**Aplicación de técnicas de clusterización para la clasificación de música Dance Electrónica.**

Estudiante: Carlos Alberto Murillo Martínez

Director: Marco Alumno

Correo: [malunno@eafit.edu.co](mailto:malunno@eafit.edu.co)

Co-director: Mauricio Toro Bermúdez

Correo: [mtorobe@eafit.edu.co](mailto:mtorobe@eafit.edu.co)

Palabras Clave: audio processing, streaming, kmeans, knn, metrics, spectogram, audio fingerprint

**Aplicación de técnicas de clusterización para la clasificación de música Dance Electrónica.**

El procesamiento de audio es una de las tareas esenciales para un científico de datos, el análisis de audio tiene aplicación en áreas muy diversas de conocimiento, como lo son: medicina, telecomunicaciones, mejorar la calidad de sonido en producciones musicales, inclusive aplicaciones militares (filtrar audio sospechoso o terrorista).

Siendo la captura del audio uno de los puntos clave para cualquier análisis, dado que es muy sensible a variaciones o ruido del ambiente, la tonalidad de una persona puede alterarse con solo una enfermedad de la garganta, entre otras.

Con este proyecto se pretende utilizar técnicas de agrupamiento concretamente Kmeans y Knn para clasificar un audio de entrada, utilizando diferentes métricas. Se utilizarán los métodos de clasificación para segmentar audios de entrada previamente procesados y de esta manera obtener una muestra de *huellas de audio (audio fingerprints)* para compararlas con una base de audios propios.

## Planteamiento del problema

Según la física el sonido es una vibración que se propaga con una onda sonora según el medio. Esta representación puede ser muestreada utilizando diferentes medios electrónicos, que transforman estos en una señal de audio que es una representación del sonido, puede ser una señal de un discurso, una canción o cualquier tipo de sonido.

La transformación de sonidos que son de naturaleza continua a una señal de audio implica algún tipo de técnica de *compresión*, para trasladarlas a un sistema de almacenamiento como un cd o un archivo de audio. El problema que intentamos resolver es que a partir de la serie de atributos que podemos obtener de la representación matricial de un sonido aplicar una métrica de distancia a un sonido para poder caracterizarlo según su género musical aplicando principalmente K-means y Knn para su clasificación.

Debido en muchos casos al ruido generado en el ambiente, pero también a la técnica que se utiliza para capturar el audio este puede mostrar muchas variaciones por este motivo los algoritmos que clasifiquen audios según su distancia pueden ser muy susceptibles a las técnicas de captura.

## Justificación

Actualmente el trabajo con audio en ciencia de datos es muy demandado debido a los múltiples mecanismos de captura de información. Por lo que se hace necesario utilizar diferentes estrategias de análisis de audio, en muchos casos el audio puede ser muy diferente por el ruido de fondo, pero su esencia es la misma.

Como se menciona al inicio del documento la aplicabilidad del análisis de audio trasciende el ámbito académico, es una de las bases de las tecnologías de *speech2text,* en el campo de la medicina se puede utilizar para medir probabilidad que una persona pueda sufrir un infarto de acuerdo con el ritmo de los latidos de su corazón.

La representación computacional de una señal de audio puede ser una matriz con coordenadas de tiempo y cada columna representa una característica del audio, como: frecuencia, bits por segundo, etc.

Este trabajo pretende aplicar una metodología robusta de clasificación tomando métricas de distancia distintas a la euclídea para estos algoritmos, por ejemplo: mediana, una combinación de métricas, etc.

## Objetivos

Comparar técnicas de clusterización como Knn y Kmeans para la clasificación de archivos de audio en música Dance Electrónica utilizando métricas robustas de clasificación.

* Calcular clústers utilizando diferentes métricas la distancia entre los diferentes audios.
* Examinar correlación entre géneros musicales (si existe)
* Crear una base de datos de sugerencias de canciones de acuerdo con su género.

## Estado del arte y Marco teórico

El sonido es definido por el Instituto Nacional de Estándares Americanos (ANSI) como:

1. *Una oscilación en presión, estrés, desplazamiento de partícula, etc. Propagado en un medio con fuerzas internas (ej. Elástico o viscoso) o de superposición de dicha oscilación propagada.*
2. *Sensación auditoria provocada por la oscilación descrita en el inciso (a)*

Esta oscilación se representa como una onda a través del tiempo. Cuando utilizamos algún método para capturar esta oscilación es lo que conocemos como una **señal de audio,** estas señales normalmente se repiten en intervalos regulares por lo que cada onda tiene la misma forma. Donde la altura muestra la intensidad del sonido y se conoce como la amplitud.

Mientras que el tiempo que tarda una señal en completar una onda completa se conoce como periodo. El numero de periodos en un segundo se le conoce como frecuencia. La frecuencia es recíproca al periodo y se mide en Hertz. (Stéphane, 2009)

Mientras que el oído humano es capaz de diferenciar sonidos de acuerdo a su “calidad”[[1]](#footnote-2) que es conocida como timbre.

Una vez visto la generalidad de lo que es el sonido en el mundo físico, se menciona la representación de este en el mundo digital, la naturaleza de una onda es continua, pero existe un límite de lo que podemos representar en una máquina por lo tanto lo que se realiza es un muestreo de la señal de audio, a este proceso se le conoce como *sampling rate* que se define como la frecuencia de muestras en una señal digital.

El estándar de muestreo para el audio en CD’s es 44.1 kilohertz (44100 hertz) por segundo (Christensson, 2015) eso significa que cada segundo hay 44100 muestras de audio por lo tanto para una muestra de 10 segundos de una canción estaríamos hablando de 441 KH, empezamos a notar que este puede llegar a convertirse rápidamente en un problema de big data. Por este motivo la elección de la muestra para calcular los centroides en el algoritmo de knn o kmeans es crucial (Han, 2010). La elección de la sección de audio que vamos a muestrear es fundamental para esto.

Para la identificación de audio es necesario tomar más cosas en consideración, por ejemplo: el modelo más utilizado en la actualidad y que mejor resultado ha dado es transformar el audio en una imagen conocida como un *espectograma* luego se procesan las imágenes utilizando *Redes Neuronales Convolucionadas (CNN or ConvNet)* (Valueva, Nagornov, Lyakhov, Valuev, & Chervyakov, 2020)*,* este acercamiento es muy útil porque permite identificar características claver para la identificación del audio, pero antes de entrar en detalle

Para entender mejor lo que es un espectograma debemos definir lo que es el espectro, recordando que un sonido puede representarse como la suma de diferentes frecuencias, el espectro se define como el conjunto de señales que sumadas producen una señal de audio. La frecuencia más baja se conoce como la *frecuencia fundamental* y los múltiplos de la frecuencia fundamental se les conoce como harmónicos.

Entonces el espectograma es un gráfico que muestra la frecuencia de las distintas señales contra el tiempo, en otras palabras, es un gráfico de sus espectros contra el tiempo por lo que algunas veces incluso se muestra en tres dimensiones, como se muestra en la imagen:

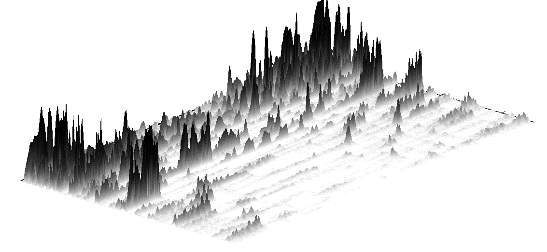


Ilustración 1 - espectograma 3d (Debianux)

Pero su versión más utilizada es en dos dimensiones:

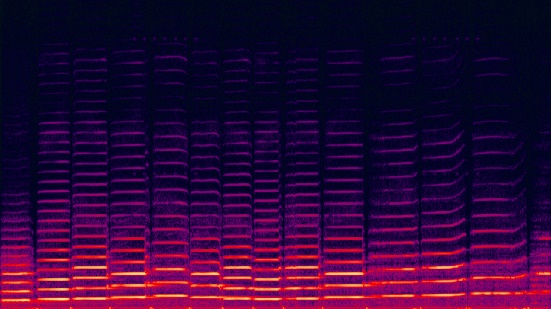


Ilustración 2 - espectograma audio de violín (Anon)

Si sabemos que un espectograma es una suma de frecuencias a lo largo de un tiempo podemos obtener una frecuencia media o una frecuencia mediana para obtener un patrón de un audio, en este punto entra la siguiente parte de la investigación.

Como se observa uno de los mayores problemas a los que se enfrenta utilizar los modelos Knn y Kmeans es la gran cantidad de datos que dispone para trabajar se deben simplificar muchas veces para analizarlos escoger, ¿Qué secciones del audio muestrear?, se debe pre procesar el audio para eliminar secciones de poca relevancia y por último determinar cuales son los centroides con los que se debería trabajar.

Los algoritmos que se van a trabajar en este proyecto se aplican en un contexto de agrupamiento más no de identificación, la distancia entre dos sonidos puede tener muchos componentes de análisis distintos, en el que incluso la preparación de datos influye, por este motivo la arquitectura más utilizada para identificación de audio es basada en CNN.

Sin embargo, otra rama que es de mucha utilidad es la clasificación de audio, es analizar su MFCC (Mel-Frequency cepstral coefficients), como nos menciona (Thiruvengatanadhan, 2017) en su artículo de uso de knn para clasificación de audio, *“El cepstrum de frecuencia mel ha demostrado ser muy eficaz en el reconocimiento de la estructura de las señales musicales y en el modelado del tono subjetivo contenido de frecuencia de las señales de audio”*.

Al aplicar esta técnica lo que se propone es realizar una serie de pasos para obtener características del audio, las cuales son principalmente (Sahidullah & Saha, 2012):

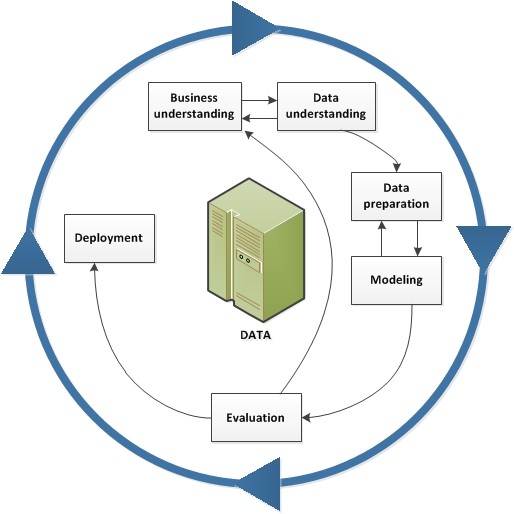
* Separar la señal en pequeños tramos.
* A cada tramo aplicarle la [Transformada de Fourier discreta](https://wikipedia.firstpartyapps.oaspapps.com/wikipedia/wikipedia_dev.html?et=%2BAA%3D%3D&_host_Info=Word$Win32$16.01$es-ES$$$$16) y obtener la potencia espectral de la señal.
* Aplicar el banco de filtros correspondientes a la [Escala Mel](https://wikipedia.firstpartyapps.oaspapps.com/wikipedia/wikipedia_dev.html?et=%2BAA%3D%3D&_host_Info=Word$Win32$16.01$es-ES$$$$16) al espectro obtenido en el paso anterior y sumar las energías en cada uno de ellos.
* Tomar el logaritmo de todas las energías de cada frecuencia mel
* Aplicarle la [transformada de coseno discreta](https://wikipedia.firstpartyapps.oaspapps.com/wikipedia/wikipedia_dev.html?et=%2BAA%3D%3D&_host_Info=Word$Win32$16.01$es-ES$$$$16) a estos logaritmos.

Podemos aplicar técnicas para robustecer las medidas de distancia en los algoritmos que se van a trabajar, como aplicar una normalización de los datos con el método MEL, normalizar la onda y eliminar secciones de amplitud baja (silencios).

También podemos aplicar diferentes medidas de distancia, como: distancia euclídea ponderada, mediana, Mahalannobis, entre otras medidas de distancia de acuerdo a la forma de los datos.

## Metodología

Para este trabajo se va a utilizar la metodología CRISP-DM:



*“CRISP-DM son las siglas de Cross-Industry Standard Process for Data Mining, es un método probado para orientar sus trabajos de minería de datos.*

* *Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.*
* *Como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos.”(IBM Knowledge center, 2021)*

Aplicado al proyecto se realizará de la siguiente manera:

* **Entendimiento del negocio:** En esta fase analizaremos el proyecto realizado por Victor Tideman (Tideman, 2022) para la obtención de la información.
* **Entendimiento de los datos:**En esta sección, se trata la descripción de los datos, cuantas variables y el tamaño del archivo. Finaliza explicando, cuál es el procedimiento para crear las variables necesarias para el modelo.
* **Preparación de datos:**Para el análisis de audio necesitamos obtener información a partir de los archivos de audio suministrados, y con los archivos de audio que se realizará la evaluación
* **Modelado:**Aplicar los algoritmos K-means y Knn para clasificar según sus géneros musicales, utilizando diferentes métricas de distancia, algunas ponderadas para dar mayor importancia al bit rate, o a la frecuencia y de esta manera poder determinar su género musical.
* **Evaluación:**Evaluar estos resultados aplicando los resultados del modelo sobre un conjunto de evaluación, algunas de las métricas que se plantean son: RMSE, Silhouette Analysis, cohesión (**compañía del director**).
* **Despliegue:**No hace parte del alcance

## Productos esperados

El trabajo tiene los siguientes entregables:

* Documento de preparación de datos: Documento en el que se muestra el procedimiento y parámetros generados en el proceso para poder implementar el proceso con el análisis de resultados de los procedimientos y tiempos de ejecución.
* Repositorio en Git: Contiene códigos en Python de preparación de datos de manera iterativa y parametrizable y un tablero de control donde se pueden analizar resultados.
* Carpeta en MEGA: con los archivos de audio utilizados, separados en carpetas de entrenamiento y evaluación.

## Plan de Gestión de Datos

Los archivos de audio suministrados de insumo fueron proporcionados por el profesor Marco Alumno en un contexto de continuar el proyecto desarrollado por uno de sus estudiantes de maestría, si existe la solicitud de crear copias de seguridad de estos, se almacenarán copias en distintos dispositivos.

Mientras que los datos de evaluación serán proporcionados por el profesor Marco.

## Aspectos éticos

Los archivos de audio que presenta el proyecto son canciones licenciadas para no violar leyes de copyright y no serán expuestas de manera pública (internet)

* ¿cuáles son los beneficios y quién se beneficiará?

Este trabajo permite analizar el primer paso dentro de la clasificacion de archivos de audio y permitirá seleccionar una metodología que permita utilizar un mejor acercamiento.

# Bibliografía

Anon. (s.f.). espectograma violín. *CC BY-SA 3.0.*

Christensson, P. (Mayo de 2015). *Sample Rate Definition*. Obtenido de techterms: https://techterms.com/definition/sample\_rate

Debianux, D. (s.f.). Espectograma 3d. *CC BY-SA 3.0.*

Han, G. C. (2010). Improve K-means clustering for audio data by exploring a reasonable sampling rate. *Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, 1639-1642.

Sahidullah, M., & Saha, G. (2012). Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech and comunication*, 543–565. .

Stéphane, M. (2009). *A Wavelet Tour of Signal Processing (Third Edition).* (M. Stéphane, Ed.) Academic Press. doi:https://doi.org/10.1016/B978-0-12-374370-1.00014-8.

Thiruvengatanadhan, R. (2017). Speech/Music Classification using MFCC and KNN. *ISSN 0973-1873 Volume 13, 13*, 2449-2452. Obtenido de https://www.ripublication.com/Openaccess/ijcirv13n10\_14.pdf

Valueva, M., Nagornov, N., Lyakhov, P., Valuev, G., & Chervyakov, N. (2020). Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Mathematics and Computers in Simulation*, 232–243.

1. Se entiende calidad como la regularidad de la onda. [↑](#footnote-ref-2)