para realizar el entrenamiento de la tarea de clasificación, destacan modelos como el RF y el LDA que demuestran un rendimiento destacado y el k-NN también pero algo más reducido. Además, se demuestran que el uso de técnicas como la eliminación de celdas NaN, la estandarización de datos y la obtención de características más importantes son muy útiles para mejorar el rendimiento de los modelos y, por tanto, mejorar los resultados obtenidos.

Asimismo, se estudian seis tipos de experimentos durante la aplicación de los métodos de aprendizaje supervisado: se mantienen todas las características y emociones, mantener todas las características y filtrar por 3 emociones, omisión de la característica *Duration* de las demás características y filtrar por las 3 emociones, obtener las características más

importantes con valores entre 0,1 y menores a 0,15, obtener las características más importantes con valores entre 0,15 y menores a 0,2 y obtener las características más importantes con valores iguales y mayores a 0,2. Se observa que el mantener todas las características y filtrar por 3 emociones, perteneciente al experimento 2, proporciona los mejores resultados en la clasificación de emociones, siendo el LDA el método con mayor rendimiento y mayores valores de predicción de emociones. Por otro lado, el *Accuracy* cuando se analizan todas las emociones está también por encima del 50 %, siendo la característica *angry* la que mejor se detecta.

En conclusión, los resultados obtenidos en este TFG muestran que es posible realizar la clasificación de emociones en grabaciones de voz con técnicas de aprendizaje supervisado. Además se demuestra que estas técnicas son efectivas y permiten obtener resultados precisos.

6.2. Limitaciones y líneas futuras

En este estudio se han encontrado algunas limitaciones que deben tenerse en cuenta al analizar los resultados alcanzados. Estas limitaciones tienen implicaciones en la generalización de las conclusiones obtenidas.

En primer lugar, se ha utilizado una base de datos reducida con respecto a la original. La base de datos original consta de cinco corpus de voz formada por frases, palabras y pseudo-palabras, donde la elegida para este TFG está enfocada en las palabras y pseudo-palabras. Esto se debe a que la original es una base de datos muy extensa y habría que abarcar más métodos, más emociones y más características para obtener resultados más precisos.

Otra limitación es el número de parámetros acústicos. Aunque se analizaron trece, cabe la posibilidad de que otros parámetros no considerados también influyan significativamente en el reconocimiento de emociones.

También la diversidad de participantes, ya que al no ser muy variada en términos de edad, cultura o acentos, la percepción y el reconocimiento de emociones pueden variar significativamente. Sino que se limitó a un grupo de jóvenes estudiantes de Alemania.

Luego, como futuras investigaciones beneficiaría el uso de una base de datos más extensa, que aborde tanto palabras, pseudo-palabras y frases. Además, se podrían analizar datos que cuenten con una cultura más diversa, es decir, el uso de más idiomas y no solo de uno. Otra posibilidad sería ampliar el conjunto de parámetros acústicos, por ejemplo, en lo referido a características prosódicas, que se tenga en cuenta el *ritmo* o la *enfasis y acentuación* de palabras. En lo referido a características temporales, añadir la *tase de cruce por cero* que puede proporcionar información sobre la agresividad en la voz. También

en lo referido a las características relacionadas con la calidad de voz, destacar la *tensión vocal*. Y por último, se podrían analizar otros algoritmos de ML y *Deep Learning* (DL) como el aprendizaje no supervisado.

6.3. Impacto

A continuación se describe el impacto que tiene este TFG en el ámbito social, medioambiental y económico.

6.3.1. Impacto Social

La clasificación de emociones mediante grabaciones de voz tiene potencial de impacto positivo en varios aspectos para la sociedad [48]. Puede abarcar desde la salud mental y el bienestar, hasta la educación, también en lo tecnológico, entretenimiento, etc.

Algunos de estos ejemplos son la mejoría de la detección de trastornos mentales con las voces de los clientes, así como para personas con discapacidades con trastornos del espectro autista [21]. En servicios de Atención al cliente, las empresas pueden utilizar esta clasificación de emociones para evaluar la satisfacción del cliente en tiempo real. En el entretenimiento y publicidad, la clasificación de las emociones que produce el usuario a un contenido pueden ayudar a mejorar la experiencia de este.

6.3.2. Impacto Medioambiental

El desarrollo y despliegue de sistemas de clasificación de emociones en gran escala podría requerir una considerable cantidad de recursos computacionales, lo que podría aumentar el consumo de energía y la huella de carbono asociada con el funcionamiento de servidores y centros de datos [49]. Por ejemplo, el tamaño de la base de datos utilizada, donde sí influye tanto directa, como indirectamente a través de los procesos de recolección, almacenamiento, procesamiento y transmisión de datos debido al consumo indebido de energía y desperdicio de recursos. Además, la fabricación y eliminación de dispositivos electrónicos necesarios para implementar estas tecnologías también podría contribuir a la contaminación ambiental.

Sin embargo, fomenta la optimización de la atención al cliente y servicios automatizados, ya que el uso de clasificación de emociones puede ayudar a dirigir eficientemente las solicitudes de los usuarios y resolver problemas de manera más rápida y efectiva. Esto puede reducir el tiempo de llamada promedio y, por lo tanto, el consumo de energía asociado con los centros de datos y la infraestructura de telecomunicaciones. Asimismo, también fomenta la optimización de la publicidad digital y los servicios de recomendación ya que al comprender mejor las emociones y preferencias de los usuarios, los sistemas

de publicidad digital y recomendación pueden mostrar anuncios más relevantes y sugerir contenido que tenga más probabilidades de ser apreciado por los usuarios. Esto puede reducir la necesidad de publicidad no deseada y promover un consumo más consciente, lo que a su vez puede tener un impacto positivo en la eficiencia energética asociada con la entrega de contenido digital.

6.3.3. Impacto Económico

La clasificación de emociones en grabaciones de voz tiene un impacto económico significativo en la sociedad [50]. Por ejemplo, esta tecnología permitiría a las empresas comprender mejor las necesidades y deseos de los clientes, lo que produciría una mejora de la experiencia del cliente y, en última instancia, en un aumento de las ventas y la fidelidad de los clientes. Además, otro ejemplo seria en campos como la publicidad y el marketing, ayuda a personalizar los mensajes para adaptarse al estado emocional del consumidor, lo que aumenta la efectividad de las campañas.

En relación con el impacto social, medioambiental y económico, es importante destacar como este TFG se enmarca dentro de los Objetivos de Desarrollo Sostenible (ODS). Los ODS son una serie de 17 objetivos globales establecidos en 2015 por la Organización de las Naciones Unidas (ONU) como parte de su agenda 2030 para el desarrollo sostenible [51]. Estos objetivos tienen como finalidad abordar los desafíos mundiales más urgentes tratando de promover un mundo más justo, equitativo y sostenible para todos.

Este TFG podría adecuarse en el ODS 3 de *Salud y bienestar*, ya que este método de clasificación podría ayudar a monitorizar el estado emocional de personas que afrontan problemas de salud mental y así poder detectarlos de manera temprana para contribuir a un mejor bienestar emocional y psicológico. Otro en el que podría basarse sería el ODS 8 de *Trabajo decente y crecimiento económico* debido a que la personalización de la publicidad y el marketing basada en emociones puede mejorar la eficacia de las campañas comerciales. Asimismo, el ODS 9 *Industria, innovación e infraestructura* también valdría porque la aplicación de tecnologías de ML en la clasificación de emociones promueve la innovación tecnológica y la mejora de infraestructuras digitales, lo que puede beneficiar a diversos sectores industriales. Y por último, también el ODS 11 *Ciudades y comunidades sostenibles* en el ámbito de la seguridad y el bienestar de los ciudadanos en entornos urbanos, ya que al tratar de clasificar las emociones en grabaciones de voz, las autoridades podrían monitorizar el estado emocional de las personas en tiempo real, lo que podría ayudar a detectar situaciones de emergencia, como incidentes de violencia o altercados, y a responder de manera más rápida y eficaz.

En la Tabla 6.1 se recogen aquellos objetivos y metas a los que aporta un beneficio el estudio llevado a cabo en esta memoria.

Tabla 6.1: Metas específicas de los ODS que se adecuan a este TFG

Objetivo	Metas	Descripción
ODS 3: Salud y bienestar	ODS 3.4	Para 2030, reducir en un tercio la mortalidad prematura por enfermedades no transmisibles mediante la prevención y el tratamiento y promover la salud mental y el bienestar.
	ODS 3.c	Aumentar sustancialmente la financiación de la salud y la contratación, el desarrollo, la capacitación y la retención del personal sanitario en los países en desarrollo, especialmente en los países menos adelantados y los pequeños Estados insulares en desarrollo.
	ODS 3.d	Reforzar la capacidad de todos los países, en particular los países en desarrollo, en materia de alerta temprana, reducción de riesgos y gestión de los riesgos para la salud nacional y mundial.
ODS 8: Trabajo decente y crecimiento económico	ODS 8.2	Lograr niveles más elevados de productividad económica mediante la diversificación, la modernización tecnológica y la innovación, entre otras cosas centrándose en los sectores con gran valor añadido y un uso intensivo de la mano de obra.
	ODS 8.10	Fortalecer la capacidad de las instituciones financieras nacionales para fomentar y ampliar el acceso a los servicios bancarios, financieros y de seguros para todos.
	ODS 8.b	De aquí a 2020, desarrollar y poner en marcha una estrategia mundial para el empleo de los jóvenes y aplicar el Pacto Mundial para el Empleo de la Organización Internacional del Trabajo.
ODS 9: Industria, innovación e infraestructura	ODS 9.1	Desarrollar infraestructuras fiables, sostenibles, resilientes y de calidad, incluidas infraestructuras regionales y transfronterizas, para apoyar el desarrollo económico y el bienestar humano, haciendo especial hincapié en el acceso asequible y equitativo para todos.
	ODS 9.4	De aquí a 2030, modernizar la infraestructura y reconvertir las industrias para que sean sostenibles, utilizando los recursos con mayor eficacia y promoviendo la adopción de tecnologías y procesos industriales limpios y ambientalmente racionales, y logrando que todos los países tomen medidas de acuerdo con sus capacidades respectivas.
	ODS 9.5	Aumentar la investigación científica y mejorar la capacidad tecnológica de los sectores industriales de todos los países, en particular los países en desarrollo, entre otras cosas fomentando la innovación y aumentando considerablemente, de aquí a 2030, el número de personas que trabajan en investigación y desarrollo por millón de habitantes y los gastos de los sectores público y privado en investigación y desarrollo.
	ODS 9.c	Aumentar significativamente el acceso a la tecnología de la información y las comunicaciones y esforzarse por proporcionar acceso universal y asequible a Internet en los países menos adelantados de aquí a 2020.
ODS 11: Ciudades y comunidades sostenibles	ODS 11.2	De aquí a 2030, proporcionar acceso a sistemas de transporte seguros, asequibles, accesibles y sostenibles para todos y mejorar la seguridad vial, en particular mediante la ampliación del transporte público, prestando especial atención a las necesidades de las personas en situación de vulnerabilidad, las mujeres, los niños, las personas con discapacidad y las personas de edad.
	ODS 11.7	De aquí a 2030, proporcionar acceso universal a zonas verdes y espacios públicos seguros, inclusivos y accesibles, en particular para las mujeres y los niños, las personas de edad y las personas con discapacidad.
	ODS 11.a	Apoyar los vínculos económicos, sociales y ambientales positivos entre las zonas urbanas, periurbanas y rurales fortaleciendo la planificación del desarrollo nacional y regional.

6.4. Lecciones Aprendidas

Durante el desarrollo de este TFG se han ampliado los conocimientos adquiridos en la asignatura de Tratamiento Digital de Sonido. En primer lugar, se ha profundizado en el estudio de la clasificación de emociones y su análisis. Mediante la utilización de ML, se han explorado la clasificación de los tipos de algoritmos de aprendizaje supervisado así como sus métodos, lo que ha permitido comprender en mayor profundidad esta actividad. Además, se han aplicado las nociones de programación adquiridas a lo largo de la carrera, donde lo recomendable era mediante Matlab pero se decide realizarlo en Python. También se han adquirido conocimientos sobre diversas técnicas como por ejemplo la estandarización de datos o la obtención de las características más importantes, que han sido fundamentales para mejorar el rendimiento de los modelos utilizados. Por último, se ha aprendido a redactar documentos científicos empleando LaTeX, comprendiendo la estructura y la organización propias de este tipo de documentos.

Bibliografía

[1] Georgia Chronaki, Michael Wigelsworth, Marc D Pell, and Sonja A Kotz. The development of cross-cultural recognition of vocal emotion during childhood and adolescence. *Scientific reports*, 8(1):8659, 2018.

[2] Patrik N Juslin, Klaus R Scherer, and J Harrigan. Vocal expression of affect. *The new handbook of methods in nonverbal behavior research*, pages 65–135, 2005.

[3] Silke Paulmann. The neurocognition of prosody. In *Neurobiology of language*, pages 1109–1120. Elsevier, 2016.

[4] S Kitayama. Word and voice: Spontaneous attention to emotional speech in two cultures. *Cognition and Emotion*, 16:29–59, 2002.

[5] Alan S Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature human behaviour*, 3(4):369–382, 2019.

[6] Alan S Cowen, Hillary Anger Elfenbein, Petri Laukka, and Dacher Keltner. Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*, 74(6):698, 2019.

[7] Adi Lausen and Annekathrin Schacht. Gender differences in the recognition of vocal emotions. *Frontiers in psychology*, 9:359771, 2018.

[8] Daniel T Cordaro, Dacher Keltner, Sumjay Tshering, Dorji Wangchuk, and Lisa M Flynn. The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion*, 16(1):117, 2016.

[9] Rebecca Jürgens, Matthis Drolet, Ralph Pirow, Elisabeth Scheiner, and Julia Fischer. Encoding conditions affect recognition of vocally expressed emotions across cultures. *Frontiers in psychology*, 4:111, 2013.

[10] Marc D Pell, Silke Paulmann, Chinar Dara, Areej Alasseri, and Sonja A Kotz. Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics*, 37(4):417–435, 2009.

[11] Klaus R Scherer, Rainer Banse, and Harald G Wallbott. Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-cultural psychology*, 32(1):76–92, 2001.

[12] Indrit Bègue, Maarten Vaessen, Jeremy Hofmeister, Marice Pereira, Sophie Schwartz, and Patrik Vuilleumier. Confidence of emotion expression recognition recruits brain regions outside the face perception network. *Social cognitive and affective neuroscience*, 14(1):81–95, 2019.

[13] Karen J Kelly and Janet Metcalfe. Metacognition of emotional face recognition. *Emotion*, 11(4):896, 2011.

[14] Elizabeth F Chua, Daniel L Schacter, and Reisa A Sperling. Neural correlates of metamemory: a comparison of feeling-of-knowing and retrospective confidence judgments. *Journal of Cognitive Neuroscience*, 21(9):1751–1765, 2009.

[15] Martijn Goudbeek and Klaus Scherer. 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, 2010.

[16] Patrik N Juslin and Petri Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin*, 129(5):770, 2003.

[17] Tom Johnstone and Klaus R Scherer. Vocal communication of emotion. *Handbook of emotions*, 2:220–235, 2000.

[18] Adi Lausen and Kurt Hammerschmidt. Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanities and Social Sciences Communications*, 7(1):1–17, 2020.

[19] Kurt Hammerschmidt and Uwe Jürgens. Acoustical correlates of affective prosody. *Journal of voice*, 21(5):531–540, 2007.

[20] Beate Wendt and Henning Scheich. The"Magdeburger Prosodie-Korpus

. In *Speech Prosody 2002, International Conference*, 2002.

[21] Amirreza Rouhi, Micol Spitale, Fabio Catania, Giulia Cosentino, Mirko Gelsomini, and Franca Garzotto. Emotify: emotional game for children with autism spectrum disorder based-on machine learning. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*, pages 31–32, 2019.

[22] Xin Xu, Yiwei Zhang, Minghong Tang, Hong Gu, Shancheng Yan, and Jie Yang. Emotion recognition based on double tree complex wavelet transform and machine learning in internet of things. *IEEE Access*, 7:154114–154120, 2019.

[23] Embla CS Neverlien, Rose Lu, Mohit Kumar, and Marta Molinas. Decoding Emotions From EEG Responses Elicited by Videos Using Machine Learning Techniques on Two Datasets. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE, 2023.

[24] Bei Fang, Xian Li, Guangxin Han, and Juhou He. Facial expression recognition in educational research from the perspective of machine learning: A systematic review. *IEEE Access*, 2023.

[25] John Henry, Huw Lloyd, Martin Turner, and Connah Kendrick. On the robustness of machine learning models for stress and anxiety recognition from heart activity signals. *IEEE Sensors Journal*, 2023.

[26] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

[27] Michael J Owren. GSU Praat Tools: Scripts for modifying and analyzing sounds using Praat acoustics software. *Behavior research methods*, 40(3):822–829, 2008.

[28] Introducing the Oxford Vocal (OxVoc) Sounds database: a validated set of non-acted affective sounds from human infants, adults, and domestic animals, author=Parsons, Christine E and Young, Katherine S and Craske, Michelle G and Stein, Alan L and Kringelbach, Morten L. *Frontiers in psychology*, 5:562, 2014.

[29] Mustafa A Qamhan, Hamdi Altaheri, Ali Hamid Meftah, Ghulam Muhammad, and Yousef Ajami Alotaibi. Digital audio forensics: microphone and environment classification using deep learning. *Ieee Access*, 9:62719–62733, 2021.

[30] Francesco Camastra and Alessandro Vinciarelli. *Machine learning for audio, image and video analysis: theory and applications*. Springer, 2015.

[31] Vladimir Nasteski. An overview of the supervised machine learning methods. *Horizons. b*, 4:51–62, 2017.

[32] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[33] Fatemeh Noroozi, Tomasz Sapiński, Dorota Kamińska, and Gholamreza Anbarjafari. Vocal-based emotion recognition using random forests and decision tree. *International Journal of Speech Technology*, 20(2):239–246, 2017.

[34] Kurt Hammerschmidt Adi Lausen. Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanities and Social Sciences Communications volume 7, Article number: 2*, 2020.

[35] Fadwa Al-Azab, Bijan Raahemi, Gregory Richards, Natalia Jaworska, Dylan Smith, Sara de la Salle, Pierre Blier, and Verner Knott. Data mining eeg signals in depression for their diagnostic value. *BMC medical informatics and decision making*, 15:108, 12 2015.

[36] Oliver Kramer and Oliver Kramer. K-nearest neighbors. *Dimensionality reduction with unsupervised nearest neighbors*, pages 13–23, 2013.

[37] Aurélien Géron. Aprende machine learning con scikit-learn, keras y tensorflow. *España: Anaya*, 2020.

[38] Pranav Shetty and Suraj Singh. Hierarchical clustering: a survey. *International Journal of Applied Research*, 7(4):178–181, 2021.

[39] Federico Miyara. La voz humana. *Laboratorio de Acústica y Electroacústica, Escuela de Ingeniería, Electrónica, Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario, Rosario, Santa Fe, Argentina. Recuperado de: http://www. fceia. unr. edu. ar/prodivoz/fonatorio. pdf*, 1999.

[40] Julián Zafra. *Ingeniería de Sonido. Conceptos, fundamentos y casos prácticos*. Ra-Ma Editorial, 2018.

[41] Shin-Woong Cho, Chang Shik Yin, Young-Bae Park, and Young-Jae Park. Differences in self-rated, perceived, and acoustic voice qualities between high-and low-fatigue groups. *Journal of Voice*, 25(5):544–552, 2011.

[42] João Paulo Teixeira, Carla Oliveira, and Carla Lopes. Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia Technology*, 9:1112–1122, 2013.

[43] João Teixeira and André Gonçalves. Accuracy of Jitter and Shimmer Measurements. *Procedia Technology*, 16:1190–1199, 12 2014.

[44] Pascal Belin, Sarah Fillion-Bilodeau, and Frédéric Gosselin. The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior research methods*, 40(2):531–539, 2008.

[45] Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter Sendlmeier, and Benjamin Weiss. A database of German emotional speech. volume 5, pages 1517–1520, 09 2005.

[46] Silke Paulmann and Sonja A Kotz. An ERP investigation on the temporal dynamics of emotional prosody and emotional semantics in pseudo-and lexical-sentence context. *Brain and Language*, 105(1):59–69, 2008.

[47] MN Utah and JC Jung. Fault state detection and remaining useful life prediction in AC powered solenoid operated valves based on traditional machine learning and deep neural networks. *Nuclear Engineering and Technology*, 52(9):1998–2008, 2020.

[48] Ana P Pinheiro, Andrey Anikin, Tatiana Conde, João Sarzedas, Sinead Chen, Sophie K Scott, and César F Lima. Emotional authenticity modulates affective and social trait inferences from voices. *Philosophical Transactions of the Royal Society B*, 376(1840):20200402, 2021.

[49] Md Shah Fahad, Ashish Ranjan, Jainath Yadav, and Akshay Deepak. A survey of speech emotion recognition in natural environment. *Digital signal processing*, 110:102951, 2021.

[50] Mirosław Płaza, Sławomir Trusz, Justyna Kęczkowska, Ewa Boksa, Sebastian Sadowski, and Zbigniew Koruba. Machine learning algorithms for detection and classifications of emotions in contact center applications. *Sensors*, 22(14):5311, 2022.

[51] Carlos Gómez Gil. Objetivos de Desarrollo Sostenible (ODS): una revisión crítica. *Papeles de relaciones ecosociales y cambio global*, 140(1):107–118, 2018.