# Clasificación de emociones complejas en Audio de Conversaciones de Call Center de la Universidad Javeriana Cali mediante Modelos Semi Supervisados de Machine Learning

*Julián Andrés Ospina Cuesta*
*Código 8984933*

*Proyecto Aplicado para optar al título de*
*Magister en Ciencia de Datos*

Director(a)
Gloria Álvarez

Codirector(a)
Diego Linares

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, DICIEMBRE 1 DE 2024

# TABLA DE CONTENIDO

# LISTA DE FIGURAS

## LISTA DE TABLAS

## LISTA DE ANEXOS

Anexo 1-README.md de la app Análisis Emociones en formato html.
Anexo 2- Dashboard para la Captura y Gestión de Sugerencias de Etiquetado de Llamadas en el Call Center.

# INTRODUCCIÓN

La creciente digitalización de los servicios ha impulsado a las organizaciones a buscar constantemente formas innovadoras de mejorar la experiencia del cliente, dado que en un entorno cada vez más competitivo, la satisfacción del cliente se ha convertido en un factor diferenciador clave. En este contexto, los call centers desempeñan un papel crucial al ser el principal punto de contacto entre las organizaciones y sus clientes, donde cada interacción puede influir significativamente en la percepción del servicio recibido. No obstante, uno de los mayores desafíos que enfrentan estas organizaciones es entender y analizar las emociones y sentimientos de los clientes durante estas interacciones telefónicas, ya que captar las sutilezas emocionales expresadas en una conversación puede ser complejo y difícil de medir con precisión.

Este trabajo de grado se centra en abordar esta problemática mediante el análisis de las emociones en las llamadas del call center de la Universidad Javeriana Cali. El objetivo principal es identificar y clasificar las emociones predominantes en las grabaciones de llamadas, lo que permitirá proporcionar información valiosa y accionable sobre la satisfacción del cliente, contribuyendo así a la mejora continua del servicio y a una comprensión más profunda de las necesidades y expectativas de los usuarios.

Para alcanzar este objetivo, se emplean técnicas de aprendizaje automático para el reconocimiento de emociones en audios desde características acústicas. Estas técnicas involucran la extracción de características como el tono, la intensidad, la velocidad del habla y la tasa de articulación, las cuales luego se utilizan como entrada para un modelo de aprendizaje automático.

Inicialmente, se planteó un enfoque de sugerencias de etiquetado apoyado en emociones básicas[1], es decir, emociones fundamentales, innatas y universales. Sin embargo, posteriormente se solicitó migrar a categorías más alineadas con el negocio, enfocándose en emociones complejas, que son más desarrolladas, moduladas culturalmente y específicas de un contexto social y personal.

De igual forma, se propuso implementar un modelo de aprendizaje semi-supervisado autodidacta en varias etapas para etiquetar la base de datos. Este enfoque permitió ampliar las etiquetas propuestas basándose en emociones previamente sugeridas al call center, utilizando información parcial obtenida a través de un etiquetado supervisado. No obstante, esta metodología solo se aplicó exitosamente a las 'no llamadas'[2]. Para las 'llamadas', debido a la baja precisión del modelo

[1] Las emociones básicas son emociones primarias, reconocidas universalmente y compartidas por todas las culturas humanas. Las emociones básicas tienden a ser más simples y se manifiestan a través de expresiones faciales y corporales de manera similar en todo el mundo [41]

[2] 'No llamadas' corresponde a los audios que contiene contestadoras, 'ruidos' y 'Interrupciones' donde no hay una interacción entre los interlocutores

semi-supervisado, se incrementó los audios clasificados hasta llegar a un modelo completamente supervisado para asegurar una mayor exactitud en el etiquetado.

Se han desarrollado modelos que logran una precisión del 95% en la clasificación de 'No llamadas' en la detección de tipos de audios para el call center, así como una detección de emociones predominantes del 41% para emociones complejas. Se espera que estos modelos sienten las bases para mejorar la experiencia del cliente en la Universidad Javeriana Cali.

# 1. DEFINICIÓN DEL PROBLEMA

## 1.1. PLANTEAMIENTO DEL PROBLEMA

La satisfacción del cliente es vital para cualquier empresa que brinde bienes o servicios, como es el caso de la Universidad Javeriana Cali. Aquí, el centro de contacto telefónico, o call center, es fundamental en la interacción directa con los clientes. Dada la importancia de esta interacción, surge la necesidad de emplear técnicas que permitan evaluar las llamas de forma rápida y aproximar la identificación de la emoción predominante en la misma. En particular, la tarea de identificar emociones en las llamadas telefónicas del call center de la Universidad Javeriana Cali presenta un reto significativo, ya que los audios carecen de una clasificación previa y no se conoce la totalidad de los registros disponibles al inicio del proyecto. La ciencia de datos, a través del análisis de grandes volúmenes de datos y el uso de algoritmos de aprendizaje automático y semi automáticos, ofrece una alternativa potencial para analizar patrones en el tono y el contenido de las conversaciones. Estos métodos permiten identificar emociones y tendencias, lo que resulta crucial para mejorar la satisfacción del cliente y optimizar los recursos.

En la Universidad Javeriana Cali, la calidad en la atención telefónica no solo es una prioridad, sino una expresión tangible del compromiso de la universidad por brindar servicios de alto nivel en todos los aspectos de su comunidad educativa. Por lo tanto, es esencial evaluar de forma continua la satisfacción de los usuarios. El seguimiento constante de los determinantes de esta satisfacción se revela crucial para mantener estándares elevados en los indicadores de satisfacción del cliente.

En el marco de la interacción empleado-cliente, o incluso en cualquier comunicación interpersonal, la percepción de las emociones desempeña un papel fundamental. Esto teniendo en cuenta que las percepciones emocionales son interpretaciones subjetivas de estímulos, generando respuestas afectivas y cognitivas. Para el escenario en mención, la percepción auditiva.

En un entorno donde se espera manejar un promedio de 20 llamadas por hora [1], la identificación de emociones se convierte en un proceso que demanda un procesamiento asistido para determinar la emoción predominante en cada caso. Se han llevado a cabo diversos estudios que abordan la identificación automática de emociones en llamadas telefónicas de centros de contacto, utilizando enfoques lingüísticos, y algunos más recientes han explorado la problemática mediante el análisis de características acústicas [2], que es precisamente el enfoque de este trabajo.

## 1.2. FORMULACIÓN DEL PROBLEMA

En la actualidad, el call center de la Universidad Javeriana se enfrenta a la tarea crítica de gestionar un volumen considerable de llamadas telefónicas, abordando diversas consultas, inquietudes y situaciones de los usuarios. La eficacia en la interacción con los llamantes es esencial para mantener un servicio de calidad. Sin embargo, la capacidad de comprender y abordar las emociones de los interlocutores durante las conversaciones telefónicas ha sido un desafío constante.

¿Cómo puede desarrollarse un modelo basado en técnicas de machine learning para la clasificación automática de emociones en llamadas reales de un centro de contacto universitario, con el propósito de mejorar la evaluación de la satisfacción del cliente y proporcionar información complementaria?

Para lograr responder a esa pregunta, se proponen las siguientes preguntas que cumplen la función de sistematizar la formulación general.

¿Cómo preparar los datos para el desarrollo del modelo de machine learning que automatice la clasificación de las emociones predominantes en un audio?
¿Cómo aplicar técnicas de machine learning para obtener un modelo de clasificación automática de emociones en llamadas de centros de contacto?
¿Cómo evaluar la efectividad del modelo de machine learning empleado?

# 2. OBJETIVOS DEL PROYECTO

## 2.1. OBJETIVO GENERAL

Desarrollar un modelo de clasificación automática de emociones en llamadas reales de un centro de contacto universitario utilizando técnicas de machine learning, con el propósito de brindar la emoción básica predominante como nuevo elemento en la evaluación de la satisfacción del cliente.

## 2.2. OBJETIVOS ESPECÍFICOS

1. Procesar y analizar las grabaciones de audio de las llamadas telefónicas del centro de contacto de la Universidad Javeriana para extracción de la señal acústica del habla, las cuales serán utilizadas como insumos clave en el desarrollo del modelo de machine learning para la clasificación automática de emociones.
2. Desarrollar un modelo de machine learning para la clasificación automática de emociones en las llamadas del centro de contacto.
3. Evaluar la efectividad del modelo.
4. Realizar una aplicación o software que permita evaluar en nuevos audios la emoción predominante con base en el modelo obtenido.

# 3. MARCO TEÓRICO Y ANTECEDENTES

## 3.1. MARCO TEÓRICO

Teniendo en cuenta que el proyecto gira en torno al desarrollo de un modelo de machine learning que permita la clasificación de conversaciones en términos del contenido emocional de las mismas, el marco teórico del presente trabajo debe estar comprendido por: la revisión de teorías y conceptos asociados a la representación de las señales acústicas y las diferentes técnicas de machine learning disponibles y que sean potencialmente útiles para el desarrollo del modelo; también será importante identificar un modelo explicativo de las expresiones emocionales que permita fundamentar la idea de que es posible establecer una relación entre el contenido acústico de las conversaciones (señales de audio) y la expresión de una emoción en particular.

### 3.1.1. Características del Audio

La formalización cualitativa del habla implica la convolución de la respuesta en frecuencia del tracto vocal con el pulso glótico. En términos simples, esto significa que el pulso glótico, originado en las cuerdas vocales, se filtra a través del tracto vocal, dando lugar a la señal acústica. El pulso glótico, donde se origina el fundamento de la frecuencia sonora o tono, determina las frecuencias más altas. La forma específica del tracto vocal da lugar a la generación de fonemas o al timbre característico de la señal [3].

En la Figura 1 [4] se aprecian las 3 principales unidades para la generación de la voz: los pulmones (fuente de energía), las cuerdas vocales (oscilador) y el tracto vocal (resonador). A partir de la acción de los pulmones, las cuerdas vocales vibran como una fuerza oscilante y, junto con las cavidades resonantes del tracto vocal, boca y nariz, esta fuerza crea las ondas sonoras requeridas para la voz.

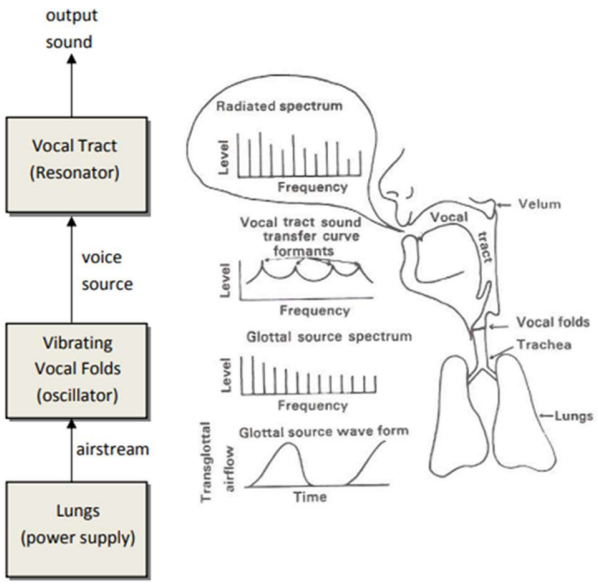



Figura 1: Esquema de la generación de sonidos de voz

El esquema de la Figura 1 representa la base de preproducción de voz, y de esta salen diferentes modelos de la reproducción de la voz humana que se pueden apreciar en la Tabla 1 [4].

Tabla 1: Modelos de reproducción de la voz

| Nombre del Modelo | Descripción del modelo | Representación matemática |
|---|---|---|
| Source-Filter Model (Modelo Fuente-Filtro) | Este modelo explica la producción del habla en dos etapas principales:<br><br>***Fuente:** Se refiere a la producción de sonido en las cuerdas vocales o glotis. La fuente genera un sonido bruto, similar a un zumbido.<br>***Filtro:** Representa la configuración de las cavidades resonantes del tracto vocal (faringe, boca, etc.). Este filtro modifica el sonido crudo generado por la fuente para crear diferentes sonidos del habla. | $S(t) = G(t) * V(t) * L(t)$<br>Donde:<br>$*\boldsymbol{S(t)}$ es la señal de voz final.<br>$*\boldsymbol{G(t)}$ es la fuente de la señal, que representa la vibración de las cuerdas vocales.<br>$*\boldsymbol{V(t)}$ es el filtro del tracto vocal, que modela las resonancias del tracto vocal (faringe, boca, etc.).<br>$*\boldsymbol{L(t)}$ es el filtro de radiación que modela el efecto de la radiación en los labios y las fosas nasales.<br><br>**Nota:** El símbolo * representa la **convolución** entre la fuente y el filtro |
| Extracción de | Las señales de voz se clasifican en | La representación de la |

| | | |
|---|---|---|
| características de una señal de voz | sonoras y sordas. Las señales sonoras son generadas por la vibración de las cuerdas vocales, como las vocales, y tienen una periodicidad definida. En contraste, las señales sordas, como "s", "p" o "ch", son producidas por el paso rápido del aire por el tracto vocal con la glotis parcialmente abierta, y tienen poca o ninguna periodicidad.<br><br>Las señales de voz son no estacionarias, lo que significa que sus características cambian con el tiempo. Para analizarlas correctamente, se segmentan en intervalos cortos (entre 10 y 30 ms), donde se consideran casi estacionarias debido a los movimientos lentos de los articuladores vocales.<br><br>Este proceso se realiza aplicando ventanas deslizantes (como Hanning o Hamming) a la señal completa, generando tramas que permiten realizar un análisis de las características acústicas. | Transformada de Fourier de Tiempo Corto (STFT o DTFT) para la trama $\boldsymbol{m}$ se define aplicando la Transformada de Fourier a una sección corta de la señal de voz, multiplicada por una ventana. La fórmula general es:<br>$$X_m\left(e^{j\omega}\right)=\sum_{n=-\infty}^{+\infty}x_m[n]e^{-j\omega n}=\sum_{n=-\infty}^{+\infty}w_m[n]x[n]e^{-j\omega n}$$<br>Donde:<br>***$X_m(e^{j\omega})$** es la Transformada de Fourier de Tiempo Corto para la trama $\boldsymbol{m}$ en el dominio de la frecuencia.<br>***$x[n]$** es la señal de voz.<br>***$w_m[n]$** es la función de ventana centrada en el tiempo $\boldsymbol{m}$, que permite seleccionar un segmento corto de la señal.<br>***$e^{-j\omega n}$** es el término de la Transformada de Fourier (con $\omega$ representando la frecuencia angular).<br><br>La idea es que para cada trama $\boldsymbol{m}$, se analiza un pequeño segmento de la señal $x[n]$ con una ventana $w[n]$, y luego se aplica la Transformada de Fourier a ese segmento. |
| Glottal Flow Models (Modelos de flujo glotal) | Los modelos de flujo glotal se centran en la generación de la señal en la fuente, específicamente en la vibración de las cuerdas vocales. Se pueden clasificar las aproximaciones en tres categorías principales:<br><br>***Métodos en el dominio del tiempo:** El ciclo glótico puede dividirse en varias fases. En los métodos basados en el dominio temporal, se identifican momentos clave, como el instante de apertura y cierre de la glotis, directamente en el pulso del flujo | **Método en el dominio del tiempo:**<br>$$OQ=\frac{T_p+T_l}{T}$$<br>*$T_p+T_l$ es el tiempo durante el cual las cuerdas vocales están abiertas.<br>*$T$ es la duración total del ciclo glótico.<br>$$ClQ=\frac{T_l}{T}$$<br>* $T_l$ es el tiempo que tarda la glotis en cerrarse completamente.<br>*T es la duración total del ciclo glótico. |

| | | |
|---|---|---|
| | glotal. A partir de estos instantes críticos, se pueden medir las duraciones de las distintas fases del ciclo, proporcionando información detallada sobre la dinámica de la fonación.<br><br>***Métodos en el dominio de frecuencia:** La parametrización del flujo glótico en el dominio de la frecuencia utiliza parámetros como la diferencia entre el primer y segundo armónico ($\Delta H_{12}$) y el factor de riqueza armónica (HRF), que dependen de la frecuencia fundamental.<br><br>***Métodos basados en modelos:** Los métodos de parametrización basados en modelos buscan capturar la forma general del flujo glotal sin considerar los detalles finos del movimiento de las cuerdas vocales. Utilizan fórmulas matemáticas para generar formas de onda artificiales que imitan los pulsos del flujo glótico | $$SQ = \frac{T_p}{T_l}$$<br>*$Tp$ es el tiempo que tarda la glotis en abrirse.<br>*$Tl$ es el tiempo que tarda la glotis en cerrarse.<br><br>**Método en el dominio de frecuencia:**<br>$$HRF = \frac{\sum_{k\geq 2} H_k}{H_1}$$<br>*$H_k$ es la amplitud del armónico<br>*$H_1$ es la amplitud del primer armónico.<br>*$N$ es el número total de armónicos considerados en la sumatoria.<br><br>**Métodos basados en modelos:** Existen varios métodos, como lo son:<br>**Modelo LF (Liljencrants-Fant)**<br>$$g'_{LF}(t) = \begin{cases} E_0 e^{\alpha t}\sin(w_g t) & , \quad 0 \le t \le t_e \\ -\frac{E_e}{\beta t_a}\left(e^{-\beta(t-t_e)} - e^{-\beta(t_c-t_e)}\right) & , \quad t_e < t \le t_c = T_0 \end{cases}$$<br>**Modelo Flant Glottal**<br>$$g_F(t) = \begin{cases} \frac{1}{2}\left(1-\cos(\omega_g t)\right) & , \quad 0 \le t \le t_p \\ K\cos\left(\omega_g\left(t-t_p\right)\right) - K + 1 & , \quad t_p < t \le t_c = t_p + \frac{1}{\omega_g} arc\cos\frac{K-1}{K} \\ 0 & , \quad tc < t \le T_0 \end{cases}$$ |

Frente a la implementación de procesamientos de audios desde librerías Python, se han tomado como apoyo en las características más prometedoras para detección de emociones dentro de las mencionadas por: el PhD. J.R. Zapata González [5], en el trabajo de señales bioacuáticas de Caycedo Paula, Ruiz Jose y Orozco Mauricio [6], Schorkhuber Christian de la University of Music and Performing Arts y Klapuri Anssi de Queen Mary University of London [7], Graz y Aguirre Fabián en su trabajo de master [8]:

- **Mel-Espectrograma:** Utilizando la escala Mel, que relaciona la frecuencia percibida con la medida real, el Mel-Espectrograma se centra en la amplitud logarítmica debido a la percepción humana logarítmica de la intensidad del sonido.

- **Transformada Wavelet**: Útil para representar señales no estacionarias con energía dispersa en un amplio rango de frecuencias. A diferencia de la Transformada de Fourier, la Wavelet proporciona información simultánea en el dominio del tiempo y la frecuencia, permitiendo una mejor identificación de patrones acústicos y detección de vocalizaciones en entornos ruidosos. [6]
- **Mel Frequency Cepstral Coefficients (MFCCs):** Los MFCCs son coeficientes utilizados en la representación del habla, basados en la percepción auditiva humana. Estos coeficientes imitan la percepción auditiva humana y capturan las características espectrales más relevantes, lo que los hace altamente efectivos para tareas como el reconocimiento de instrumentos musicales y la clasificación de sonidos ambientales. Se destacan por su capacidad para representar la envolvente del espectro de potencia de corto plazo de una señal de audio y su uso generalizado en el procesamiento de voz y música, debido a su precisión en entornos controlados. [9] Se centran en información valiosa, excluyendo detalles irrelevantes como el ruido de fondo, emociones, volumen y tono, y son comúnmente empleados para describir el timbre.
- **Características Espectrales:** Para la clasificación de sonidos, se utilizan momentos espectrales como el centroide, ancho de banda, asimetría, curtosis, entre otros, que proporcionan información sobre la distribución espectral. El centroide espectral, por ejemplo, indica la frecuencia central de la energía espectral, similar a una media ponderada.
- **Centroide espectral (Spectral Centroid):** Se usa para describir la tonalidad percibida de un sonido, ya que valores altos indican una mayor presencia de frecuencias agudas y valores bajos sugieren predominio de frecuencias graves. Su cálculo considera el espectro como una distribución de probabilidad, donde la amplitud en cada frecuencia actúa como un peso relativo. [8]
- **Transformada Q constante (Constant-Q transform)**: Representación tiempo-frecuencia en la que los bins de frecuencia están espaciados geométricamente y todos tienen el mismo factor Q (relación entre la frecuencia central y el ancho de banda). A diferencia de la Transformada de Fourier Discreta, que tiene bins equidistantes en frecuencia, el CQT ofrece mejor resolución en bajas frecuencias y mejor resolución temporal en altas frecuencias, lo que lo hace particularmente útil en el análisis de señales musicales. [7]

### 3.1.2. Las emociones humanas

El estudio de las emociones humanas constituye un fascinante campo multidisciplinario que involucra a la psicología, la neurociencia, la filosofía y otras disciplinas afines. La complejidad de las emociones radica en su capacidad para influir en el pensamiento, el comportamiento y la experiencia subjetiva de los individuos. Los investigadores buscan comprender la naturaleza de las emociones, su origen evolutivo, y cómo interactúan con la cognición y el cuerpo humano. Desde la alegría hasta el miedo, las emociones desempeñan un papel crucial en la toma de decisiones, las relaciones interpersonales y la salud mental. El avance tecnológico, incluyendo técnicas de imagen cerebral y análisis de datos, ha permitido un mayor conocimiento sobre la

base neural de las emociones. Este campo en constante evolución no sólo arroja luz sobre los misterios de la mente humana, sino que también tiene aplicaciones prácticas en la mejora de la salud mental, la inteligencia artificial y el diseño de interacciones humanas más efectivas [10].

#### 3.1.2.1. Emociones básicas

Paul Ekman es un autor que defiende la idea de la existencia de un grupo de emociones básicas universales, cuya clasificación es ampliamente utilizada con fines de clasificación discreta de las emociones en distintas áreas, incluyendo recientemente la mayoría de los trabajos en ciencia de datos que buscan la identificación automática de las emociones. Estas emociones básicas son: alegría, tristeza, miedo, ira, desagrado y sorpresa [11].

En cuanto al enfoque principal del abordaje de la expresión emocional del presente trabajo, el principal referente teórico es Ekman, quien defiende una postura acerca de la existencia de formas de expresión emocional universales a toda la especie humana, lo que incluye, además de una fuerte evidencia acerca de la universalidad de las expresiones faciales, unas hipótesis fuertes alrededor de la universalidad en todos los canales de expresión emocional, incluyendo la señal acústica del habla, que además implicaría una universalidad de dichos patrones acústicos [12].

#### 3.1.2.2. Psicolingüística emocional y sus claves para la detección de emociones

La ciencia que estudia los sentimientos y la comunicación verbal se conoce como psicolingüística emocional. Esta disciplina se enfoca en comprender cómo las emociones influyen en la producción y percepción del lenguaje, así como en cómo el lenguaje puede afectar el estado emocional de las personas. En la psicolingüística se define la emoción humana como la capacidad de comunicación oral (o gestual, en el caso de los sordomudos) exclusiva de los seres humanos, que permite transmitir eventos internos y externos usando un sistema de señales convencionales organizadas en dos niveles: sonido y sentido. [13]

En autor Jorge Fernández en el libro ‘lenguaje cuerpo y mente: Psicolingüística emocional’ define varias características del lenguaje [13], en donde menciona característica acústica como lo son: Canal vocal-auditivo, transmisión irradiada y recepción direccional, pero menciona muchas más relacionadas a la semántica del mensaje como lo son: Desvanecimiento rápido, intercambialidad, retroalimentación total, semanticidad, arbitrariedad, entre otros. Planteando nuevamente la pregunta: ¿Qué tan buena es la representación acústica en solitario de las emociones complejas en los audios? Con base en trabajos relacionados y desde el punto de la Psicolingüística emocional faltarían más elementos para una buena representación de la señal.

### 3.1.3. Machine learning

Algunos modelos empleados para la clasificación de audios empleando procesamiento de las características del Audio son los siguientes:

- **Máquinas de soporte vectorial (SVM):** Es un algoritmo de aprendizaje automático diseñado para tareas de clasificación y regresión, que busca encontrar el "hiperplano" que

mejor separa los datos en diferentes clases en un espacio multidimensional. Su principal objetivo es maximizar la distancia (o margen) entre las clases, lo que contribuye a mejorar la precisión del modelo. A diferencia de otros métodos de clasificación, SVM clasifica las alternativas evaluadas a través de una función de puntuación que no es ni lineal ni paramétrica. [14].

- **Árboles de decisión:** Modelo que se asemeja a una estructura en forma de árbol, donde los nodos internos representan características o atributos, las ramas denotan reglas de decisión, y cada nodo hoja indica el resultado. La selección de nodos se realiza mediante conceptos como la entropía o la ganancia de información, propiedades estadísticas que evalúan cómo un atributo específico separa los ejemplos de entrenamiento según su clasificación objetivo [15].
- **Random Forest (RF):** Modelo que involucra varios árboles de decisión combinados con bagging. Al usar bagging, cada árbol ve distintas porciones de los datos, ninguno usa todos los datos de entrenamiento. Esto hace que cada uno se entrene con distintas muestras para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros, obteniendo una predicción que generaliza mejor [15].
- **Redes neuronales de perceptrón multicapa (FFNN):** Una red neuronal se presenta como un procesador masivamente distribuido y en paralelo con una inclinación inherente para capturar conocimiento experimental y hacerlo accesible para su aplicación. Este modelo guarda similitudes con el cerebro en dos aspectos fundamentales: 1. La red adquiere conocimiento a través de un proceso de aprendizaje. 2. Las fuerzas de conexión entre neuronas, denominadas ponderaciones sinápticas, se emplean para retener la información adquirida [24]. El perceptrón, la forma más básica de red neuronal utilizada para clasificar patrones linealmente separables, se expande en el perceptrón multicapa, incorporando múltiples capas. Este diseño incrementa su capacidad para abordar problemas que no presentan una separación lineal clara [16].
- **Redes neuronales recurrentes (RNN):** Una red neuronal recurrente (RNN) es un tipo de red neuronal artificial diseñada para trabajar con datos secuenciales o series de tiempo. Estos algoritmos de aprendizaje profundo son especialmente eficaces para abordar problemas temporales u ordinales. Las RNN aprenden a partir de datos de entrenamiento y se distinguen por su capacidad de 'memoria', ya que pueden retener información de entradas anteriores y utilizarla en las entradas actuales y en los resultados. A diferencia de las redes neuronales tradicionales, que suponen que los datos de entrada y los resultados son independientes, las RNN consideran que los resultados dependen de elementos anteriores dentro de una secuencia. Aunque los eventos futuros podrían ser relevantes para determinar los resultados, las RNN unidireccionales no pueden incorporar esta información en sus predicciones [17].
- **Redes neuronales convolucionales (CNN):** Las redes neuronales son un componente fundamental del aprendizaje automático y del aprendizaje profundo. Estas redes están compuestas por capas de nodos que incluyen una capa de entrada, varias capas ocultas y una capa de salida. Cada nodo cuenta con un peso y un umbral que determinan su activación. Las redes neuronales convolucionales (CNN) son especialmente efectivas en el

procesamiento de imágenes, voz y audio, y se estructuran en tres tipos principales de capas:

- Capa convolucional: La primera capa que extrae características de la entrada.
- Capa de agrupamiento: Reduce la dimensionalidad de los datos.
- Capa totalmente conectada (FC): Capa final que combina la información para la clasificación.

A medida que los datos atraviesan las capas, las CNN identifican primero características simples, como colores y bordes en el caso de las imágenes, y gradualmente reconocen formas y elementos más complejos, hasta llegar a identificar el objeto previsto [18]

### 3.1.4. Métricas de evaluación de aprendizaje semi supervisado

A continuación, se presentan algunas de las técnicas utilizadas para evaluar modelos de aprendizaje semi-supervisado [19] [20]:

- **Exactitud:** Esta métrica representa la proporción de predicciones correctas en relación con el total de predicciones realizadas. Se calcula dividiendo la suma de verdaderos positivos y verdaderos negativos entre el número total de casos.
- **Precisión:** La precisión mide la proporción de predicciones positivas que son acertadas. Su cálculo se realiza dividiendo los verdaderos positivos entre la suma de verdaderos positivos y falsos positivos.
- **Recall o Sensibilidad:** Esta métrica evalúa la proporción de casos positivos reales que son correctamente identificados por el modelo. Se calcula dividiendo los verdaderos positivos entre la suma de verdaderos positivos y falsos negativos.
- **F1-score:** El F1-score combina precisión y recall en una única medida. Su cálculo se realiza mediante el doble del producto de precisión y recall dividido por la suma de precisión y recall.

### 3.1.5. Técnicas de etiquetado semi automático

Para la generación de etiquetas para los datos no etiquetados a partir de los etiquetados, se cuenta con las siguientes técnicas [21]:

- **Co-entrenamiento:** Consiste en entrenar simultáneamente dos o más modelos, cada uno utilizando conjuntos de características diferentes para un mismo ejemplo. Luego, las predicciones de cada modelo se emplean para etiquetar los datos no etiquetados del otro modelo, eliminando la necesidad de revisión manual por parte de un experto, lo que reduce costos y tiempos significativamente. [22] . Este enfoque se basa en la suposición de que las características son independientes y se complementan entre sí.
- **Aprendizaje auto-didacta:** En el aprendizaje auto-didacta, se entrena un modelo utilizando datos etiquetados, y luego se emplean las predicciones de dicho modelo para etiquetar los datos que carecen de etiquetas. La premisa subyacente es que el modelo tiene la capacidad de generalizar de manera efectiva a partir de los datos etiquetados.

- **Aprendizaje basado en grafos:** La metodología de aprendizaje basado en grafos implica la representación de datos en forma de un grafo, donde los ejemplos son nodos y las similitudes entre ellos se expresan mediante aristas. Posteriormente, las etiquetas de los nodos etiquetados se propagan a los nodos no etiquetados a través de las aristas, utilizando algún criterio de optimización o inferencia.
- **Aprendizaje basado en modelos generativos:** Este enfoque implica modelar la distribución conjunta de las características y las etiquetas. Luego, se utiliza el algoritmo de Expectation-Maximization (EM) para estimar los parámetros del modelo y asignar etiquetas a los datos no etiquetados. Este método parte de la suposición de que los datos siguen una distribución probabilística específica.

### 3.2. ANTECEDENTES

A continuación, se enumeran los antecedentes más importantes encontrados en lo referente a la identificación y clasificación de emociones a partir de señales auditivas y luego se presenta la evidencia de lo revisado hasta el momento que tiene relación directa con el contexto de los centros de contacto.

En el campo del reconocimiento emocional por medio de señales de voz, es importante mencionar el trabajo de tesis doctoral de Humberto Pérez [23], quien utilizó un modelo de clasificación continuo de las emociones, haciendo uso de un auto entrenamiento semi-supervisado, el uso de agrupamiento difuso mediante la técnica Fuzzy C-means (FCM) y la técnica probabilística de campos aleatorios de Markov (CAM), con el fin de generar un método de predicción y reconocimiento de patrones emocionales espontáneos. De este trabajo puede ser útil el modelo de clasificación continuo de las emociones y su uso en el posible etiquetado automático del audio.

También vale la pena mencionar el trabajo de grado de Sánchez [24], quien utiliza una técnica de aprendizaje automático para entrenar un modelo a partir de imágenes y voz e identificar emociones discretas. De este trabajo cabe resaltar que los resultados arrojaron una mejor predicción de las imágenes que de la voz y una mejor predicción cuando se mezclaban imágenes y voz que cuando se hacían por separado. De este trabajo se valdrán para sustentar la importancia de desarrollar modelos más confiables de predicción con base en voz, en el contexto de los centros de contacto.

Por otra parte, el trabajo de Bello y colaboradores [25], ilustra un abordaje del reconocimiento de las emociones básicas en audios en condiciones controladas por medio del análisis de fragmentos de la voz que utiliza la transformada rápida de Fourier (FFT) y posteriormente aplicando coeficientes de correlación de Pearson. El trabajo resulta interesante, además, porque presenta una metodología completa fundamentada en evidencia científica para lograr etiquetar las emociones, sustentada en parámetros observables de la señal del habla.

En otro trabajo, se utilizó la base de datos RAVDESS (The Ryerson Audio-Visual Database of

Emotional Speech and Song) [26] para entrenar un modelo a través de técnicas de aprendizaje supervisado, redes neuronales artificiales y redes neuronales convolucionales. El modelo logró un nivel de precisión del 85% de clasificación de las emociones de personas en videos con audio que mantienen una configuración parecida a las de la base de datos de entrenamiento [27]. Este trabajo permitirá sustentar e informar sobre el uso de técnicas de aprendizaje supervisado y redes neuronales, con el fin de seleccionar la más adecuada para el contexto de nuestro trabajo con audios de conversaciones reales.

Por otra parte, el trabajo de Pervaiz y Ahmed [28] se enfoca en un modelo para el reconocimiento de emociones utilizando características prosódicas, temporales y lingüísticas, demostrando que esta combinación mejora significativamente la precisión. Se emplea un modelo de dos etapas: en la primera, se extraen emociones basándose en características prosódicas y temporales; en la segunda, se utilizan la segmentación de palabras y características lingüísticas para la identificación emocional. Los resultados indican que la combinación de ambas etapas produce un rendimiento superior en comparación con el uso de cada etapa de forma independiente, aumentando la precisión sin afectar negativamente el rendimiento general. Además, los experimentos demostraron que la precisión en la clasificación mejora al considerar el factor de edad durante el entrenamiento del clasificador, ya que no tenerlo en cuenta resulta menos efectivo. En nuestro trabajo, hemos identificado esta diferencia como un aspecto a considerar, ya que nuestro enfoque actual se limita a referencias acústicas y no incorpora características sociodemográficas de los interlocutores.

De igual manera, el proyecto de reconocimiento de emociones en la voz de Lerache y Elkfury [29] aborda la dificultad de analizar emociones en el discurso hablado en español rioplatense, especialmente en entornos interactivos con robots. Se reconoce que la voz humana tiene limitaciones en la transmisión de emociones, y se propone utilizar espectrogramas y técnicas de aprendizaje profundo, como redes neuronales convolucionales (CNN) y recurrentes (RNN), para desarrollar un clasificador de emociones. Además, se busca mejorar la comunicación persona-máquina mediante la creación de una API para la explotación del modelo, la comparación de rendimiento entre los clasificadores y la integración en un framework multimodal. Se plantea un método para la conversión de enfoques categóricos a dimensionales y se establecen bases para la mejora continua mediante la adquisición de nuevos datos para el entrenamiento del modelo. En conjunto, la investigación busca avanzar en la comprensión y aplicación de las emociones en el discurso hablado, contribuyendo a la mejora de la interacción emocional en entornos tecnológicos. Este trabajo es de alta relevancia, puesto que trabaja con particularidades propias de la señal de audio en idioma español, lo cual es escaso en la literatura.

Asimismo, los mismos autores abordan [30] el reconocimiento de emociones en señales de voz mediante el uso de redes neuronales profundas. La identificación precisa de emociones en el habla es crucial para diversas aplicaciones, y el estudio se centra en evaluar los efectos de funciones de pérdida y técnicas de aumento de datos en la clasificación de siete emociones. Motivado por la colaboración con la empresa SEAT para evaluar sistemas basados en deep learning en situaciones de conducción, el proyecto tiene como objetivo principal implementar un sistema eficaz de reconocimiento de emociones en señales de voz, aprovechando información

espectral. Las conclusiones resaltan el éxito de la implementación, subrayando la necesidad de evaluar configuraciones y técnicas para mejorar la precisión del modelo y destacando su potencial aplicabilidad en situaciones prácticas, incluyendo el reconocimiento de emociones en contextos de conducción.

La investigación realizada por M. Pervaiz y T. Ahmed Khan (2016) enseña la combinación de dos modelos para la detección de emociones en audios: modelo de emoción de audio usando características prosódicas. tales como: Zero Crossing Rate (ZCR), Cepstral Coefficients (MFCC) y la Linear Preddiction Coefficient (LPC) en un modelo Support Vector Machine (SVM), en conjunto con un modelo que contenga lingüística (NLP), permitiéndoles mejorar la precisión desde 50%-60% de los modelos individuales a un 70% en conjunto. [28] Este enfoque proporciona las metodologías utilizadas en el análisis de audios, así como las limitaciones y soluciones durante la construcción del modelo.

El trabajo de grado realizado por M. Patricio G. y A. Berlanga en 2022 se enfocó en la implementación de un análisis de emociones a través de la conversión de audio a texto. Este proceso de análisis fue llevado a cabo mediante el entrenamiento de un modelo utilizando un corpus de EMOFILM y la aplicación de herramientas de Procesamiento del Lenguaje Natural (NPL). El resultado final consistió en la implementación de un chatbot que permite a las personas enviar archivos de audio y recibir la clasificación emocional correspondiente [31]. Este documento presenta nuevos enfoques de modelos utilizados para el análisis de audios, centrándose en las características específicas de grabaciones de corta duración.

Finalmente, en un trabajo muy cercano a los intereses del presente proyecto, Bolo y colaboradores establecieron correlación entre los niveles de satisfacción reportada por el cliente y la detección automática de patrones emocionales de la señal del habla en una base de datos de 160630 llamadas de call center, analizando únicamente la parte de la llamada correspondiente al cliente y haciendo uso de técnicas de reconocimiento automático de texto y de etiquetado manual y técnicas de regresión, con el fin de generar modelos predictivos que combinaron el texto y la señal auditiva para predecir la valencia emocional y la ira, en particular [2].

### 3.2.1. Aplicaciones disponibles en el mercado

La empresa alemana Audeering ofrece en su portafolio de servicios una serie de dispositivos para análisis de voz, detección de emociones, biomarcadores de voz que permiten la detección de enfermedades como parkinson y Covid-19. Sus soluciones para empresas son implementadas para la detección de emociones en videojuegos de realidad aumentada, call center e investigaciones de mercado. Han sido pioneros desde hace 20 años en este tipo de investigaciones, sin embargo, al ser un producto costoso por la alta tecnología de sus dispositivos, no se ha podido masificar su comercialización, al menos en Latinoamérica [32].

La Unión Europea en 2020 financió el proyecto Mixed Emotions desarrollado por diferentes

colaboradores entre los que se encuentra el Dr. Paul Buitelaar como director del Proyecto. Han tenido grandes avances en aplicaciones multilingües y multimodales de análisis de grandes datos, logrando determinar un perfil muy completo de la emoción del usuario, aunque contiene código abierto y han realizado aplicaciones con fines comerciales, han sido como proyectos piloto y hacen la advertencia de que lograr la interpretación de diferentes usuarios y diferentes fuentes de información, estilos e idiomas es un reto muy amplio y aún no se ha estandarizado para poder realizarlo en un contexto industrial [33].

# 4. PROCESAMIENTO DE AUDIOS DEL CALL CENTER

Iniciaremos con la metodología empleada para convertir las emociones humanas en predicciones de clasificación de emociones, la cual se basa en un proceso sistemático compuesto por varios pasos clave. Esta metodología abarca desde la captura y el preprocesamiento de señales de audio hasta la aplicación de técnicas avanzadas de aprendizaje automático para la clasificación emocional. Este enfoque se ilustra de manera detallada en la Figura 2, que proporciona una visión integral del flujo de trabajo y de las técnicas utilizadas en cada etapa del proceso.


Figura 2: Esquema de procesamiento de audios del call center

El proceso comienza con la extracción de características relevantes de las grabaciones de audio, que comprenden parámetros acústicos específicos para el presente proyecto. A continuación, se selecciona una representación del audio, la cual consiste en una transformación diseñada para optimizar el aprendizaje del modelo y reducir el peso computacional asociado con los datos de audio. Simultáneamente, se realiza la clasificación de los audios.

Para este proyecto, se emplearía un esquema de etiquetado sugerido al call center, que cuenta con etiquetas basadas en las recomendaciones proporcionadas por los directores que interactuaron con el call center, considerando un enfoque inicialmente semi-supervisado. Posteriormente, se entrenarían los modelos de predicción utilizando el etiquetado sugerido y la representación que ha demostrado ser más efectiva con nuestros datos. Finalmente, se llevaría a cabo la optimización del modelo, lo cual permitirá aplicar las clasificaciones emocionales a nuevos audios.

### 4.1. Entendimiento de los audios

Se comenzó el análisis de 2,860 audios de llamadas del centro de atención al cliente y usuarios, correspondientes al periodo del 1 de abril de 2023 al 26 de agosto de 2023. Estos audios fueron entregados en 3 secciones o grabaciones (Ver Tabla 2) en las siguientes fechas: 24 de febrero de 2024, 7 de marzo de 2024 y 22 de abril de 2024.

Tabla 2: Registros de grabaciones del Call Center según paquete compartido

| Envíos de grabaciones | Número de audios |
|---|---|
| Grabaciones 1 | 529 |
| Grabaciones 2 | 1222 |
| Grabaciones 3 | 1109 |
| **Total** | 2860 |

Los audios corresponden a dos tipos de llamadas:

- Grabaciones de llamadas reales entre el call center y los interesados en inscribirse en los diversos programas y cursos que ofrece la universidad. Estas llamadas buscan resolver dudas y/o proporcionar orientación sobre los aplicativos de inscripción o financiamiento disponibles a través de terceros.
- Audios de estudiantes que llamaron al call center para solicitar orientación sobre diversos procesos de la universidad, como refinanciación, matrícula y otros procedimientos.

A continuación, destacan las características principales:

### 4.1.1. Duración de las llamadas:

La duración es una característica puede influir significativamente en el rendimiento de ciertos modelos de análisis, lo que hace fundamental su tratamiento adecuado durante el proceso de extracción de características. En la Figura 3, se presenta una representación visual de esta variabilidad, permitiendo identificar la distribución y dispersión de la duración de las llamadas:

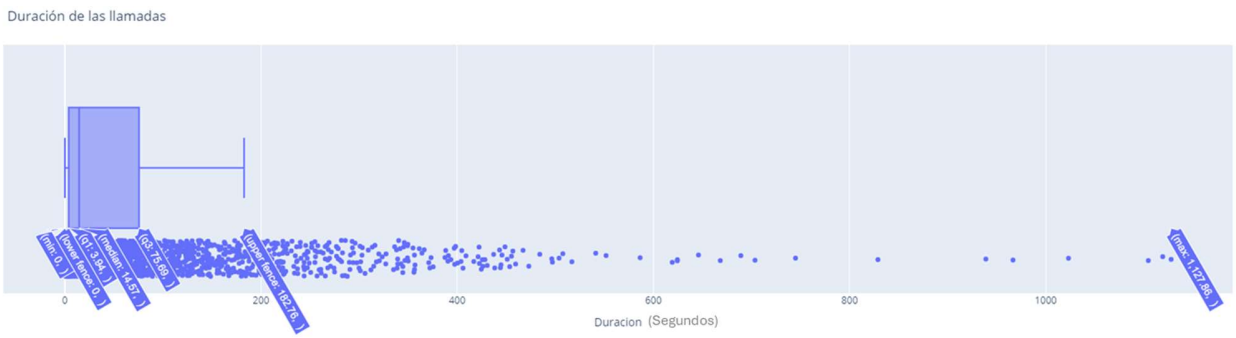



Figura 3: Diagrama de cajas y bigotes de la duración de llamadas

Basándonos en la premisa de que una llamada necesita un mínimo de 30 segundos para expresar adecuadamente las emociones, la figura anterior revela una importante limitación en la calidad de los insumos para el entrenamiento de los modelos. Se observa que un 58.9% de los audios tienen una duración inferior a 14.57 segundos, lo cual resulta insuficiente para capturar el flujo

emocional completo de una conversación, presentando un desafío significativo para el análisis de emociones.

En cuanto a los audios cuya duración supera el tercer cuartil de 75 segundos, podría considerarse la opción de fragmentarlos en segmentos más pequeños, en donde cada uno representando diferentes emociones. Sin embargo, el alcance de este proyecto está enfocado en analizar la emoción predominante de cada llamada: sin realizar fragmentaciones. En el caso de ampliar el conjunto de datos mediante la fragmentación de audios, implicaría dividir los archivos en intervalos de tiempo arbitrarios, lo cual no se contempla en el procesamiento de datos, ya que podría comprometer la integridad emocional que se busca preservar en el análisis.

### 4.1.2. Umbral del mínimo de duración de una conversación

Se lleva a cabo una categorización de la duración con el fin de identificar un umbral de tiempo que permita distinguir la duración de las llamadas de interés para nuestro proyecto. Este proceso se detalla en la Tabla 3:

Tabla 3: Grabaciones del Call Center según paquete compartido y duración en segundos

| **Duración en segundos** | **Grabaciones-1** | **Grabaciones-2** | **Grabaciones-3** | **Total** | **%Total** |
|---|---|---|---|---|---|
| De 0 – 30 seg | 265 | 703 | 718 | 1686 | 58.95 |
| De 30 – 60 seg | 51 | 148 | 93 | 292 | 10.21 |
| De 60 – 90 seg | 50 | 129 | 98 | 277 | 9.69 |
| De 90 – 120 seg | 29 | 73 | 41 | 143 | 5.00 |
| De 120 – 150 seg | 30 | 48 | 33 | 111 | 3.88 |
| De 150 – 180 seg | 16 | 27 | 24 | 67 | 2.34 |
| De 180 – 210 seg | 19 | 24 | 18 | 61 | 2.13 |
| De 210 – 240 seg | 9 | 12 | 28 | 49 | 1.71 |
| De 240 – 270 seg | 12 | 18 | 10 | 40 | 1.4 |
| De 270 – 300 seg | 6 | 9 | 12 | 27 | 0.94 |
| Mayor a 5 minutos | 42 | 31 | 34 | 107 | 3.74 |
| **Total > 30 seg** | 264 | 519 | 391 | 1174 | 41.01 |

Se observa que el umbral asumido de 30 segundos abarca el 58.95% del conjunto de datos entregado, representando un volumen de datos significativo que no debe descartarse para posibles replicaciones de clasificaciones. Por otro lado, el 41.04% de los audios supera los 30 segundos y estos se consideraron inicialmente el foco del análisis de audios.

#### 4.1.2.1. Proporción de energía entre canales

Los audios proporcionados en el dataset están en formato estéreo, lo que permite diferenciar las señales de las personas involucradas en la llamada. Para más detalles, consulte la Figura 4.

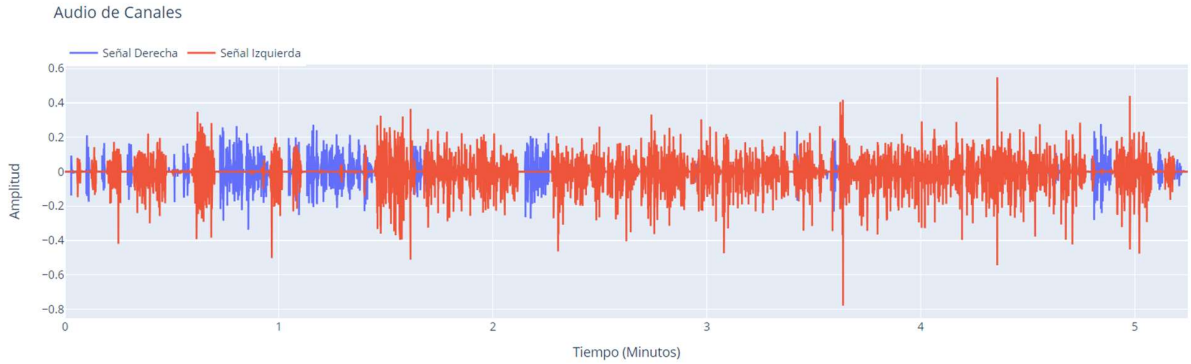



Figura 4: Señales de un audio del dataset

Esta característica permite relacionarla con la participación de cada una de las señales, característica que se pueden asociar con la **energía**. obteniendo el gráfico comparativo, Figura 5:

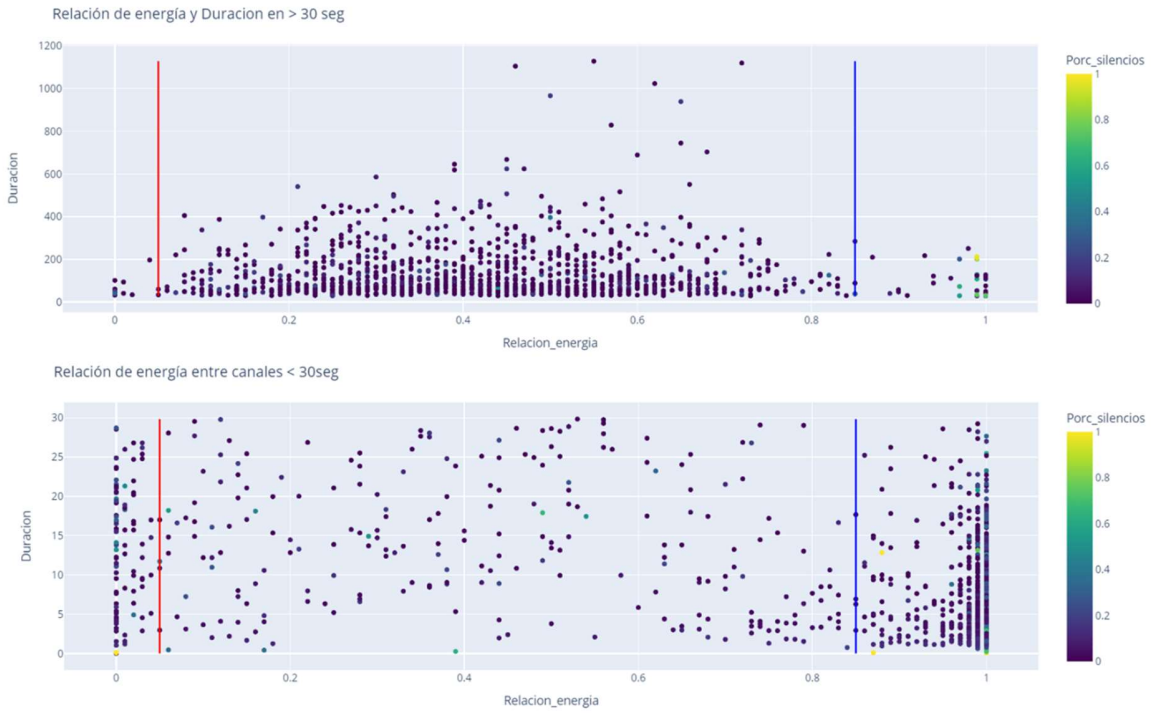



Figura 5: Comparación de relación de energía entre audios de menos 30 segundos vs audios con más de 30 segundos

Se compara la relación de energía entre los canales del audio con el umbral mínimo de 30 segundos establecido para considerar una llamada. A partir de esta comparación, se destacan las siguientes observaciones:

- Un audio de call center tiene una relación de energía entre los interlocutores que varía entre el 5% y 85% de relación con la señal del estudiante (señal derecha); según aproximaciones detectadas a partir de una selección aleatoria de los audios en regiones extremas.
- En audios de menos de 30 segundos, existen clasificaciones de audios que no son de interés para el entrenamiento, dada la ausencia de interacción entre interlocutores.
- En audios mayores de 30 segundos, se pueden observar algunos audios con comportamientos extremos en la relación de energía, dando a entender que la problemática presente en los audios de menos de 30 segundos debe ser tenida en cuenta al clasificar las emociones, surge una categoría llamada '**No llamadas**', que representa audios no relevantes para la clasificación de una llamada.
- En el gráfico se observa una barra de porcentaje de silencio, esta categoría surge bajo la premisa de que algunos audios tienen largos tiempos en donde no hay interacción. Por lo que se recurre a una nueva característica que identifica la relación de porciones de 140 ms en el audio, que tiene una amplitud menor al 0.01. detectando audios, en su mayoría con extrema relación de energía, que presentan proporciones elevadas de silencio durante el audio.

### 4.2. Características acústicas en audios

La señal de audio presenta diversas características acústicas, en las que, según el marco teórico construido, la característica que se suele emplear es el Coeficientes Cepstrales en la escala de Mel [34], para una mejor apreciación, ver el ejemplo Figura 6:

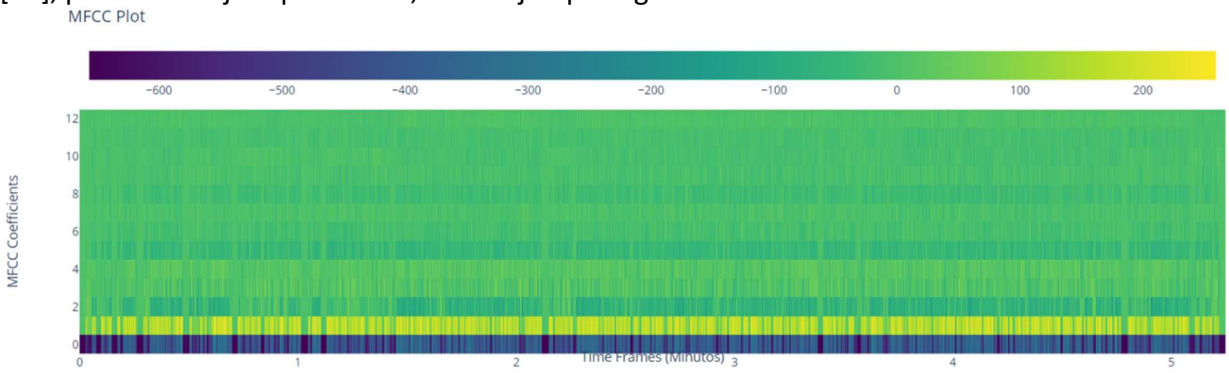



Figura 6: Gráfico de MFCC de un audio del dataset

Se observa que esta característica cuenta con la particularidad que crear ventanas de tiempo en las cuales calcula el coeficiente de Mel; es decir, una representación del espectro del sonido que se puede observar como una gráfica o imagen. Esta metodología implica varios pasos: 1. Segmentación de la señal 2. Transformada de Fourier discreta (DFT) 3. Banco de filtros Mel. 4. Logaritmo de las energías Mel. Transformada de coseno discreta (DCT) [35]. Es muy utilizada en

el procesamiento de audio y música para la representación de señales en diversos modelos, como SVM, bosques aleatorios, y modelos más complejos, como las redes neuronales convolucionales. [9]

Debido a que estas metodologías de representación de la señal requieren de unos parámetros para el cálculo de los coeficientes de mel, se plantea emplear los sugeridos en algunos trabajos similares al presente. Se decide iniciar con la representación de características acústicas, sugeridas por Diego Calvo para el trabajo de clasificación de sonido en la fase de extracción, en donde se emplean los siguientes parámetros: número de mfcc: 40 , los valores por defecto de ancho de venta a empleada por la librería librosa de Python para el calculo de MFCC y el relleno de ayudas de diversas longitudes para estandarización de la duración. [36]

### 4.3. Clasificación de emociones de audios

Dado el número limitado de audios y la falta de definición en cuanto al volumen de estos al inicio del proyecto, se optó por implementar la metodología de etiquetado no supervisado autodidacta. Esta metodología permite etiquetar los nuevos audios basándose en la clasificación establecida con los primeros. Para llevar a cabo esta metodología, se proponen las siguientes etapas:

#### 4.3.1. Etapa 1: Exploración inicial

- Crear una herramienta que permita facilitar el etiquetado y monitorear la clasificación que se tiene sobre los audios. Para ver mas detalles sobre esta herramienta por favor ver el 'Anexo 2- Dashboard para la Captura y Gestión de Sugerencias de Etiquetado de Llamadas en el Call Center'.
- Plantear una clasificación de emociones de la primera sección o paquete de audios compartidos por el call center, que consistían en 529 audios para determinar el comportamiento categorías de preclasificación y propuesta de clasificación de emociones.
- Colocar especial cuidado a las clasificaciones de audios de mayores de 30 segundos.
- Por sugerencia de los directores, se plantea iniciar con análisis de emociones complejas [37] ya que se relacionan directamente con intereses del negocio. Estas emociones propuestas son:
    - **Duda:** Todo audio en el cual el cliente presenta expresiones y tono que indica confusión, falta de claridad o que manifiesta claramente alguna duda frente al proceso.
    - **Interés:** Muestras explicitas de interés del cliente, tono emocional de la conversación que deja claro que hay una intención de buscar más información o de pasar al siguiente nivel de compromiso con el proceso.

- **Desagrado:** Es una expresión, la mayor parte de las veces expresada por los familiares del estudiante/cliente, en el sentido de criticar o expresar desacuerdo con alguna parte del proceso de admisión.
- **Desmotivado**: Planitud en el tono durante la mayor parte del audio por parte del cliente, además, expresión es verbales y paraverbales que indican desmotivación o falta de interés acerca del proceso.
- **Amabilidad**: expresiones amables durante la conversación, pero no necesariamente indicadoras de interés, probablemente relacionadas con un estado general del cliente y no relacionadas con el proceso.
- **Decepción**: expresiones de sorpresa negativa o de aterrizar expectativas que no van a ser satisfechas (Desmotivado en parte).

Para audios de '**No llamadas**' se plantean las siguientes categorías:

- **Ruido:** Llamadas sin ningún tipo de conversación o interacción humana entre los interlocutores.
- **Contestadora:** Contestadora automática.
- **Interrumpido:** Usuarios y operadora no realizan una interacción en la llamada.

- Construir modelos para las preclasificaciones, que permita tener una sugerencia en audios que no sean de interés para nuestro análisis e ir confirmando con nuevos audios.
- Tener primeros acercamientos con el modelo de clasificación y validar algunos modelos planteados en el marco teórico con aplicabilidad al enfoque de características netamente acústicas.

### 4.3.2. Resultados de implementación Etapa 1 -clasificación inicial:

Se realiza la clasificación obteniendo los siguientes resultados en audios mayores a 30 segundos, con un total de 284 audios, con la siguiente clasificación Tabla 4:

Tabla 4: Clasificación de audios Grabaciones-1 en audios de más de 30 segundos

| Categoría | Número de audios |
|---|---|
| Amabilidad | 97 |
| Duda | 69 |
| Interés | 42 |
| Decepción | 31 |
| Contestadora | 21 |

| | |
|---|---|
| Desmotivado | 9 |
| Interrumpido | 8 |
| Ruido | 4 |
| Desagrado | 3 |

Se observa un desequilibrio en las categorías y la presencia de audios correspondientes a situaciones que no involucran llamadas, lo que sugiere la necesidad de considerarlos en la agrupación de ciertas clasificaciones del modelo. Además, se plantea la implementación de un proceso de aprendizaje autodidacta para identificar estos últimos y clasificarlos preliminarmente.

### 4.3.3. Desarrollo de ventanas especiales en dash plotly para sugerencias en etiquetado:

Debido a la importancia de la generación de sugerencias de etiquetado de los audios y al número de volumen de audios que se esperaba recibir. Se desarrollo un tablero de control de alistamiento de datos, detallado en el Anexo 'Dashboard para la Captura y Gestión de Sugerencias de Etiquetado de Llamadas en el Call Center', responde a la necesidad de optimizar el etiquetado de audios y garantizar una estructura de datos consistente en este proyecto de grado. Este tablero estandariza y agiliza la recopilación de etiquetas para los audios recibidos y enviados desde el centro de llamadas, permitiendo una gestión eficaz de los datos.

Los detalles completos del tablero de control, incluyendo sus tres secciones principales:

- **Tablero de control - Vista General:** Vista general de la cantidad de audios, fechas de grabación, duración y dispersión.
- **Tablero de control - Detalle del audio:** Información detallada del audio seleccionado, incluyendo su identificación, duración, categoría de emoción, comentarios, opciones de reproducción y visualización de señal.
- **Tablero de control - Estado de calificaciones:** Estado de la clasificación por categoría y metodología (supervisada o semi supervisada), con desglose entre llamadas y no llamadas, y resumen de emociones sugeridas.

Debido a la carga de trabajo del personal del call center, el área correspondiente no revisará las sugerencias de etiquetado. Asimismo, por recomendación de los directivos y considerando la confidencialidad de los datos del proyecto, no se contó con personal especializado de la universidad para una segunda revisión o apoyo en el etiquetado. En consecuencia, las sugerencias quedarán como propuestas preliminares para los audios entregados durante el desarrollo del proyecto.

### 4.3.4. Resultados de análisis de propuesta de etiquetados

Inicialmente el proyecto contemplaba extender la clasificación semi supervisada obtenida en cada

sección previas a las nuevas, no obstante, como se verá en el capítulo 5 de modelado, debido a dificultades en los entrenamientos del modelo se requirió ampliar la base de etiquetas para obtener un modelo confiable para dicha labor, aplicando en la segunda y tercera sección. Obteniendo una clasificación supervisada a lo largo de los audios compartidos. A continuación, en la Figura 7 se aprecia la cantidad de audios por emoción sugerida al call center.

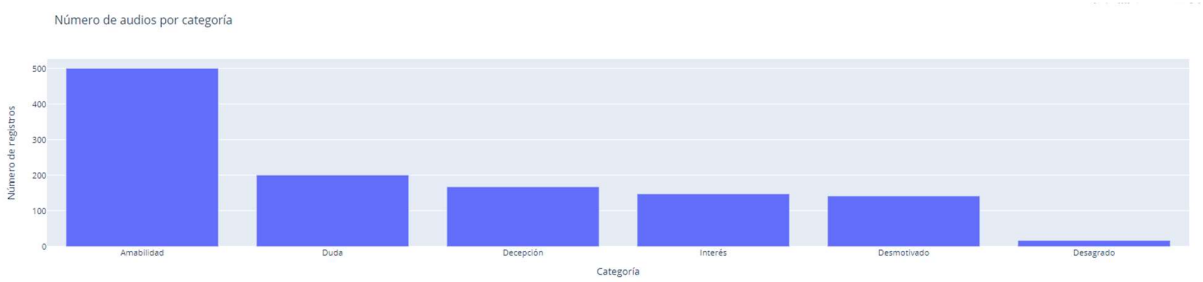



Figura 7: Tablero de control – Estado de calificaciones

La emoción predominante en los audios del call center es ‘Amabilidad’, con un total de 501 registros, seguida por Duda (201), Decepción (168), Interés (148), Desmotivación (142) y, finalmente, Desagrado (17). Debido al desbalance de esta última categoría en comparación con el resto, se propone agruparla con la emoción más cercana en características descriptivas, Decepción (ver Figura 7).

# 5. DESARROLLO DEL MODELO MACHINE LEARNING PARA DETECCIÓN DE EMOCIONES

Una vez se finalizó con el procesamiento de los audios y la sugerencia de etiquetado, se procede con el siguiente paso para el desarrollo del modelo: la identificación de la mejor representación del audio seguido de la identificación del mejor modelo de clasificación. Y para poder realizar una comparación de las representaciones se requiere del desarrollo de un modelo inicial para su comparación, se plantea iniciar con un modelo CNN con pocas capas, inspirados en las referencias de trabajos con audios que se encontraron en puntos posteriores.

Tabla 5: Modelos de clasificación empleado en comparación de representaciones

| Bloques | Tipo de Capa | Filtros / Neuronas | Kernel size | Activación | Pooling | Pool_size |
|---|---|---|---|---|---|---|
| Bloque 1 | Convolución | 16 | (2) | relu | MaxPooling 2D | 2 |
| Bloque 2 | Convolución | 32 | (2) | relu | MaxPooling 2D | 2 |
| Bloque 3 | Flatten | | | | | |
| Bloque 4 | Densa | 32 | | relu | | |
| Bloque 5 | Densa | 16 | | relu | | |
| Bloque 6 | Densa – salida | # Categorías | | softmax | | |

El modelo CNN empleado en la comparación de representaciones está compuesto por dos capas convolucionales con 16 y 32 neuronas, respectivamente, seguidas de una capa de aplanamiento (flatten) y dos capas densas con un número de neuronas similar al de las capas convolucionales anteriores (ver Tabla 5).

## 5.1. Identificando la representación de la mejor representación del audio modelo call center

Se plantean las cuatro representaciones mencionadas en el marco teórico, específicamente en el capítulo '3.1.1 Características del Audio', y se comparan utilizando el esquema CNN descrito al inicio de este capítulo. El objetivo es identificar, con la misma selección de audios y la misma división entre los conjuntos de entrenamiento y validación, cuál representación proporciona el mejor rendimiento para continuar con el desarrollo del modelo.

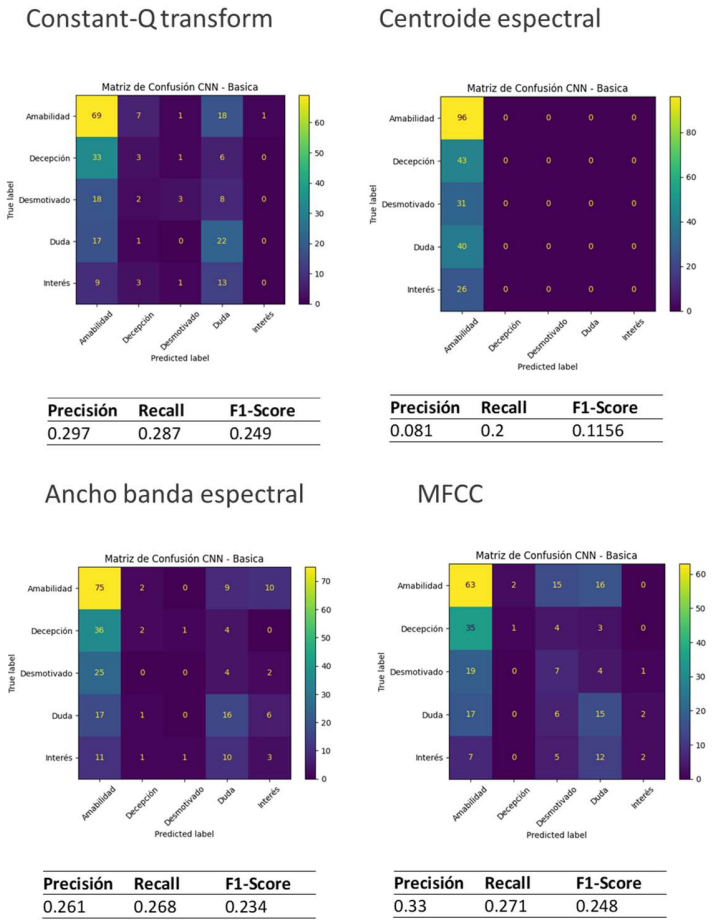



Figura 8: Matriz de confusión e indicadores de diferentes presentaciones del audio.

La representación MFCC alcanza la mayor precisión en comparación con las demás representaciones, destacándose como la opción más efectiva. Además, es la más utilizada en la industria para el procesamiento de audio (ver Figura 8).

Se debe mencionar que, aunque se consideraron representaciones que trabajaban con los datos en su forma más pura, como el Ventaneo de la señal y la Transformada Discreta de Wavelet (DWT), se decidió no utilizarlas en el proyecto debido a que excedían las capacidades de cómputo disponibles para el procesamiento (capacidad máxima de 300 GB de RAM en una TPU de Google).

### 5.2. Esquema de funcionamiento del modelo call center

En la Figura 9. se plantea el esquema del modelo a implementar en el Call center. Este modelo se implementaría en el aplicativo de interfaz que se entregaría al Call center de la Universidad Javeriana de Cali

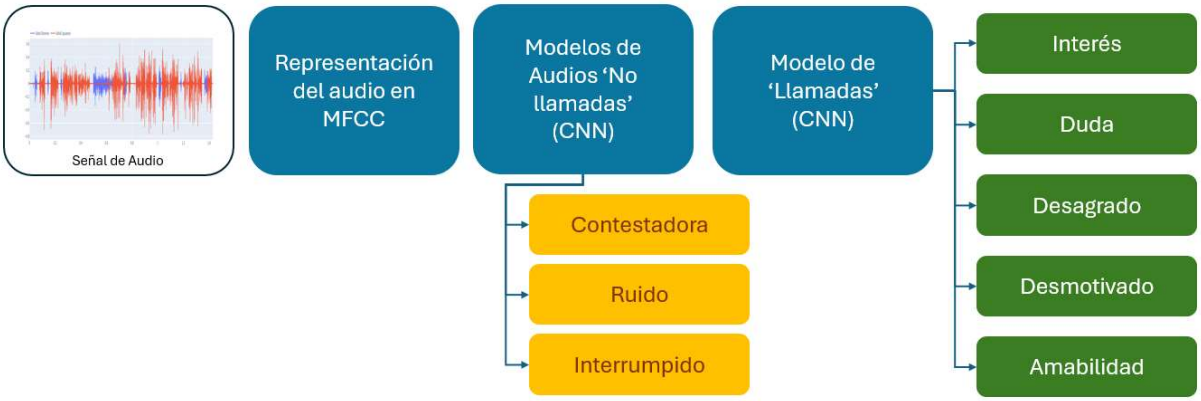



Figura 9: Modelo planteado para el análisis de emociones en audios del Call center

En esta figura, se presenta un proceso detallado que comienza con la fase de representación de las señales de audio utilizando la técnica MFCC. A continuación, se aplica un modelo de depuración para identificar y depurar las 'No llamadas', asegurando que únicamente se conserven los audios correspondientes a 'Llamadas' reales. Estos audios depurados son luego utilizados como entrada para el modelo de clasificación de emociones, que tiene como objetivo categorizar las interacciones en cinco emociones complejas: Interés, Duda, Desagrado, Desmotivación y Amabilidad.

### 5.3. Etapa 1: Modelos Exploración inicial

Se inicia la etapa de modelos etapa 1, con la Tabla6, para la implementación de la metodología semi supervisada y extensión del etiquetado en esta etapa.

Tabla 6: Estado de clasificación del dataset para etapa1

| Categoría | Audios revisados |
|---|---|
| Contestadora | 81 |
| Interrumpido | 47 |
| Llamada | 267 |
| Pendiente | 1336 |
| Ruido | 26 |

#### 5.3.1. Modelos de preprocesamiento em audios 'No llamadas'

Dado la importancia de contar con una depuración de las llamadas a análisis emociones, se crea un preprocesamiento antes de la clasificación de emociones a modo de limpieza de datos. Este proceso consiste en a partir de modelos enfocados a la detección de audios sin ningún tipo de conversación, se plantea un modelo que trabaje con características de coeficientes MFCC planteadas por trabajos similares mencionados en el capítulo 4.2 y un modelos como el de

RandomForest que no tiene problemas con los desbalanceo. Obteniendo los siguientes resultados con modelos sin ajustar hiper parámetros:

Tabla 7: Modelos empleados en audios 'No llamadas'

| Modelo | Categoría | Precisión |
|---|---|---|
| Random Forest | Contestadoras | 0.9523 |
| Random Forest | Ruido | 0.9889 |
| Random Forest | Interrumpido | 0.9667 |

En los resultados presentados en la Tabla 7, se evidencia que el modelo alcanzó una alta precisión, situándose entre un 95% y 98%. Con base en estos resultados, se procedió a utilizar dicho modelo para predecir los audios en las etapas subsecuentes, implementándolo como un enfoque semi supervisado para la detección de audios que corresponden a 'No llamadas'. Este proceso permitió una depuración más eficiente de los datos, asegurando que el conjunto de audios analizados contuviera solo las interacciones relevantes para el análisis emocional.

### 5.3.2. Modelos preliminares en secciones 1 y 2 de categorías de emociones

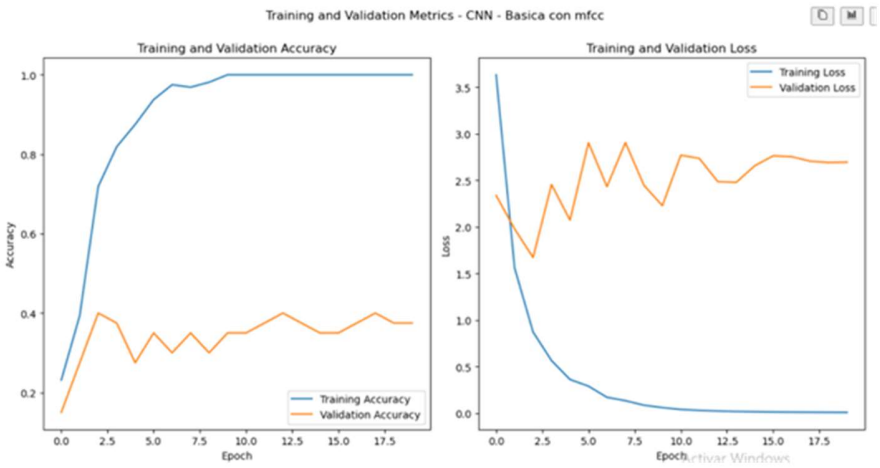
En los modelos de categorías de emociones se realizan diversas arquitecturas recomendadas por el marco teórica con miras a identificar una aproximación que tuviera resultados aceptables para realizar modelos semi supervisados. Se inicio con los modelos de redes neuronales planteados en las siguientes figuras: Figura 10, 11 y 12.


| Bloques | Tipo de Capa | Filtros / Neuronas | Kernel size | Activación |
|---|---|---|---|---|
| Bloque 1 | Flaten | | | |
| Bloque 2 | Densa | 64 | | relu |
| Bloque 3 | Densa | 32 | | relu |
| Bloque 4 | Densa - salida | # Categorías | | softmax |

Figura 10: Modelo de red Neuronal RNN con desempeño en clasificación emociones sección 1


| Bloques | Tipo de Capa | Filtros / Neuronas | Kernel size | Activación |
|---|---|---|---|---|
| Bloque 1 | LSTM | | | |
| Bloque 2 | LSTM | 32 | | |
| Bloque 3 | Densa - salida | # Categorías | | softmax |

Figura 11: Modelo de red Neuronal LSTM con desempeño en clasificación emociones sección 1




| Bloques | Tipo de Capa | Filtros / Neuronas | Kernel size | Activación | Pooling | Pool_size |
|---|---|---|---|---|---|---|
| Bloque 1 | Convolución | 32 | (2) | relu | MaxPooling 2D | 2 |
| Bloque 2 | Convolución | 64 | (2) | relu | MaxPooling 2D | 2 |
| Bloque 3 | Flaten | | | | | |
| Bloque 4 | Densa | 64 | | relu | | |
| Bloque 5 | Densa | 32 | | relu | | |
| Bloque 6 | Densa - salida | # Categorías | | softmax | | |

Figura 12: Modelo de red Neuronal CNN con desempeño en clasificación emociones sección 1

A pesar de que la estructura de la red no era muy compleja ni contaba con una gran cantidad de neuronas, el modelo no lograba converger, como se evidencia en la divergencia entre las curvas de entrenamiento y validación a lo largo de las épocas (ver Figuras 10, 11 y 12). Además, la precisión máxima alcanzada por los modelos no supera el 45 % en el conjunto de prueba, lo que indica dificultades significativas para clasificar las emociones de manera efectiva.

| Bloques | Tipo de Capa | Filtros / Neuronas | Kernel size | Activación | Pooling | Pool_ size |
|---|---|---|---|---|---|---|
| Bloque 1 | Convolución | 64 | (3) | | Max_Pooling2D | 2 |
| Bloque 2 | Convolución | 64 | (3) | | Max_Pooling2D | 2 |
| Bloque 3 | Convolución | 128 | (3) | | | |
| Bloque 4 | Flatten | | | | | |
| Bloque 5 | Densa | 128 | | Relu | | |
| Bloque 6 | Densa -Salida | # Categorias | | softmax | | |

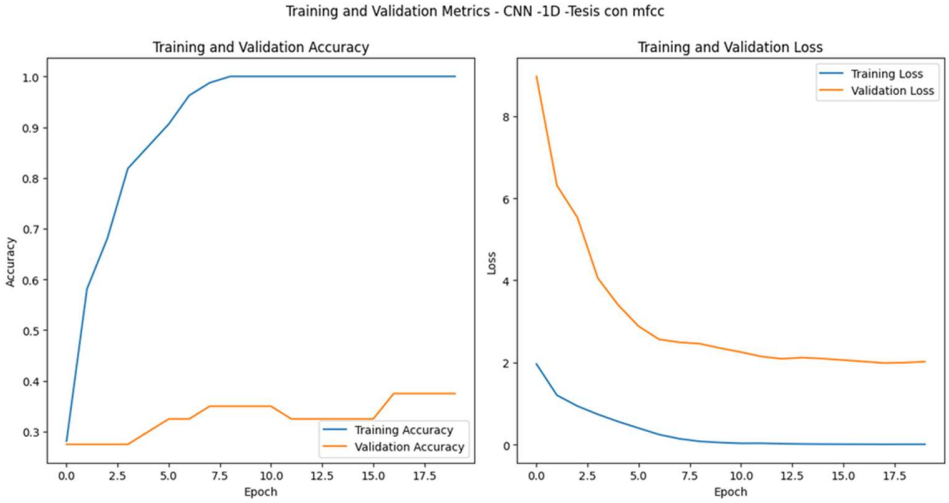



Figura 13: Modelo Redes Neuronales – clasificación emociones sección 2

| Bloques | Tipo de Capa | Filtros / Neuronas | Kernel size | Activación | Pooling | Pool_ size |
|---|---|---|---|---|---|---|
| Bloque 1 | Convolución | 32 | (3) | Relu | Max_Pooling2D | 2 |
| Bloque 2 | Convolución | 64 | (3) | Relu | Max_Pooling2D | 2 |
| Bloque 3 | Flatten | | | | | |
| Bloque 4 | Densa | 128 | | Relu | | |
| Bloque 5 | Dropout | 0.5 | | | | |
| Bloque 6 | Densa -Salida | # Categorias | | softmax | | |

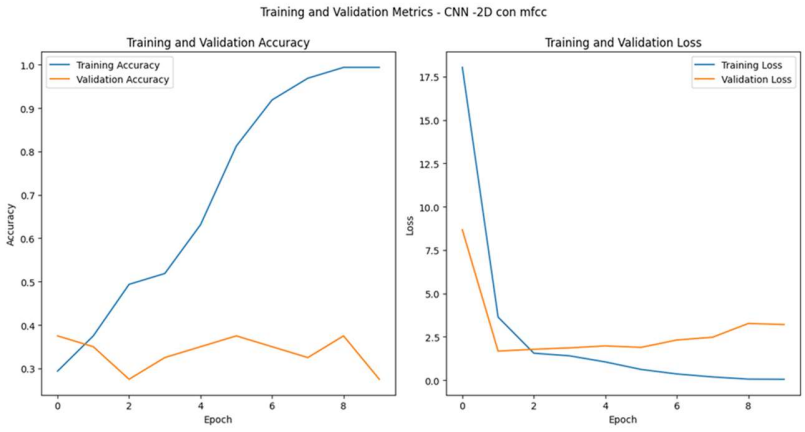



Figura 14: Modelo Redes Neuronales – clasificación emociones sección 3

Los modelos presentados en las Figuras 13 y 14, se presentan dos modelos de redes convolucionales con estructuras más complejas en comparación con el primer modelo. Estos modelos incorporan métodos de normalización, capas de flatten tras las capas convolucionales, y el uso de dropout en las capas densas para reducir la divergencia. Sin embargo, los resultados mostraron que los modelos no lograron converger, manteniendo valores de pérdida (loss) muy elevados, superiores a 1. Esto llevó a la decisión de no continuar con una metodología semi supervisada para la clasificación, lo que obligará a implementar un proceso de etiquetado manual en las siguientes etapas o secciones de los audios.

### 5.3.3. Evaluación de modelos – Etapa 1

Con base en las clasificaciones de emociones frente al dataset de llamadas etiquetas de modo supervisado, se evidencia que el modelo tiene problemas de extracción de datos frente a las categorías de emociones, esto plantea dos posibilidades:

- Las emociones como lo plantea a la psicolingüística emocional, desde características acústicas esta faltante de elementos para una buena representación
- Se requiere mejorar los modelos a través de mejoramiento de características del modelo tales como:
    - Un mayor número de audios, que mejoren el aprendizaje del modelo.
    - Una mejora en la representación de la señal ya sea incluyendo más canales al MFCC, cambiando los parámetros que otorga la librería Librosa u obteniendo una representación mejor para la representación de la señal.
    - Una mejor arquitectura de los modelos para que aprendan las clasificaciones de emociones brindadas.
    - Revisando la calidad de etiquetado sugerido.

## 5.4. Comparación de diversos modelos de machine learning de clasificación múltiple

Para mejorar el desempeño observado en las pruebas con la metodología semi supervisada, se decidió evaluar 16 modelos de clasificación múltiple. El objetivo era determinar si la arquitectura o el modelo utilizado para la clasificación no era el más adecuado. Los resultados obtenidos se documentaron en la Tabla 8.

Tabla 8: Modelos de clasificación en audios llamada

| Modelo | Accuracy | Loss |
|---|---|---|
| Lenet 5 | 0.434 | 1.483 |
| SVM | 0.430 | - |
| CNN Diego Calvo | 0.430 | 1.459 |

| | | |
|---|---|---|
| Gradient Bossting | 0.430 | - |
| Random Forest | 0.407 | - |
| VGG16 | 0.398 | 1.493 |
| Fully connect | 0.398 | 1.493 |
| Regresión Logistica | 0.367 | - |
| CNN 3 capas | 0.367 | 3.681 |
| RNN LSTM | 0.330 | 2.734 |
| Árbol de decisión | 0.330 | - |
| RNN | 0.294 | 9.410 |
| Alex net | 0.290 | 10.420 |
| Bayes | 0.249 | - |
| KNN | 0.240 | - |
| Perceptron | 0.213 | - |

Se aprecia que los modelos evaluados para las emociones en llamadas no superan el 45% de precisión, en donde se destacan los modelos como: CNN Lenet -5, un modelo lineal de SVM y el modelo del científico de datos Diego Calvo [36]. Llama la atención el valor de perdida al final del entrenamiento de ambos modelos, con valores de los que superan los 1.4 en perdida, dando a entender que no son modelos confiables.

En búsqueda de tener una comparación con este ejercicio, se plantea compararlos empleando las pre categorías o preprocesamiento de las no llamadas con la identificación de llamadas reales de estudiantes y Call center. Obteniendo la Tabla 9.

Tabla 9: Modelos de clasificación en audios No llamada

| **Modelo** | **Accuracy** | **Loss** |
|---|---|---|
| CNN 3 Capas | 0.940 | 0.633 |
| Randomforest | 0.937 | - |
| SVM | 0.928 | - |
| GradientBoosting | 0.926 | - |
| AlexNet | 0.926 | 0.360 |

| | | |
|---|---|---|
| CNN Diego Calvo | 0.926 | 0.207 |
| Lenet 5 | 0.918 | 0.387 |
| Árbol de decisión | 0.902 | - |
| Regresión logística | 0.900 | - |
| Perceptron | 0.886 | - |
| KNN | 0.884 | - |
| Fully connect | 0.881 | 0.335 |
| VGG16 | 0.877 | 0.383 |
| RNN_LSTM | 0.860 | 0.654 |
| Bayes | 0.849 | - |
| RNN | 0.818 | 1.291 |

Se observa que todos los modelos presentan precisiones superiores al 80%. Los modelos con menor pérdida son los mismos que destacaron en el ejercicio con llamadas: el modelo del científico de datos Diego Calvo, el modelo CNN Lenet-5 y el modelo SVM lineal. Cabe destacar que estos modelos tienen precisiones superiores al 92%.

A continuación, se presenta en detalle estos 3 modelos, su arquitectura y los gráficos de pérdida y matriz de confusión.

### 5.4.1. Modelo CNN Lenet - 5

En la Tabla 10, se puede apreciar la arquitectura que compone esta red neuronal, en donde destaca 3 capas convolucionales, un flatten y dos capas densas.

Tabla 10: Arquitectura del CNN lenet -5

| Bloques | Tipo de Capa | Filtros / Neuronas | Kernel size | Activación | Pooling | Pool_ size | Paso |
|---|---|---|---|---|---|---|---|
| Bloque 1 | Convolución | 6 | (5) | tanh | AveragePooling 2D | 2 | 2 |
| Bloque 2 | Convolución | 16 | (5) | tanh | AveragePooling 2D | 2 | 2 |
| Bloque 3 | Convolución | 120 | (5) | tanh | | | |
| Bloque 4 | Flatten | | | | | | |
| Bloque 5 | Densa | 84 | | tanh | | | |
| Bloque 6 | Densa -Salida | # Categorias | | softmax | | | |

Los desempeños obtenidos por este modelo para los audios de llamadas para las clasificaciones de las emociones fueron los que se aprecian en la Figura 15.

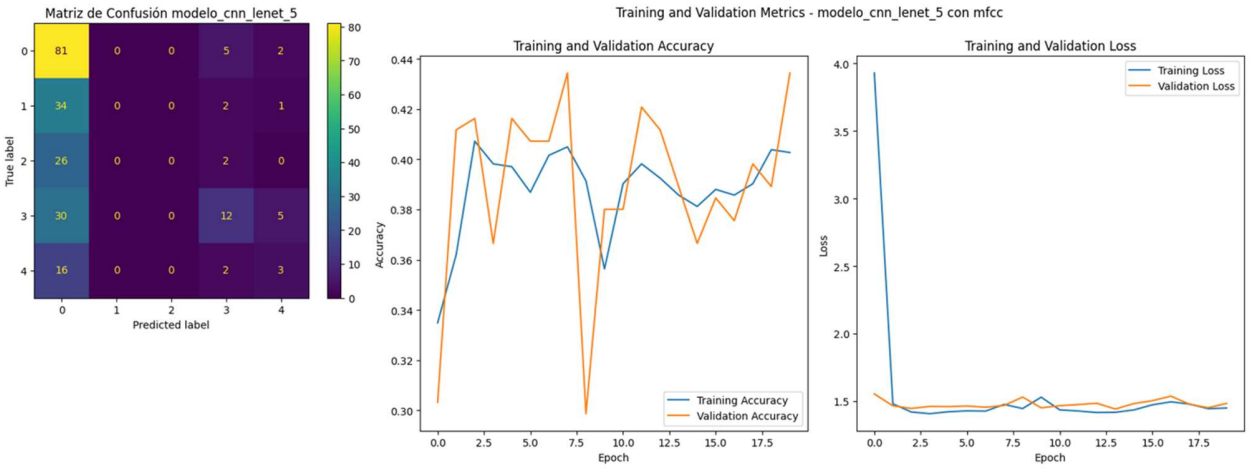



Figura 15: Modelo CNN Lenet -5 con llamada

El modelo presenta una pérdida elevada y no mejora a partir de la época 3, con una precisión cercana al 40% y una matriz de confusión que muestra dificultades para diferenciar correctamente entre la etiqueta 0 – Amabilidad y las demás emociones.

Si revisamos los mismos gráficos, pero con los audios de las no llamadas y su clasificación de pre categorías, observamos la Figura 16.


Figura 16: Modelo CNN Lenet -5 con No Llamadas

En este modelo se observa un overfiting que puede aplicarse metodologías de regularización para mejor su aprendizaje. La matriz de confusión muestra un buen desempeño en general, con excepción de las categorías de interrumpido y Llamada, pues el modelo no es capaz de diferenciarlos con buena precisión.

### 5.4.2. Modelo SMV

En el modelo de SMV se empleó la mayoría de hiper parametros que trae la librería de sklearn por defecto con un kernel 'linear'. Obteniendo las siguientes matrices de confusión para los datos de llamadas y no llamadas, apreciadas en la Figura 17.

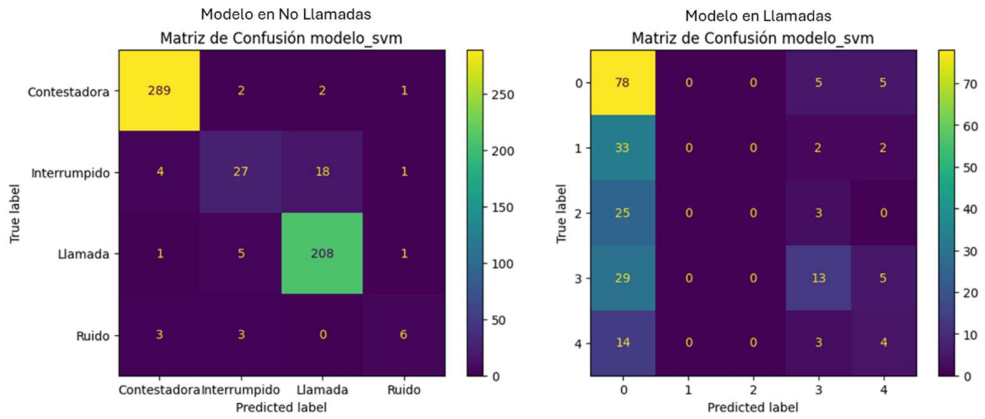



Figura 17: Modelo SVM linear para 'llamadas' y 'No Llamadas'

En la matriz de confusión de llamadas vemos que se repite la dificultad de identificar las emociones 0 – Amabilidad con el resto de las emociones. Frente al modelo SVM con las 'No llamadas' vemos que tiene una buena clasificación y que frente al modelo Lenet -5 vemos una mejora en la precisión entre la clasificación e Interrumpido y llamada.

### 5.4.3. Modelo CNN por científico datos Diego Calvo

La arquitectura empleada en el trabajo de clasificación de audios por el científico de datos Diego Calvo empleo la siguiente arquitectura [36] que se aprecia en la tabla Tabla 11.

Tabla 11: Arquitectura del CNN por Diego Calvo

| Bloques | Tipo de Capa | Filtros / Neuronas | Kernel size | Activación | Pooling | Pool_size | Dropout |
|---|---|---|---|---|---|---|---|
| Bloque 1 | Convolución | 16 | (2) | relu | MaxPooling 2D | 2 | 0.2 |
| Bloque 2 | Convolución | 32 | (2) | relu | MaxPooling 2D | 2 | 0.2 |
| Bloque 3 | Convolución | 64 | (2) | relu | MaxPooling 2D | 2 | 0.2 |
| Bloque 3 | Convolución | 128 | (2) | relu | MaxPooling 2D | 2 | 0.2 |
| Bloque 4 | GlobalAveragePooling | | | | | | |

| | | | |
|---|---|---|---|
| Bloque 5 | Densa - salida | # Categorías | softmax |

Esta arquitectura emplea 4 capas convolucionales, un global average pooling y una capa densa. Los resultados que se obtuvieron en el dataset de 'llamadas' son los apreciados en la Figura 18.


Figura 18: Modelo CNN por Diego Calvo en llamada

En el gráfico de pérdida, se evidencia que el modelo no muestra un overfiting, pero con un loss tan alto, el modelo no está aprendiendo, que es justo lo que confirma el gráfico de pérdida, quedando en la precisión 40%. La matriz de confusión muestra que el modelo no es capaz de diferenciar la etiqueta 0 – Amabilidad de las demás emociones.

Si se compara con el dataset de No llamadas, se observa la Figura 19.

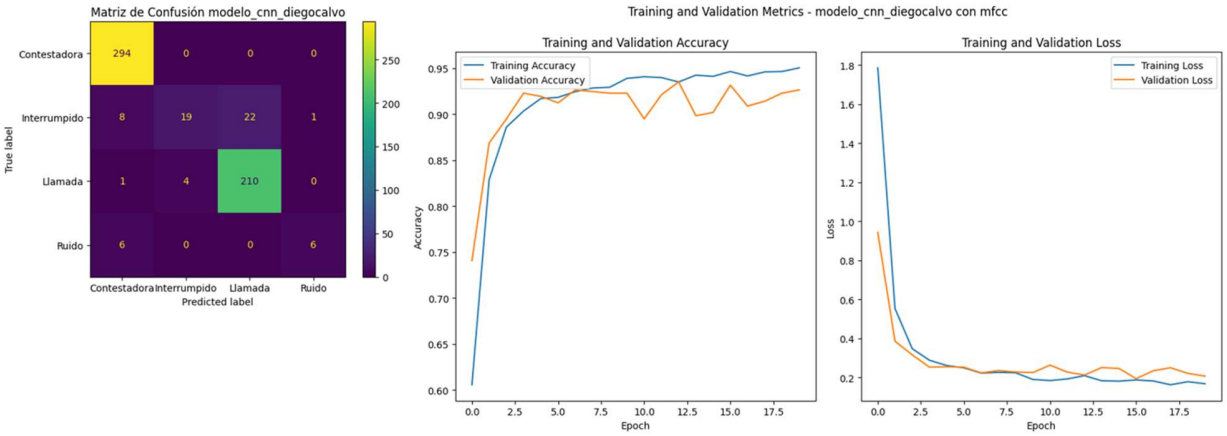



Figura 19: Modelo CNN por Diego Calvo en No llamada

En este modelo, la gráfica de pérdida muestra un normal comportamiento y un modelo de precisión que ronda los 90% - 95% de precisión. Se infiere que el modelo tiene problemas en la separación de interrumpido con la Llamada.

### 5.5. Entrenamiento del modelo con los mejores audios por emoción y representación en estéreo

Dado que el dataset tiene un etiquetado sugerido, se adoptó la recomendación de seleccionar los mejores 20 audios por categoría que mejor representen las emociones en llamadas. Esto se hizo para mitigar problemas de etiquetado y evaluar si esta selección influye en la representación de la emoción como señal. Se observa en la Tabla 12.

Tabla 12: Registro de audios elite por emoción que mejor representan cada categoría

| Categoría | Número audios |
|---|---|
| Desmotivado | 20 |
| Duda | 20 |
| Decepción | 20 |
| Amabilidad | 20 |
| Interés | 20 |

Se inicia con el modelo de CNN Diego Calvo para entrenar inicialmente estos datos, dado su buen desempeño en el modelo general. Obteniendo el entrenamiento que se aprecia en la Figura 20

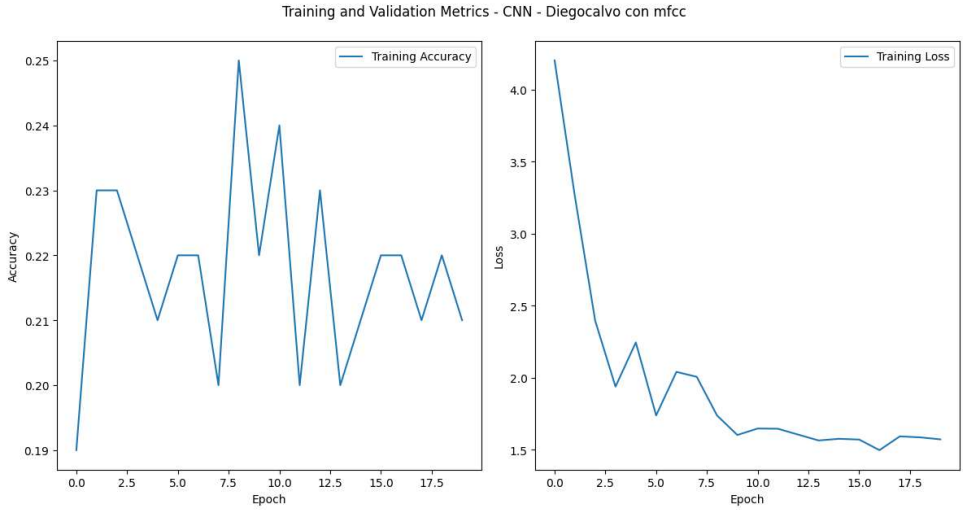



Figura 20: Modelo de audios elite representación mono con mfcc 40

El modelo presenta un valor de pérdida elevado, superior a '1.5'. Para mejorar esta representación, se emplea una modificación en la extracción, empleando una representación en estéreo que aumenta el número de atributos al doble dado los dos canales existentes en los audios. Obteniendo la Figura 21.

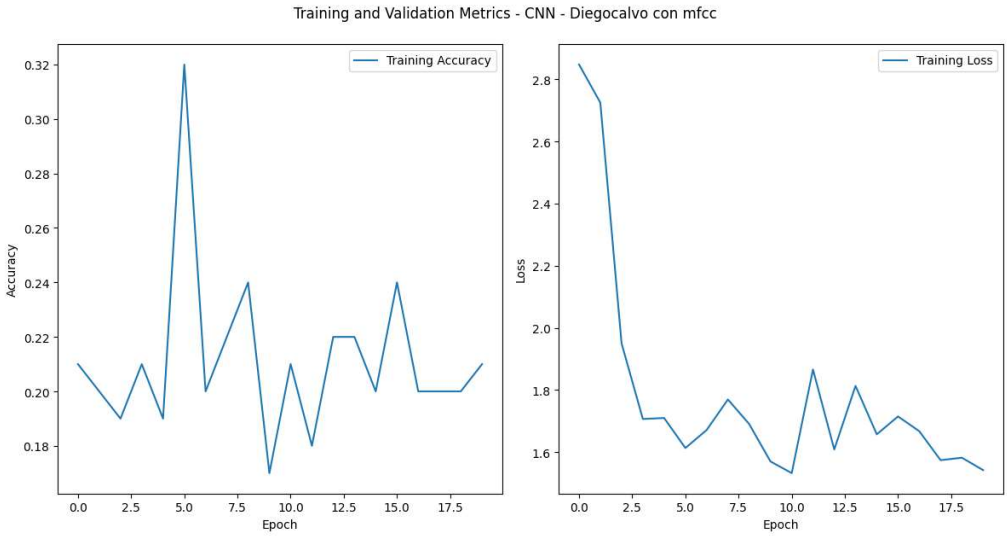



Figura 21: Modelo de audios élite representación estereo mfc 40

El desempeño general de la extracción mono se mantiene, con una pérdida superior a 1.6 y una precisión promedio del 22%.

A continuación, se observa una nueva prueba cambiando a la representación de la señal en el número de coeficientes MFCC, empleando 20 y 13. En la Figura 22.

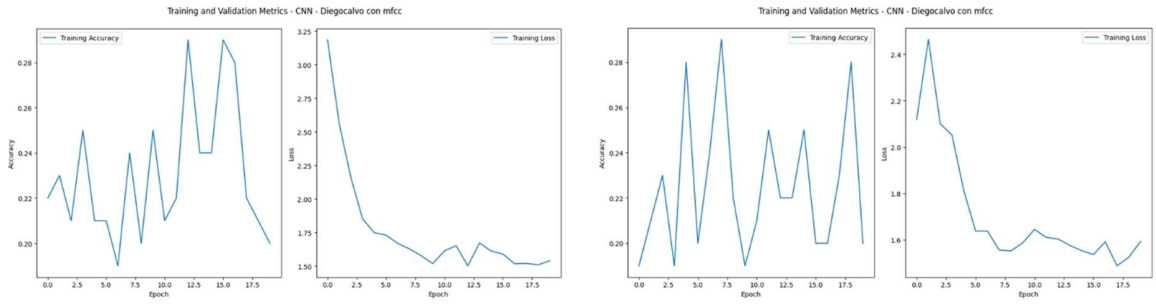



Figura 22: Modelo de audios elite representación mono y estéreo mfcc 20

Se observa los entrenamientos tanto para Mono como para estéreo. El comportamiento de alta pérdida y una precisión del 24% se mantiene constante con este dataset.

En la Figura 23, se observa una mejora la precisión del modelo uy una pedida que baja muy lentamente a medida que se incrementan las épocas. Cabe mencionar, que la representación dio resultados fue la estéreo con una pérdida más estable, aunque igualmente alta.

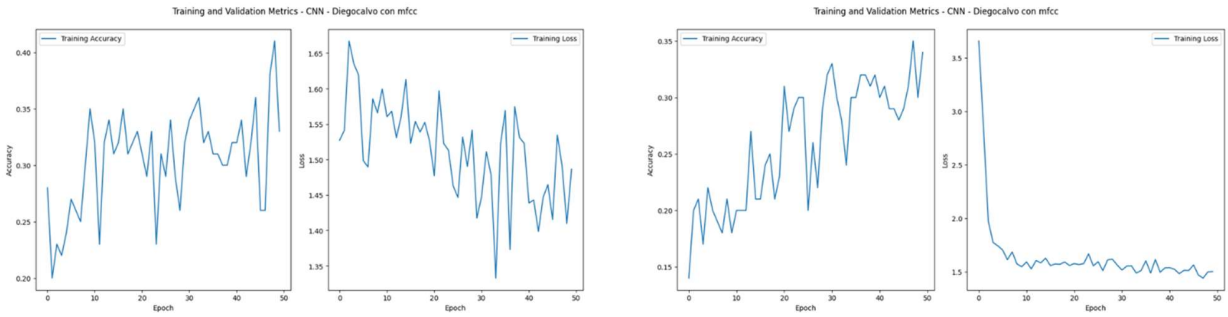



Figura 23: Modelo de audios elite representación mono y estéreo mfcc 13

Se decide emplear otra arquitectura diferente, pero que haya sido empleada con éxito en problemas de clasificación de emociones básicas, encontrando el empleado en un proyecto de grado de Francisco Pastor Naranjo [38] que empleo la siguiente arquitectura ilustrada en la Tabla 13.

Tabla 13: Arquitectura del CNN por Francisco Naranjo

| **Bloques** | **Convolución** | | | **Pooling** | |
|---|---|---|---|---|---|
| | **Filtros** | **Kernel_size** | **Paso** | **Kernel_size** | **Paso** |
| Bloque 1 | 64 | (3,3) | (1,1) | (2,2) | (2,2) |
| Bloque 2 | 64 | (3,3) | (1,1) | (2,2) | (2,2) |
| Bloque 3 | 128 | (3,3) | (1,1) | (2,2) | (2,2) |
| Bloque 4 | 128 | (3,3) | (1,1) | (4,4) | (4,4) |

En esta estructura se evidencia 4 capas convolucionales, un flatten y una capa densa softmax, generando el siguiente entrenamiento que se puede ver en la Figura 24.

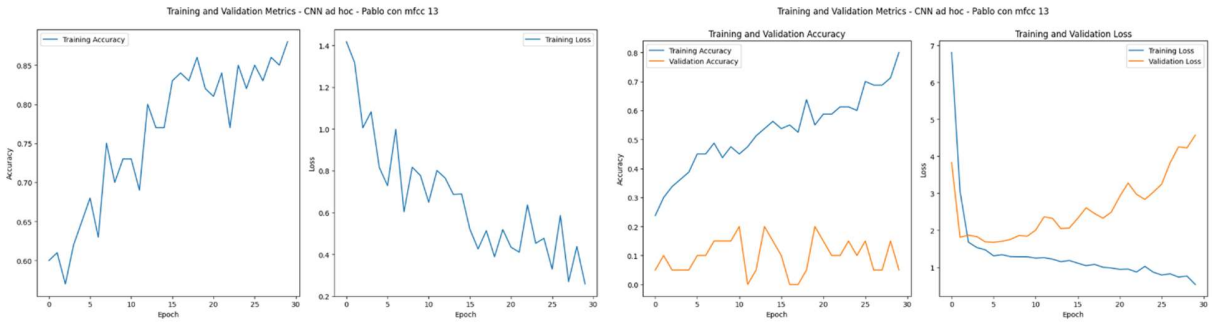



Figura 24: Modelo de audios elite representación estéreo mfcc 13 CNN Fráncico Naranjo

El modelo tanto en solo training como en la versión con validación del 20%, tuvo un overiting que ocasionaba que mostrara una precisión alta, con una perdida baja en el entrenamiento, pero en con los datos de Test no tuviera buenos comportamientos.

### 5.6. Búsqueda de hiper parámetros para la representación de la señal

El proceso de búsqueda de hiper parámetros se lleva a cabo en dos fases. La primera fase se enfoca en la optimización de los hiper parámetros de la representación MFCC, utilizando el mismo modelo inicial empleado para la comparación de representaciones como punto de referencia. Una vez identificados los mejores valores para los hiper parámetros de la representación, se procede a la segunda fase, que consiste en la búsqueda de los hiper parámetros óptimos para el modelo de clasificación.

Debido a que los resultados obtenidos con diversos modelos de clasificación no fueron satisfactorios, se decidió ampliar el enfoque probando con los tres modelos más prometedores de la fase de 'No llamadas'. Se realizó una búsqueda de hiper parámetros en estos tres modelos con el objetivo de mejorar el rendimiento general y obtener resultados más precisos en la clasificación de las emociones presentes en los audios.

#### 5.6.1. Representación de la señal MFCC

En relación con los hiper parámetros de la representación MFCC, se propone evaluar utilizando los siguientes parámetros de la Tabla 14:

Tabla 14: Hiper parámetros de representación de la señal en MFCC

| Hiper parámetro de representación | Valores propuestos a evaluar |
|---|---|
| n_mfcc (Número de coeficientes MFCC): | 13, 20 y 40 |
| largo de ventana (Window length): | 512, 1024, 2048 |
| Hop_length_relacion<br>(hop_length = largo ventana //<br>Hop_length_relacion) | 2,3,4 |

| | |
|---|---|
| n_mels (Número de bandas de filtro de Mel): | 40, 64, y 128 |
| Dct_type : | 1,2 y 3 |
| Representación: | Monofónica y estero fónica |

Cada una de las representaciones se compara con un mismo modelo CNN y se extrae los resultados obtenidos de accuracy y loss para su comparación. La estructura del modelo CNN empleado cuenta con pocas capas densas y convolucionales con pocas neuronas, como se muestra en la siguiente Tabla 15.

Tabla 15: Modelo CNN con pocas capas para comparación de representaciones MFCC

| **Bloques** | **Tipo de Capa** | **Filtros / Neuronas** | **Kernel size** | **Activación** | **Pooling** | **Pool_size** | **Dropout** |
|---|---|---|---|---|---|---|---|
| Bloque 1 | Convolución | 16 | (2) | elu | MaxPooling 2D | 2 | 0.2 |
| Bloque 2 | Convolución | 32 | (2) | Relu | MaxPooling 2D | 2 | 0.2 |
| Bloque 3 | Flatten | | | | | | |
| Bloque 3 | Densa | 32 | | Relu | | | |
| Bloque 4 | Densa | 16 | | Relu | | | |
| Bloque 5 | Densa - salida | # Categorías | | softmax | | | |

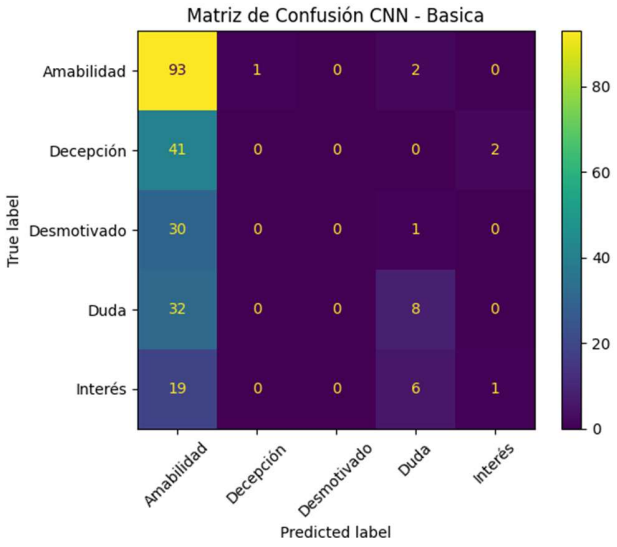
Las comparaciones de las diferentes representaciones de las señales dieron los siguientes resultados observados en Tabla 16:

Tabla 16: Comparación de representación de señal con MFCC

| **Posición** | **Accuracy** | **Loss** | **Representación** | **N_mfcc** | **N_fft** | **Hop_length relación nfft** | **N_mels** | **Dct_type** |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.4364 | 2.9413 | Estéreo | 13 | 512 | 2 | 40 | 2 |
| 0 | 0.4110 | 3.6512 | Mono | 13 | 512 | 2 | 128 | 2 |
| 4 | 0.4110 | 2.1610 | Estéreo | 13 | 512 | 2 | 64 | 2 |
| 6 | 0.4110 | 2.3203 | Mono | 13 | 512 | 4 | 128 | 1 |
| 2 | 0.4110 | 3.5127 | Estéreo | 13 | 512 | 4 | 128 | 2 |
| 24 | 0.4067 | 1.5489 | Mono | 40 | 2045 | 2 | 128 | 2 |
| 5 | 0.4067 | 1.4365 | Mono | 13 | 512 | 2 | 64 | 3 |
| 11 | 0.4067 | 1.5491 | Mono | 20 | 512 | 4 | 128 | 2 |
| 12 | 0.4067 | 1.5422 | Mono | 20 | 1024 | 2 | 128 | 2 |
| 13 | 0.4067 | 1.5449 | Mono | 20 | 1024 | 3 | 128 | 2 |

La mejor representación es Estereo con **n_mfcc=13, n_mels=2 ,dct_type = 2, n_fft=512 y un hop_length_relación_nfft = 2**, esto quiere decir que el stride = 512/2 = 256 . bajo este modelo se observa la siguiente Figura en matriz de confusión y curvas de precisión y pérdida.

Tabla 17: Matriz de confusión CNN básica con mejor representación




Se evidencia en Tabla 17, bajo este modelo CNN con pocas capas, no es capaz de diferencias las diferentes emociones de los modelos, lo que se refleja con la precisión baja del 0.436. Ahora veamos los siguientes gráficos de entrenamiento y validación durante en entrenamiento:

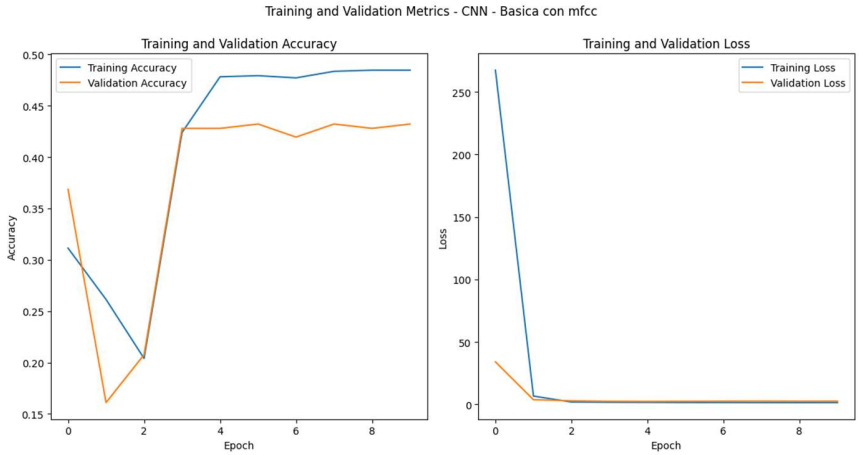



Figura 25: Modelo de CNN básica con mejor representación

En la Figura 25, observamos la baja precisión que observamos en la matriz de confusión y un alto loss en el grafico que no disminuye con el paso de las épocas: 4.3. lo que nos permite concluir que a pesar de que nuestra representación tiene la mejor precisión y un bajo valor de loss frente al resto de representaciones, no es una buena representación o al menos bajo este modelo que empleamos.

## 5.7. Búsqueda de hiper parámetros para el TOP 3 de modelos de clasificación con representación MFCC

En cuanto a los hiper parámetros de los modelos de clasificación de emociones utilizando la mejor representación, se aplicará una búsqueda de hiper parámetros para los siguientes tres modelos:

Tabla 18: Matriz de hiper parámetros en los 3 modelos de clasificación

| Parámetro | Modelo CNN | Modelo SVM | Random Forest |
|---|---|---|---|
| Capas convolucionales | 3-4 capas | N/A | N/A |
| Número de filtros | 32, 64, 128 | N/A | N/A |
| Tamaños de filtros | 2, 3, 5, 7 | N/A | N/A |
| Tipos de pooling | MaxPooling, AveragePooling | N/A | N/A |
| Tamaño de pooling | (2x2), (3x3) | N/A | N/A |
| Funciones de activación | ReLU, Sigmoid, Tanh, Leaky ReLU | N/A | N/A |
| Capas densas | 1-2 capas | N/A | N/A |
| Neuronas en capas densas | 32, 64, 128 | N/A | N/A |
| Kernel | N/A | ‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’ | N/A |
| C | N/A | 0.001, 1, 1000 | N/A |
| Degree (solo 'poly') | N/A | 2, 3, 5 | N/A |
| Coef0 (solo 'poly' y 'sigmoid') | N/A | 0, 5, 10 | N/A |
| Gamma (solo 'rbf', 'poly', 'sigmoid') | N/A | 1, 0.01, 0.0001 | N/A |
| n_estimators | N/A | N/A | 100, 200, 500 |
| max_depth | N/A | N/A | 10, 30, None |
| min_samples_split | N/A | N/A | 2, 5, 10 |
| max_features | N/A | N/A | 4, 0.3, ‘sqrt’, ‘log2’ |

Se utilizarán técnicas de búsqueda RandomSearch para la optimización de hiper parámetros, explorando las diferentes combinaciones posibles de estos para cada uno de los modelos seleccionados. Esta estrategia permitirá identificar la mejor configuración para cada modelo, garantizando un ajuste óptimo de los hiper parámetros a partir del mismo conjunto de datos. Al aplicar RandomSearch, se incrementa la probabilidad de encontrar la combinación que ofrezca el mayor rendimiento en la clasificación de emociones, maximizando así la precisión y efectividad

del modelo final.

### 5.7.1. Resultados de la búsqueda de hiper parámetros

Los resultados del mejor modelo de clasificación obtenido por cada uno de los modelos son los siguientes:

Tabla 19: Matriz de confusión del mejor modelo por hiper parámetros por CNN

| Clase | Precisión | Recall | F1-Score | Soporte |
|---|---|---|---|---|
| Amabilidad | 0.41 | 1.00 | 0.58 | 96 |
| Decepción | 0.00 | 0.00 | 0.00 | 43 |
| Desmotivado | 0.00 | 0.00 | 0.00 | 31 |
| Duda | 0.00 | 0.00 | 0.00 | 40 |
| Interés | 0.00 | 0.00 | 0.00 | 26 |
| **accuracy** | | | **0.41** | **236** |
| **Macro avg** | **0.08** | **0.20** | **0.12** | **236** |
| **Weighted avg** | **0.17** | **0.41** | **0.24** | **236** |

Tabla 20: Matriz de confusión del mejor modelo por hiper parámetros por RamdomForest

| Clase | Precisión | Recall | F1-Score | Soporte |
|---|---|---|---|---|
| Amabilidad | 0.58 | 0.41 | 0.48 | 96 |
| Decepción | 0.00 | 0.00 | 0.00 | 43 |
| Desmotivado | 0.00 | 0.00 | 0.00 | 31 |
| Duda | 0.29 | 0.05 | 0.09 | 40 |
| Interés | 0.00 | 0.00 | 0.00 | 26 |
| **Micro avg** | **0.55** | **0.17** | **0.26** | **236** |
| **Macro avg** | **0.17** | **0.09** | **0.11** | **236** |
| **Weighted avg** | **0.29** | **0.17** | **0.21** | **236** |
| **Samples avg** | **0.17** | **0.17** | **0.17** | **236** |

Tabla 21: Matriz de confusión del mejor modelo por hiper parámetros por SVM

| Clase | Precisión | Recall | F1-Score | Soporte |
|---|---|---|---|---|
| Amabilidad | 0.43 | 0.98 | 0.59 | 96 |
| Decepción | 0.00 | 0.00 | 0.00 | 43 |
| Desmotivado | 0.00 | 0.00 | 0.00 | 31 |
| Duda | 0.50 | 0.17 | 0.26 | 40 |
| Interés | 0.00 | 0.00 | 0.00 | 26 |

| | | | | |
|---|---|---|---|---|
| **Micro avg** | **0.43** | **0.43** | **0.43** | **236** |
| **Macro avg** | **0.19** | **0.23** | **0.17** | **236** |
| **Weighted avg** | **0.26** | **0.43** | **0.29** | **236** |
| **Samples avg** | **0.43** | **0.43** | **0.43** | **236** |

Se presentan las siguientes conclusiones frente a los resultados de las tablas: Tabla 19, 20 y 21:

- El F1-Score global para CNN es efectivamente 0.41, pero la precisión del modelo varía mucho entre las clases, con "Amabilidad" obteniendo buenos resultados (F1-Score 0.58), mientras que otras clases como "Decepción" y "Desmotivado" tienen un desempeño nulo. Esto sugiere que el CNN tiene dificultades para manejar las clases minoritarias y que se debe poner atención en mejorar la generalización para todas las clases.

- Es cierto que el recall promedio ponderado es bajo en los tres modelos, como lo indicaste. Por ejemplo, en el CNN, aunque "Amabilidad" tiene un recall perfecto (1.00), otras clases no son detectadas en absoluto. Este patrón se repite en los otros modelos. Así que la conclusión sobre la dificultad para diferenciar las clases es correcta, y podría estar relacionado con la representación del audio o con desequilibrios en las clases.

- La tendencia de los modelos a asignar la mayoría de los casos a la clase "Amabilidad" es coherente con los resultados obtenidos, especialmente en los modelos CNN y SVM, donde "Amabilidad" muestra un rendimiento significativamente superior en comparación con las demás clases. Esto sugiere un claro sesgo hacia la clase mayoritaria. Sin embargo, al recordar lo discutido en la sección 5.5 sobre la selección de audios representativos de las emociones predominantes, a pesar de haber garantizado una distribución equitativa de los audios entre las clases, el modelo no logra converger. Esto refuerza la idea de que la representación utilizada no permite una discriminación efectiva entre las emociones. Tal como se mencionó en la sección 3.1.2.2 desde la perspectiva psicolingüística, el análisis basado únicamente en el audio resulta insuficiente para captar las emociones de manera precisa. Es necesario incorporar otras características lingüísticas para mejorar la representación y, en consecuencia, lograr un modelo más robusto.

- Se propone la construcción de un modelo de clasificación de emociones a partir de audios, utilizando un número reducido de épocas de entrenamiento. Esto tiene como objetivo evitar una convergencia excesiva del modelo, aprovechando la estructura de

hiper parámetros óptima obtenida en la Tabla 19 con el modelo CNN. Esta estrategia busca mejorar el equilibrio entre el ajuste del modelo y su capacidad de generalización.

- En relación con la mejora del desempeño de los modelos, se adoptó un enfoque gradual que se detalla a lo largo de varias secciones del trabajo. En la sección 5.4, se aborda la mejora del desempeño de los modelos preliminares mediante un análisis detallado, comenzando con la evaluación de los resultados obtenidos a partir de un etiquetado altamente confiable. Posteriormente, en la sección 5.5, se amplía la evaluación con pruebas específicas que consideran diferentes tipos de canales, como mono y estéreo, para observar el impacto de estas variables en el rendimiento del modelo. Finalmente, en la sección 5.6, se lleva a cabo una búsqueda exhaustiva de hiper parámetros tanto para la representación de la señal en 5.6.1 como para los modelos planteados, con el fin de optimizar su desempeño. Estas acciones son esenciales para el proceso de mejora continua de los modelos y buscan maximizar la precisión y efectividad de las predicciones.

# 6. INTERFAZ DE USUARIO PARA EVALUAR NUEVOS AUDIOS

Se presenta la interfaz creada para el despliegue de los modelos del proyecto de grado, esta propuesta cuenta con las siguientes características:

- El aplicativo elegido es en forma de archivo ejecutable. Permitiendo rápido despliegue en usuario final sin necesidad de realizar instalaciones en equipo final ni incurrir en costos de infraestructura
- La arquitectura de la interfaz se basa en módulos para cargar los modelos con las clasificaciones de los audios que se coloquen anexo al ejecutable, volviéndolo modular en caso de querer mejorar los modelos sin cambiar las clasificaciones de emociones definidas.
- Permite evaluar nuevos audios sin necesidad de tener un conocimiento técnico.

**Nota**: Para un mayor detalle ver Anexo 1-README.md de la app Análisis Emociones en formato html.

## 6.1. Elementos empleados en el aplicativo

- **Módulo de selección de modelos:** Este módulo permite seleccionar un modelo previamente guardado en formato HDF5 ('.h5'), que es utilizado principalmente para compartir modelos entrenados de Keras y TensorFlow. Este formato conserva tanto la arquitectura del modelo como los pesos y la configuración de entrenamiento. La intención es facilitar al usuario la posibilidad de reemplazar el modelo por uno nuevo, siempre que este mantenga la misma configuración de representación de audios en MFCC, sin necesidad de intervenir en la interfaz de usuario.
- **Módulo de clasificación de audios:** Este módulo permite seleccionar un audio a la vez a través del explorador de archivos. La selección permitida corresponde a archivos en formato Waveform Audio File Format ('.wav'), que es el formato utilizado para compartir los audios del call center. Una vez que el usuario haya realizado la selección, deberá presionar el botón 'Procesar' para iniciar la extracción del audio. Esto permitirá generar su representación utilizando MFCC, previamente configurada, y llevar a cabo la identificación de la emoción predominante con base en el modelo cargado en el módulo anterior. **Visualización de interfaz usuario – APP análisis emociones**

El aplicativo está empaquetado en un archivo ejecutable que, al iniciarse, desplegará una ventana principal con dos ventanas auxiliares que permiten acceder a módulos específicos: el primero está dedicado a la selección y configuración del modelo, mientras que el segundo se enfoca en la clasificación de audios. Estas ventanas ofrecen una interfaz intuitiva para que el usuario pueda navegar fácilmente entre las opciones y realizar las tareas necesarias de forma eficiente y fluida.

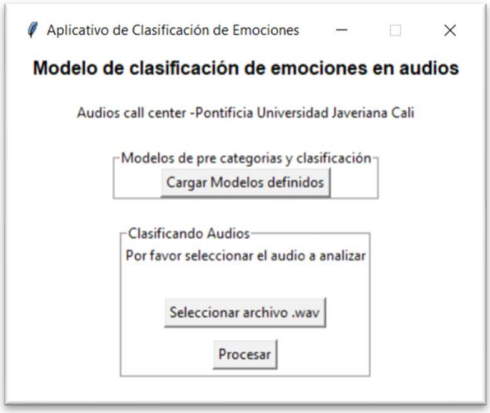
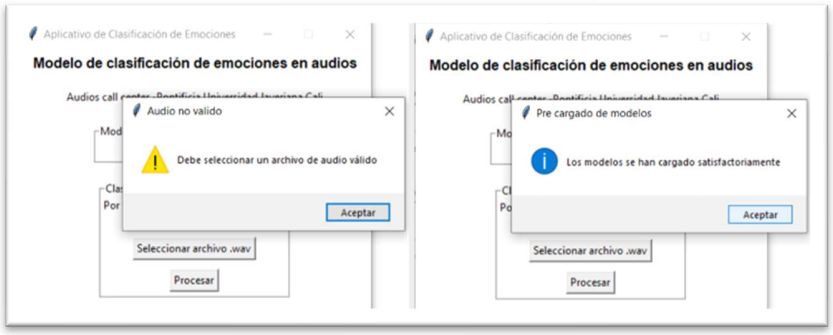



Figura 26: Ventana principal del aplicativo de interfaz de usuario

En la Figura 26, se observa dos recuadros que enmarcan los módulos del programa, el primer módulo cuenta con un botón para cargar los modelos definidos en el momento de la entrega y el segundo modulo cuenta con dos botones, el primero es para selecciona el audio a evaluar y el segunda es para obtener la clasificación de emoción predominante en el audio.

Figura 27: Controles del proceso de cargue de audios




En la Figura 27, se destacan los distintos controles que el usuario puede gestionar durante el proceso. Estos controles incluyen alertas que se activan si, en el momento de la clasificación, no se ha seleccionado correctamente el modelo o el audio. También se presentan mensajes de confirmación tanto para la selección del audio como del modelo, brindando al usuario una verificación clara de que el proceso se ha completado de manera adecuada.

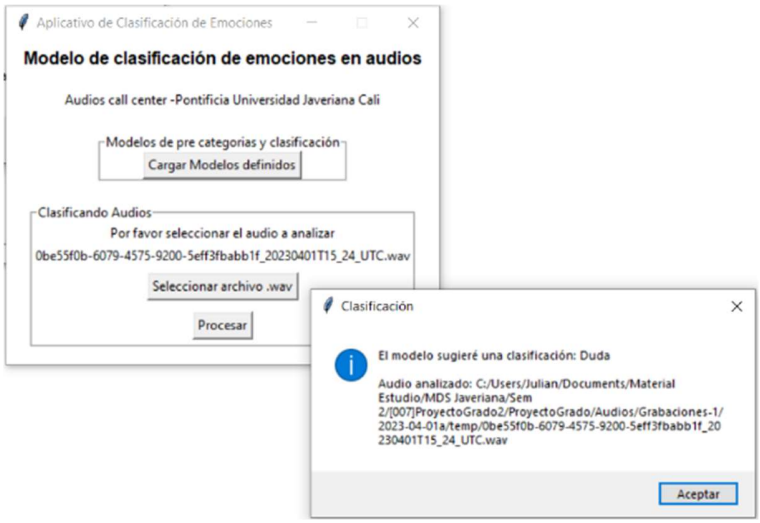



Figura 28: Clasificación de nuevos audios

En la Figura 28, se muestra el resultado proporcionado al usuario sobre la clasificación emocional de las llamadas. En caso de que el audio corresponda a un contestador, ruido o una llamada intervenida, también se presenta una preclasificación. El mensaje incluye tanto la emoción predominante identificada en la llamada como la ruta completa del archivo de audio seleccionado, brindando al usuario una información clara y detallada.

# 7. CONCLUSIONES Y TRABAJOS FUTUROS

## 7.1. CONCLUSIONES

- Respecto a los objetivos específicos, se logró procesar y analizar los audios de llamadas telefónicas para desarrollar un modelo de machine learning que clasifica las emociones predominantes en las llamadas. Estos modelos fueron evaluados y se integraron en una aplicación que permite la clasificación de nuevos audios mediante el modelo creado.
- En relación con el rendimiento de los modelos para la clasificación de audios con emociones complejas predominantes, se observa que abordar emociones de esta naturaleza podría requerir un mayor volumen de datos, un enfoque centrado en emociones básicas o la inclusión de características adicionales como lo son el contexto lingüístico o señales visuales de gesticulación. Por otro lado, una posible estrategia de dividir los audios en segmentos más pequeños, con ventanas que contengan emociones claramente definidas (no predominantes), podría ser considerada para mejorar los resultados. Este enfoque excede el alcance de este proyecto y por tanto no se consideró dentro de la metodología aplicada.
- El modelo de redes convolucionales (CNN) que obtuvo los mejores resultados en la clasificación en la etapa 1 de datos, logrando una precisión del 98 % para audios de 'no llamadas' y una precisión del 41 % para audios de 'llamadas', como se muestra en la sección 5.4.
- Los modelos, en particular CNN y SVM, muestran una tendencia a clasificar la mayoría de los casos como "Amabilidad", lo que indica un sesgo hacia esta clase mayoritaria. A pesar de contar con una distribución equitativa de audios entre las diferentes clases, la implementación de pruebas con canales estéreo y mono (sección 5.5), así como la búsqueda de los mejores hiper parámetros para cada modelo y diversas representaciones de la señal (sección 5.6.1), el modelo no logra converger adecuadamente. Esto refuerza la idea de que la representación basada únicamente en audio no es suficiente para diferenciar emociones complejas con precisión.
- Ante la limitación de disponer de un volumen mayor de audios y únicamente características acústicas, se propone construir un modelo de clasificación de emociones utilizando un número reducido de épocas de entrenamiento. Esto busca evitar una convergencia excesiva y mejorar la capacidad de generalización del modelo.

## 7.2. TRABAJOS FUTUROS

- La integración de modelos híbridos que combinen redes neuronales convolucionales (CNN) con técnicas de procesamiento de lenguaje natural (NLP) podría aumentar la capacidad del sistema para interpretar tanto señales acústicas como contenido lingüístico, proporcionando así una clasificación de emociones más robusta.
- Dado el alto grado de variabilidad en los audios, la diversidad de voces y las distintas regiones de origen en el país, se recomienda ampliar la base de audios de "llamadas" en el conjunto de datos procesados. Esto permitirá observar y capturar una mayor cantidad de parámetros, mejorando así

la precisión y efectividad del modelo en el entrenamiento y su capacidad para generalizar en distintos contextos acústicos.

# 8. REFERENCIAS BIBLIOGRÁFICAS


[1] V. Pachón, «En los contact center, cada queja, reclamo o venta deja $1,000 por llamada [32],» La República - May 2015. [En línea]. Available: https://www.larepublica.co/empresas/en-los-contact-center-cada-queja-reclamo-o-ventadeja-1-000-por-llamada-2258586.. [Último acceso: 30 11 23].

[2] B. E, S. M. , S. N. y C. M., "Quietly Angry, Loudly Happy: Self-Reported Customer Satisfaction Vs. Automatically Detected Emotion In Contact Center Calls"[4], Interaction Studies, vol. 24, no. 1, pp. 168-192, . DOI: 10.1075/is.22038.b, 2023.

[3] B. T. P. .. R. B. Pittala, «'Study of Speech Recognition Using CNN' DOI: 10.1109/ICAIS53314.2022.9743083. [20],» February 2022. [En línea]. Available: https://ieeexplore.ieee.org/document/9743083. [Último acceso: 30 11 2023].

[4] S. O. Dias, «ESTIMATION OF THE GLOTTAL PULSE FROM SPEECH OR SINGING VOICE,» Monograph of SCHOOL OF ENGINEERING OF THE UNIVERSITY OF PORTO, Oporto, Portugal, Jul, 2011.

[5] J. R. Zapata, «"Extracción de Características - STFT (Transformada Corta de Fourier)" Curso de Minería de Audio,[22],» [En línea]. Available: https://joserzapata.github.io/courses/mineria-audio/extraccion_caracteristicas/#stft---transformada-corta-de-fourier--short-time-fourier-transform.. [Último acceso: 23 11 2023].

[6] J. F. R.-M. y. M. O.-A. Paula Catalina Caycedo-Rosales, «Reconocimiento automatizado de señales bioacústicas: Una revisión de métodos y aplicaciones,» Universidad EAFIT , Ingeniería y Ciencia, Colombia, 5 11 2013. [En línea]. Available: https://www.redalyc.org/pdf/835/83529050011.pdf. [Último acceso: 31 01 2025].

[7] C. S. y. A. Klapuri, «CONSTANT-Q TRANSFORM TOOLBOX FOR MUSIC PROCESSING. university CORE,» [En línea]. Available: https://core.ac.uk/download/pdf/144846462.pdf. [Último acceso: 31 01 2025].

[8] F. A. Martín, «Desarrollo y análisis de clasificadores de señales de audio. Master en Ingenieria Acustica UNIVERSIDAD POLITECNICA DE VALENCIA,» UNIVERSIDAD POLITECNICA DE VALENCIA, 1 06 2017. [En línea]. Available: https://riunet.upv.es/bitstream/handle/10251/90005/Aguirre%20-%20Desarrollo%20y%20an%C3%A1lisis%20de%20clasificadores%20de%20se%C3%B1ales%20de%20audio..pdf?sequence=1. [Último acceso: 31 01 2025].

[9] M. J. M. S. I. K. N. e. A. S. N. B. y. M. A. R. Karim Mohammed Rezaul, «Enhancing Audio Classification Through MFCC Feature Extraction and Data Augmentation with CNN and RNN Models,» International Journal of Advanced Computer Science and Applications, 01 01 2024. [En línea]. Available: https://thesai.org/Downloads/Volume15No7/Paper_4-Enhancing_Audio_Classification_Through_MFCC.pdf. [Último acceso: 21 01 2025].

[10] C. E. Izard, "Emotion theory and research: Highlights, unanswered questions, and

emerging issues,"[27], Annual Review of Psychology, vol. 60, pp. 1-25, 2009.

[11] P. Ekman, "An argument for basic emotions," [11], Cognition and Emotion, vol. 6, no. 3/4, 1992.

[12] P. Ekman, "Universals and cultural differences in facial expressions of emotion,"[10], J. Cole, Ed. Lincoln: University of Nebraska Press, 1972.

[13] J. F. Jaen, Lenguaje, cuerpo y mente: claves de la Psicolingüística, Dialnet, 2007.

[14] L. M. R. A. Auria, "Support vector machines (SVM) as a technique for solvency analysis" [13], 2008.

[15] J. O. Alvear, «Arboles de decisión y Random Forest [14],» 16 12 2018. [En línea]. Available: https://bookdown.org/content/2031/. [Último acceso: 23 11 2023].

[16] IBM SPSS Statistics, «"Redes neuronales: perceptrón multicapa," [23],» [En línea]. Available: https://www.ibm.com/docs/es/spss-statistics/saas?topic=networks-multilayer-perceptron. [Último acceso: 23 11 2023].

[17] IBM., «"Redes neuronales recurrentes," [25],» [En línea]. Available: https://www.ibm.com/es-es/topics/recurrent-neural-networks.. [Último acceso: 23 11 2023].

[18] IBM., «"Redes neuronales convolucionales,"[26],» [En línea]. Available: https://www.ibm.com/es-es/topics/convolutional-neural-networks. [Último acceso: 23 11 2023].

[19] X. Font, «“Técnicas de clasificación supervised learning,” 1st ed., Barcelona: FUOC [33],» 2019. [En línea]. Available: https://openaccess.uoc.edu/bitstream/10609/147174/9/AnaliticaDeDatos_Modulo4_TecnicasDeClasificacionSupervisedLearning.pdf.. [Último acceso: 23 11 2023].

[20] I. Herrera y A. Figueroa, «"Aprendizaje Semi-Supervisado de Múltiples Vistas para Detectar Temporalidad de Preguntas," ResearchGate, [36],» [En línea]. Available: https://www.researchgate.net/profile/Alejandro-Figueroa-15/publication/306082751_Aprendizaje_Semi-Supervisado_de_Multiples_Vistas_para_Detectar_Temporalidad_de_Preguntas/links/57aeef3708ae0101f176ff3f/Aprendizaje-Semi-Supervisado-de-Multiples-Vistas-para-. [Último acceso: 12 01 2023].

[21] IBM., «"¿Qué es el etiquetado de datos?" [34],» [En línea]. Available: https://www.ibm.com/es-es/topics/data-labeling. [Último acceso: 7 12 2023].

[22] F. D. V. M. J. S. ,. F. P. Damaris Pascual González, «Detección de ruido en aprendizaje semisupervisado con el uso de flujos de datos,» Departamento de Lenguajes y Sistemas Informáticos, Universitat Jaume I. Castellón, España Univ. Antioquia, 01 06 2014. [En línea]. Available: http://www.scielo.org.co/pdf/rfiua/n71/n71a05.pdf. [Último acceso: 31 01 2025].

[23] H. P. Espinosa, "Reconocimiento de Emociones a Partir de Voz Basado en un Modelo Emocional Continuo," [5], Tesis de Doctorado, Instituto Nacional de Astrofísica, Óptica y

Electrónica, Tonantzintla, Puebla,, 2013.

[24] D. Sánchez Angón, "Reconocimiento de emociones a partir de imagen y voz,"[6], Tesis de Licenciatura, 2017.

[25] M. M. A. J. A. M. V. y. J. M. H. B. V. B. Ambario, "Reconocimiento de emociones a través del análisis de la voz," [7], Memorias del Congreso Internacional de Investigación Académica Journals Celaya 2017, Academia Journals, 2017.

[26] F. A. R. Steven R. Livingstone, «The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English [8],» 16 05 2018. [En línea]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391. [Último acceso: 23 11 2023].

[27] L. F. Correa Pinto, "Reconocimiento automático de emociones en audio y video usando Machine Learning," [9], Proyecto Fin de Carrera, Universidad de los Andes, Facultad de Ingeniería, Departamento de Ingeniería Eléctrica y Electrónica,, 2019.

[28] M. Pervaiz y T. Ahmed, «“Emotion Recognition from Speech using Prosodic and Linguistic Features,” International Journal of Advanced Computer Science and Applications, vol. 7,. doi: 10.14569/IJACSA.2016.070813. [18],» 08 2016. [En línea]. Available: https://www.researchgate.net/publication/309624815_Emotion_Recognition_from_Speech_using_Prosodic_and_Linguistic_Features. [Último acceso: 23 11 2023].

[29] F. E. J. Ierache, “Reconocimiento de emociones en la voz empleando redes neuronales y su integración en frameworks multimodales de educación emocional” [16], XXIII Workshop de Investigadores en Ciencias de la Computacion 623.

[30] V. E. H. L., “Emociones en Señales de Voz: Reconocimiento con Redes Neuronales Profundas" [17], Universitat Politècnica de Catalunya, Barcelona., 2021.

[31] A. M.Patricio G., «Análisis de Sentimiento en Audio mediante Inteligencia Artificial orientado al idioma Español [21],» Tesis grado,.Adm.Emp.Ing Sistemas., Universidad Carlos III de Madrid,, Madrid,España, 2022.

[32] Audeering., «"Devaice - Audeering." [28],» [En línea]. Available: https://www.audeering.com/products/devaice/. [Último acceso: 22 11 2023].

[33] MixedEmotions., «MixedEmotions Project, [29],» [En línea]. Available: https://mixedemotions-project.eu/. [Último acceso: 23 11 2023].

[34] E.-B. T.-C. J.-A. J. A. J.-B. S.Bedoya-Jaramillo, «Automatic Emotion Detection in Speech Using Mel frequency Cesptral Coefficients. Departamento de Ingeniería Electrónica y Telecomunicaciones - Universidad de Antioquia,» [En línea]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6340558. [Último acceso: 21 04 2024].

[35] Z. K. A. y. A. K. AL-TALABANI, «Mel Frequency Cepstral Coefficient and Its Applications: A Review,» 18 11 2022. [En línea]. Available: https://www.academia.edu/94221686/Mel_Frequency_Cepstral_Coefficient_and_its_Ap

plications_A_Review. [Último acceso: 2024 05 27].

[36] D. Calvo, «Clasificación de sonido con Redes Neuronales Convolucionales,» 29 06 2020. [En línea]. Available: https://www.diegocalvo.es/clasificacion-de-sonido-con-redes-neuronales-convolucionales/. [Último acceso: 2024 02 24 ].

[37] I. A. Pinedo Cantillo y J. Yáñez-Canal, «Emociones básicas y emociones morales complejas: claves de comprensión y criterios de clasificación desde una perspectiva cognitiva,» Tesis Psicológica, vol. 15, núm. 2, 2020, Julio-Diciembre, pp. 1-33.Fundación Universitaria los Libertadores, 2020.

[38] F. P. Naranjo, «Análisis de audio mediante técnicas de aprendizaje. Trabajo de grado,» UNIVERSITAT POLITÈCNICA DE VALÈNCIA, Politecnica, 2021.

[39] F. Santiago., «"Regresión Logística." [12],» *Tesis de grado. Universidad Autónoma de Madrid, ,* 2011. Madrid..

[40] M. Pervaiz y T. Ahmed, «“Emotion Recognition from Speech using Prosodic and Linguistic Features" International Journal of Advanced Computer Science and Applications, vol. 7,. doi: 10.14569/IJACSA.2016.070813 [18],» 8 2016. [En línea]. Available: https://www.researchgate.net/publication/309624815_. [Último acceso: 23 11 2023].

[41] K. Cherry, «Very well Mind, The 6 Types of Basic Emotions and Their Effect on Human Behavior,» 01 12 2022. [En línea]. Available: https://www.verywellmind.com/an-overview-of-the-types-of-emotions-4163976. [Último acceso: 23 05 2024].

[42] L. L. K. K. y. O. K. Martin Dietz, «Spectral Band Replication, a novel approach in audio coding,» Audio Engineering Society, 10 05 2002. [En línea]. Available: https://eva.fing.edu.uy/pluginfile.php/506737/mod_folder/content/0/AES112th_SBR.pdf. [Último acceso: 31 01 2025].