

**UNIVERSIDAD “MAYOR DE SAN SIMÓN”**

**FACULTAD DE CIENCIAS Y TECNOLOGÍA**

**Departamento de Informática y Sistemas**

**Aprendizaje No Supervisado**

**Reconocimiento de Emociones por Voz**

**INTEGRANTES: Brandon Brayan Goytia Cano**

**Bryan Henry López Maldonado**

**Henry Gabriel Montaño Uriona**

**MATERIA: Ciencia de Datos y Machine Learning**

**DOCENTE: Patricia Erika Rodriguez Bilbao**

**Junio 2022**

**COCHABAMBA - BOLIVIA**

**ÍNDICE**

[**Introducción**](#_kqwdxnt0p6x1) **3**

[**Objetivos**](#_pfa9ozpqhw1j) **3**

[Objetivo general](#_z8k1ebwddce7) 3

[Objetivos específicos](#_c66l3h7srhql) 3

[**Marco teórico**](#_jybtm01tennl) **4**

[Emociones](#_z8qph25ma6ly) 4

[Reconocimiento de las emociones](#_ewovdmbbnw3c) 4

[Selección y extracción de características](#_h64glp58yyul) 4

[Aprendizaje No supervisado](#_z6tm4pkjrwh) 5

[Clustering](#_mh4meuo5luvh) 5

[**Marco referencial**](#_dkem1owa5ltc) **7**

[**Ingeniería**](#_luv15wb6kdh9) **8**

[Referencia de las tecnologías a utilizar (Lenguaje, Librerías, otros.)](#_7xhhm7w7q5m4) 8

[Código Python Colab](#_un4oym27rxu3) 9

[Descripción del Dataset](#_bekrxblr94lp) 9

[Enfoque 1](#_nfv2mqo5kz7) 9

[Representación Esquemática](#_hx38yzbd5mns) 9

[Extracción de Features](#_7fv42bvklppc) 10

[Reducción de Dimensionalidad y Preprocesamiento](#_43cc9b7593el) 11

[Modelo de Arquitectura](#_qj0g5m93yl0j) 12

[Métrica/Evaluación](#_kdr45uya1hs5) 15

[Enfoque 2](#_9t5d2dw0agp7) 16

[Representación Esquemática](#_3pe1qoa6xmfe) 16

[Extracción de Features](#_ur89u8ydnbmf) 16

[Modelo de Arquitectura](#_8z30twgyfxs) 17

[Métrica/Evaluación Enfoque Espectrograma](#_ay76pysk3ok5) 19

[**Referencias**](#_d2v2nycvnax7) **20**

# 

# Introducción

Desde hace varios años, el campo de la inteligencia artificial ha tenido un crecimiento acelerado. Los avances de la inteligencia artificial han generado varias tecnologías que involucran la interacción humano-computadora que son de suma importancia. Vivimos en un mundo donde Siri Y Alexa están físicamente más cerca de nosotros que otros humanos.

Como seres humanos, el habla es una de las formas más naturales de expresarnos. Dependemos tanto de él que reconocemos su importancia cuando recurrimos a otras formas de comunicación como correos electrónicos y mensajes de texto donde a menudo usamos emojis para expresar las emociones asociadas a los mensajes. Dado que las emociones juegan un papel vital en la comunicación, la detección y el análisis de las mismas es de vital importancia en el mundo digital actual de la comunicación remota. La detección de emociones es una tarea desafiante, porque las emociones son subjetivas. No existe un consenso común sobre cómo medirlos o categorizarlos.

El reconocimiento de emociones del habla es una colección de metodologías que procesan y clasifican las señales de audios para detectar emociones en ella. Un modelo de este tipo puede encontrar uso en una amplia variedad de áreas de aplicación, como el asistente interactivo basado en voz o el análisis de conversación entre el agente y la persona que llama.

Desde una perspectiva de aprendizaje automático, el reconocimiento de emociones del habla es un problema de clasificación en el que una muestra de entrada (audio) debe clasificarse en unas pocas emociones. Por este motivo en el presente estudio, intentamos realizar un modelo basado en el aprendizaje no supervisado para detectar las emociones del hablante mediante el análisis de características acústicas de los datos de audio grabados.

# Objetivos

## Objetivo general

Desarrollar un modelo para reconocer emociones a partir del habla basado en el aprendizaje no supervisado.

## Objetivos específicos

* Analizar la transmisión, aplicación y manipulación de las señales de voz para poder utilizarlas en el modelo de reconocimiento de emociones.
* Extracción de características de la señal de voz.
* Generar un modelo de aprendizaje para el reconocimiento de emociones..
* Evaluar el funcionamiento y rendimiento del modelo.

# Marco teórico

## Emociones

Las emociones son un proceso del organismo humano que se activa mediante la

detección de alguna amenaza, peligro o desequilibrio, esto permite poner en acción los

recursos de cada persona para poner al frente alguna situación. Cada individuo puede

experimentar una emoción de forma diferente debido a las experiencias que viven o la

forma en la que tienen un diálogo con alguna otra persona. A veces las personas pueden

ser gobernadas por las emociones y estas hacen que se tomen acciones positivas o

negativas hacia uno mismo o hacia otra persona, es importante tener un claro control

sobre estas emociones.

## Reconocimiento de las emociones

El reconocimiento de voz es una tecnología que permite que una máquina comprenda el idioma hablado y lo traduzca a un formato legible por máquina. También se conoce como reconocimiento de voz por computadora o reconocimiento de voz automática. Es una forma de comunicarse con una computadora, y una computadora puede realizar una tarea específica según el comando.

El habla es el método más fundamental de comunicación humana. Los humanos pueden detectar naturalmente la emoción en el discurso que se les presenta. Sin embargo, no es tan sencillo para las máquinas. Aquí es donde se destaca la importancia del reconocimiento de emociones del habla (SER). Un sistema SER utiliza archivos que contienen datos de voz y los clasifica en varias emociones independientemente del contenido semántico.

Para crear un modelo de reconocimiento de emociones, se pueden emplear modelos de aprendizaje automático y aprendizaje profundo. Esto implica extraer características significativas de los datos sin procesar y utilizarlos para hacer que las máquinas entiendan las emociones humanas a través del entrenamiento de modelos. Durante el proceso de entrenamiento, el modelo aprende a clasificar la información y producir los resultados deseados mientras mantiene un cierto nivel de precisión.

La tarea de reconocimiento de emociones del habla se divide en dos secciones principales: selección y extracción de características, y clasificación.

### **Selección y extracción de características**

En este estudio se realizó el uso de tres características extraídas de los datos de audio, MFCC (coeficientes cepstrales de frecuencia de Mel), espectrograma de Mel y chroma.

* **MFCC (coeficientes cepstrales de frecuencia de Mel)**

En el análisis convencional de señales de tiempo, cualquier componente periódico (por ejemplo, ecos) aparece como picos agudos en el espectro de frecuencia correspondiente (es decir, espectro de Fourier. Esto se obtiene aplicando una transformada de Fourier en la señal de tiempo). Cualquier característica del cepstrum se obtiene aplicando la transformada de Fourier en un espectrograma. La característica especial de MFCC es que se toma en una escala Mel, que es una escala que relaciona la frecuencia percibida de un tono con la frecuencia real medida. Escala la frecuencia para que coincida más estrechamente con lo que el oído humano puede escuchar. La envolvente del espectro de potencia temporal de la señal del habla es representativa del tracto vocal y MFCC representa con precisión esta envolvente.

* **Espectrograma de Mel**

Se calcula una transformada rápida de Fourier sobre segmentos de ventana superpuestos de la señal, y obtenemos lo que se denomina espectrograma. Este es solo un espectrograma que representa la amplitud que se mapea en una escala de Mel.

* **Chroma.**

Un vector Chroma es típicamente un vector de características de 12 elementos que indica cuánta energía de cada clase de tono está presente en la señal en una escala cromática estándar.

## Aprendizaje No supervisado

El aprendizaje no supervisado se asocia con el aprendizaje sin supervisión o capacitación. Esto se realiza sin necesidad de etiquetar objetos. El objetivo del aprendizaje no supervisado es identificar las relaciones implícitas que puedan tener los datos. En el aprendizaje no supervisado, los algoritmos se entrenan con datos que no están etiquetado ni clasificado. En el aprendizaje no supervisado, el agente necesita aprender patrones sin valores de salida correspondientes.

Clasificación y regresión fueron los dos modelos usados en aprendizaje supervisado y los modelos de aprendizaje no supervisado son: **clustering, asociación y reducción de la dimensionalidad.**

### Clustering

La agrupación en clusters es una técnica de minería de datos que agrupa datos no etiquetados en grupos basados en similitudes y diferencias. Por ejemplo, el algoritmo de K-means divide puntos de datos similares en grupos, donde el valor “K” indica el tamaño y la granularidad del agrupamiento. Este método es útil para la segmentación del mercado, la compresión de imágenes y otras aplicaciones.

Una de las tareas que podemos realizar en un conjunto de datos sin etiquetas es encontrar grupos de datos en nuestro conjunto de datos que sean similares entre sí a los que llamamos clústeres. Un cluster se refiere a una colección de puntos de datos agregados debido a ciertas similitudes.

Luego, los clusters se posicionan como puntos y todas las observaciones o puntos de datos se asocian con el conglomerado más cercano, se calculan, se ajustan y luego el proceso comienza de nuevo usando los nuevos ajustes hasta que se alcanza el resultado deseado.

Se define un número objetivo k, que se refiere a la cantidad de centroides que necesita en el conjunto de datos. Un centroide es la ubicación imaginaria o real que representa el centro del grupo. Cada punto de datos se adjunta a cada uno de los grupos mediante la reducción de la suma de cuadrados dentro del grupo. El algoritmo identifica k número de centroides y luego adjunta cada punto de datos al cluster más cercano, mientras que al mismo tiempo mantiene los centroides lo más pequeños posible.

El proceso del algoritmo se compone principalmente de 4 pasos principales: Se colocan K puntos o cruces en el espacio de datos del objeto que representa el grupo inicial de centroides. Luego, cada objeto o punto de datos se asigna a la k más cercana. Luego, una vez asignados todos los objetos, se vuelven a calcular las posiciones de los k centroides. Finalmente este proceso se repite hasta que las posiciones de los centroides ya no se mueven.



https://stanford.edu/~cpiech/cs221/img/kmeansViz.png

La librería sklearn.cluster se puede usar para agrupar clusters que no han sido etiquetados. Cada algoritmo de agrupación en clústeres se divide en dos partes: una clase que implementa el método de ajuste para aprender los clústeres a partir de los datos de entrenamiento y una función que devuelve una matriz de etiquetas de enteros correspondientes a los diferentes clústeres cuando se proporcionan los datos del entrenamiento.

El uso de conjuntos de datos etiquetados es la principal diferencia entre los dos enfoques. El aprendizaje no supervisado no utiliza datos de entrada y salida etiquetados, mientras que el aprendizaje supervisado si lo hace. En lugar de hacer predicciones iterativas sobre conjuntos de datos y ajustar la respuesta correcta con un algoritmo de entrenamiento de aprendizaje, los modelos de aprendizaje no supervisados funcionan por sí solos para descubrir la estructura inherente de los datos no etiquetados.

# Marco referencial

**¿Para qué existen los sistemas de reconocimiento de emociones?**

Los sistemas de reconocimiento de emociones existen para mejorar y coadyuvar las labores donde es necesario una mayor atención a las personas, como por ejemplo el reconocimiento de emociones puede usarse en el campo de la medicina para priorizar la atención la atención de los pacientes que requieran atención urgente o en el campo de la seguridad puede usarse para evitar la violencia, ya sea doméstica o de otro tipo.

**Sistema de reconocimiento de emociones cuando se utiliza porque se utiliza donde se utiliza, como y a quiénes se va tomar en cuenta?**

El reconocimiento de emociones proporciona beneficios a muchas instituciones y aspectos de la vida. Es útil e importante para fines de seguridad y atención médica. El reconocimiento emocional se puede basar en el reconocimiento facial, el reconocimiento de voz y otros. El reconocimiento de emociones del habla es el acto de intentar reconocer las emociones humanas y los estados afectivos del habla. Usamos el hecho de que la voz a menudo refleja la emoción subyacente a través del tono y el tono para nuestro modelo de aprendizaje.

Los proveedores de atención médica pueden usar el reconocimiento de emociones para priorizar a sus pacientes analizando las expresiones del habla en la sala de espera, especialmente en los centros de atención de urgencia donde las personas no programan citas. Determinación del cansancio en el caso de conducción y alerta previa, detección de miedo en los casos en que una persona quiere abrir un sistema seguro como una casa o un cajero automático, donde el sistema puede bloquearse automáticamente cuando se detecta miedo en el habla y otros.

# Ingeniería

## Referencia de las tecnologías a utilizar (Lenguaje, Librerías, otros.)

Las herramientas y tecnología que vamos a utilizar son el Lenguaje de Programación

Python, el Colaboratory o “Colab” online de Google para notebooks, para que luego nos

ayude a desarrollar el algoritmo K-Means para el modelo de Clustering no supervisado.

Usaremos python no solo porque nos brinda una gran variedad de bibliotecas para

entrenar un modelo de aprendizaje automático, sino que también es compatible con

Google Colab para un notebook de Python en línea. Google Colab es un producto de

Google Research que permite a cualquier persona escribir y ejecutar código Python en

vivo a través del navegador y está especialmente diseñado para el aprendizaje

automático, el análisis de datos y la educación.

Entre las librerías a utilizar están:

**Pandas:** Es una librería de código abierto dentro de los desarrolladores de Python, y sobre todo dentro del ámbito de Data Science y Machine Learning, ya que ofrece unas estructuras muy poderosas y flexibles que facilitan la manipulación y tratamiento de datos.

**Numpy:** Es una librería de Python especializada en el cálculo numérico y el análisis de datos, especialmente para un gran volumen de datos.

**Sklearn:** Es una biblioteca de aprendizaje automático de software libre que cuenta con varios algoritmos de clasificación , regresión y agrupamiento, y está diseñado para interactuar con las bibliotecas numéricas y científicas de Python como NumPy y Pandas.

**IPython:** proporciona un completo conjunto de herramientas para ayudarlo a aprovechar al máximo el uso interactivo de Python posee soporte para visualización de datos interactivos y uso de kits de herramientas GUI.

**Librosa:** Librosa es una librería de python para analizar música y audio, también proporciona los componentes básicos necesarios para crear sistemas de recuperación de información musical. También podemos usar algoritmos de extracción de features de audio con mfccs, chroma y mel.

**cv2:** Es una biblioteca de enlaces de Python diseñada para resolver problemas de visión por computadora. Usaremos métodos como imread() para cargar imágenes para nuestro enfoque de espectrograma.

**matplotlib:** Es una librería de Python que nos permite hacer visualizaciones estáticas, animadas e interactivas. Esto nos va a permitir crear las imágenes de los espectrogramas.

**Keras:** Es una librería para desarrollar y evaluar modelos de aprendizaje profundo. Usaremos esta biblioteca para instanciar y usar el modelo VGG16 para el enfoque de espectrograma.

## Código Python Colab

### Descripción del Dataset

Nuestros archivos de audio de conjuntos de datos, categorizados en 7 emociones diferentes son: 'ira', 'disgusto', 'miedo', 'felicidad', 'neutral', 'tristeza' y 'sorpresa' o a para ‘anger, d para ‘disgust, f para ‘fear’, h para ‘happiness', n para 'neutral', sa para 'sadness’' y su para 'surprise'. Los nombres de los archivos de audio están etiquetados como a01.wav para mostrar que el archivo está etiquetado como ira (anger). Como tal, d01.wav se etiquetará como repugnancia (disgust), etc. En total, hay 480 archivos de audio etiquetados en nuestro conjunto de datos.

### Enfoque 1

En el presente trabajo nosotros propusimos dos enfoques para extraer dos tipos de características. En el primer enfoque trabajamos directamente con los archivos de audio para obtener un número de características.

En esta sección se describen los detalles

#### Representación Esquemática

Nos hemos acercado a la extracción de características del conjunto de datos de voz con el agrupamiento de KMeans. Usamos Mel, Mfcc y Chroma para la extracción de características que preprocesamos, escalamos con StandardScaler y analizamos usando PCA. Más tarde recibimos los resultados entrenados después de usar el enfoque de agrupación en clústeres de KMeans.



#### Extracción de Features

En este enfoque, trabajamos directamente en el conjunto de datos de audio para obtener características numéricas. Para reconocer la emoción de un discurso dado, se deben identificar y extraer una o más características significativas. Al extraer las características relevantes de las señales de voz, se puede adaptar un conjunto de datos de audio. Con la ayuda de la biblioteca Librosa, usamos la combinación de algoritmos Mel, Mfcc y Chroma para la extracción de características.

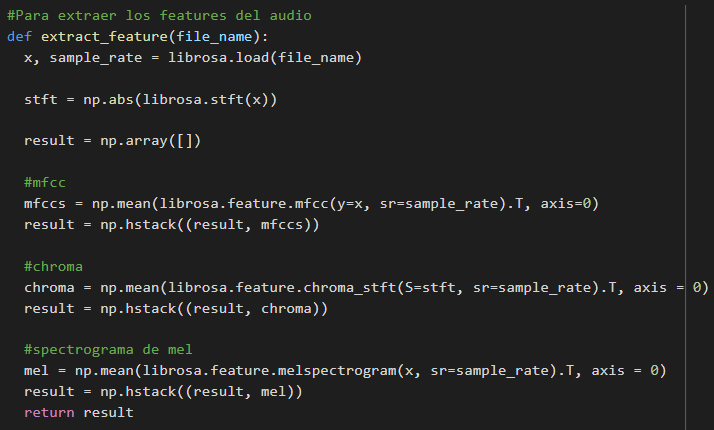


Fig 3: Extracción de Features/Características

Después de la extracción, las características de cada archivo junto con las etiquetas se han convertido en un vector de características 2D.

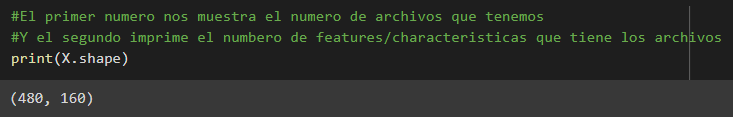


Fig 4: Vector de Características

#### Reducción de Dimensionalidad y Preprocesamiento

Se han extraído un total de 160 características de los archivos de audio. Para eliminar la escasez y la alta dimensionalidad del conjunto de datos, se realizó un preprocesamiento adicional de los datos con el método StandardScaler para la normalización de los datos y el método PCA que redujo el sobreajuste al eliminar variables altamente correlacionadas.

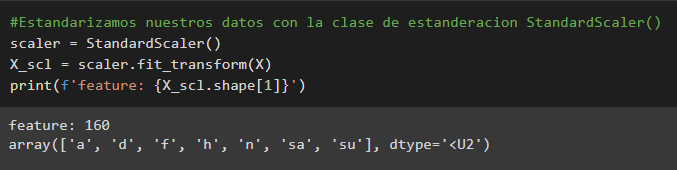


Fig 5: Normalización con StandardScalar

Con la función PCA() (Principal Component Analysis) obtenemos la relación de la varianza explicada para luego poder sacar los features más importantes. Utiliza la reducción de la dimensionalidad lineal mediante la descomposición de valores singulares de los datos para proyectarlos a un espacio dimensional más bajo. El método expalined\_variance\_ratio\_ utiliza el porcentaje de varianza explicado por cada uno de los componentes seleccionados. Se almacenan todos los componentes y la suma de las proporciones es igual a 1.0.

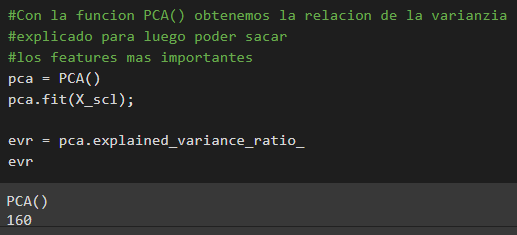


Fig 6: Porcentaje de Varianza Explicado

Usando PCA, se han seleccionado 14 características importantes para permitir un entrenamiento y una evaluación efectivos.

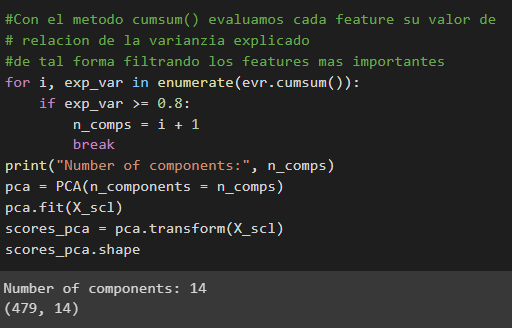


Fig 7: Reducción de Características Importantes con PCA

#### Modelo de Arquitectura

El algoritmo KMeans agrupa los datos tratando de separar muestras en n grupos de igual varianza, minimizando un criterio conocido como inercia o suma de cuadrados dentro del grupo. Este algoritmo requiere que se especifique la cantidad de grupos, en esto usamos la cantidad de funciones que hemos etiquetado en nuestro audio, que son 7 en total ( 'ira', 'disgusto', 'miedo', 'felicidad', 'neutral', 'tristeza' y 'sorpresa'). Luego entrenamos los datos de características reducidas usando PCA en el método de ajuste KMeans fit().

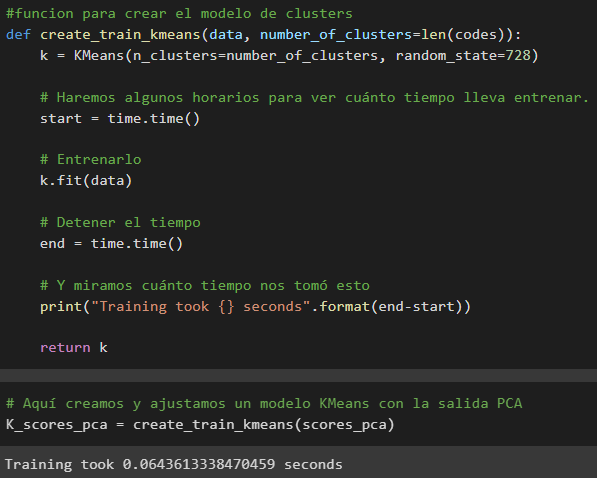


Fig 8: Entrenamiento del Modelo KMeans Clustering

Después de obtener los resultados del entrenamiento del modelo KMeans con PCA, predecimos el grupo más cercano al que pertenece cada muestra en X con el método de predicción KMeans. Usamos los resultados previstos de KMenas con los resultados de y\_train para contar la cantidad de etiquetas que hay por grupo.

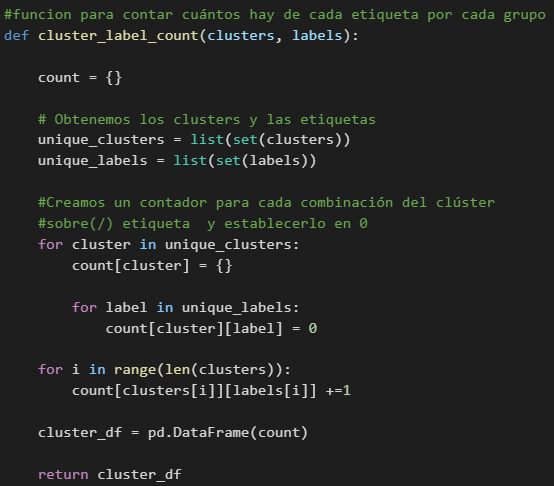


Fig 9: Cantidad de Etiquetados por Cluster

Luego, podemos asignar manualmente los códigos de clúster de acuerdo con el conteo de etiquetas de clúster para que podamos usarlo para evaluar nuestro modelo con el valor y\_train más adelante.

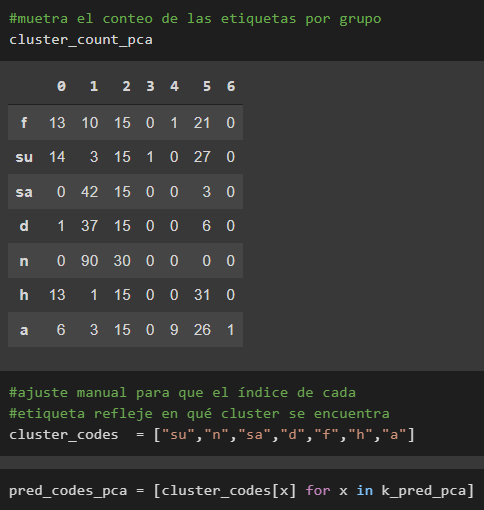


Fig 10: Asignación Manual de Codigo de Cluster

#### Métrica/Evaluación

Ahora que tenemos dos matrices, una con las etiquetas de predicción de clúster manual y otra con las etiquetas verdaderas o y\_train en nuestro caso, usamos el método accuracy\_score para calcular la precisión de nuestro modelo. En nuestro caso, tenemos un puntaje de precisión de 0.20489491, lo que muestra que el modelo de agrupamiento no supervisado con KMeans no es muy adecuado para el reconocimiento de voz.

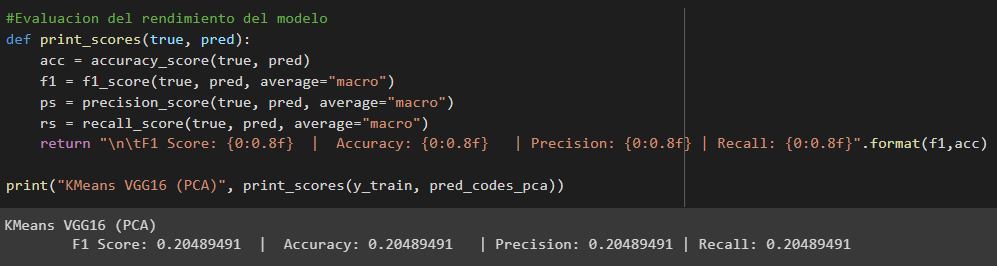


Fig 11: Evaluación del Rendimiento del Modelo

### Enfoque 2

En esta sección se describe el segundo enfoque utilizando la extracción de espectrogramas como imágenes.

#### Representación Esquemática

Otro modelo que podemos usar es el Espectrograma con Imagenet y el modelo VGG16 que se muestra en el diagrama de la figura 12. Usaremos el mismo dataset del anterior modelo de aprendizaje. Al convertir los archivos de audio en espectrogramas, lo convertimos en un dataset de imágenes.

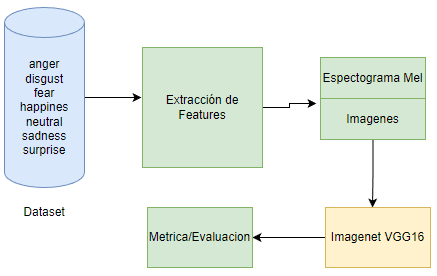


Fig 12: Diagrama de Representación de Esquemática con Espectrograma

#### Extracción de Features

Con la librería de librosa cargamos los archivos de audio para luego usar los métodos stft (utilizado en la extracción de características de audio para la descomposición de frecuencia de tiempo) y amplitud\_to\_db (convierte un espectrograma de amplitud en un espectrograma escalado en dB.) para luego poder convertir los archivos audios en imágenes/espectrogramas en lo cual podemos utilizarlo como features.

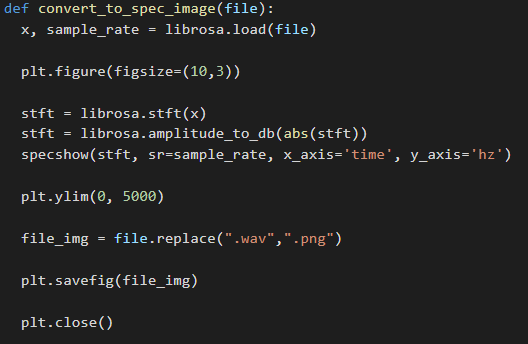


Fig 13: Conversión de Audio a Imagen

Reducción de Dimensionalidad y Preprocesamiento

Normalizamos las imágenes y códigos de etiquetado de los espectrogramas con numpy para luego usarlos en la cuenta de etiquetas de clusters.

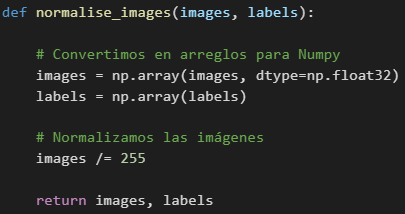


Fig 14: Normalización de Datos

#### Modelo de Arquitectura

Usamos la librería de Keras para crear el modelo VGG16, con el modelo pre entrenado imagenet, que utiliza un algoritmo de detección y clasificación de objetos con una arquitectura de red neuronal de convolución (CNN)



Fig 15: Modelo VGG16

Entrenamos el modelo con el método de KMeans fit() con la predicción del modelo VGG16.

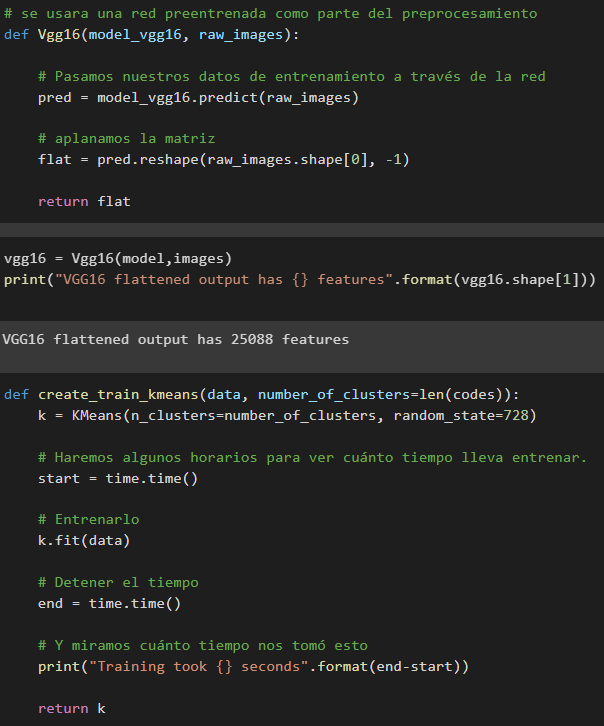


Fig 16: VGG16 y entrenamiento KMeans

Luego, podemos asignar manualmente los códigos de clúster de acuerdo con el conteo de etiquetas de clúster para que podamos usarlo para evaluar nuestro modelo más adelante.

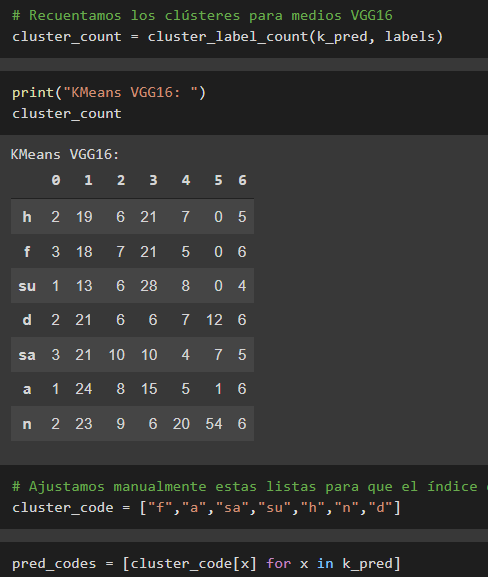


Fig 17: Cantidad de Clusters por medios VGG16

#### Métrica/Evaluación Enfoque Espectrograma

Ahora que tenemos dos matrices, una con las etiquetas de predicción de clúster manual y, usamos el método accuracy\_score para calcular la precisión de nuestro modelo. En nuestro caso, tenemos un puntaje de precisión de 0.23371822.

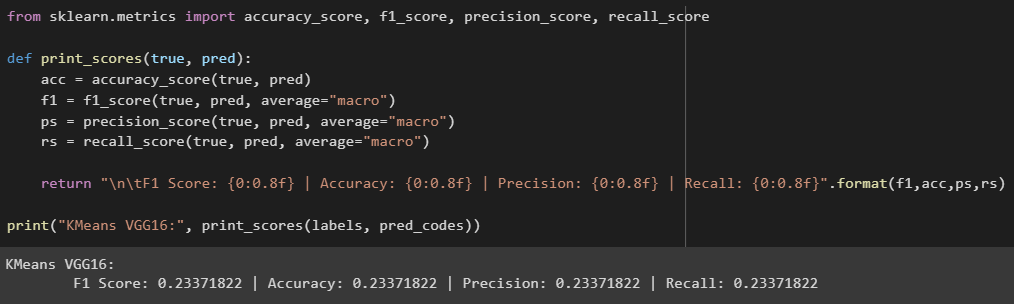


Fig 18: Evaluación del Rendimiento del Modelo

# Referencias

1. Joy, J, Kannan, A, Ram, S, & Rama, S. “Speech emotion recognition using Neural Network and MLP classifier.” International Journal of Engineering, Science, and Computing, vol 10, iss. 4, pp. 25170-25172, Apr. 2020.
2. Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., & Anbarjafari, G. Audio-visual emotion recognition in video clips. IEEE Transactions on Affective Computing,2017
3. Aggarwal, A.; Srivastava, A.; Agarwal, A.; Chahal, N.; Singh, D.; Alnuaim, A.A.; Alhadlaq, A.; Lee, H.-N. Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning. Sensors 2022,22, 2378. https://doi.org/10.3390/s22062378
4. <https://medium.com/analytics-vidhya/speech-emotion-recognition-using-machine-learning-df31f6fa8404>
5. https://stanford.edu/~cpiech/cs221/handouts/kmeans.html