DATA SCIENCE BOOTCAMP SEPTIEMBRE-DICIEMBRE 2019

MEMORIA DEL PROYECTO

# Redes neuronales para la clasificación de música utilizando atributos y espectrogramas

Ricardo Sousa Freitas

# Motivación

Muchos de los sistemas de recomendación de música tienden a agrupar canciones por idioma, país del artista, popularidad.

Esto hace difícil descubrir nueva música en base a cómo suena. Por ejemplo, si a una persona le gusta música rítmica y con *Vocoder*, seguramente las recomendaciones entrarán en el bucle del *Regaetton* si ya ha guardado algunas canciones de este género.

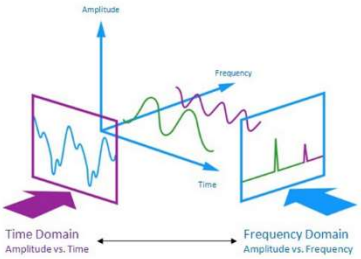
Recomendar y agrupar canciones en base a sus características, tanto numéricas como de imágenes (espectrogramas), podría permitir conocer nueva música, aunque no sea muy popular o cercana geográficamente.

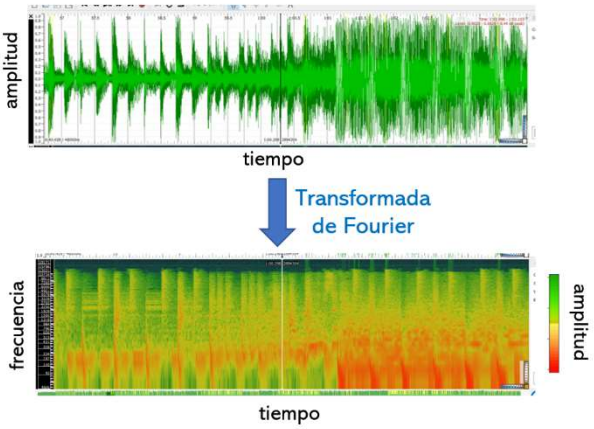
# Objetivos

Evaluar la precisión de los algoritmos de Machine Learning, específicamente de Redes Neuronales y K-Means, para clasificar la música por género y por tipo de música (más melódica y armónica o más rítmica y enérgica). Todo ello extrayendo variables numéricas e imágenes (espectrogramas) a partir de muestras de audio.

# Señales de audio

Las señales de audio se recogen como muestras de amplitudes en función del tiempo. La transformada de Fourier permite extraer el espectro de frecuencias y obtener más información sobre las características de la señal.




# Metodología

Fase de investigación

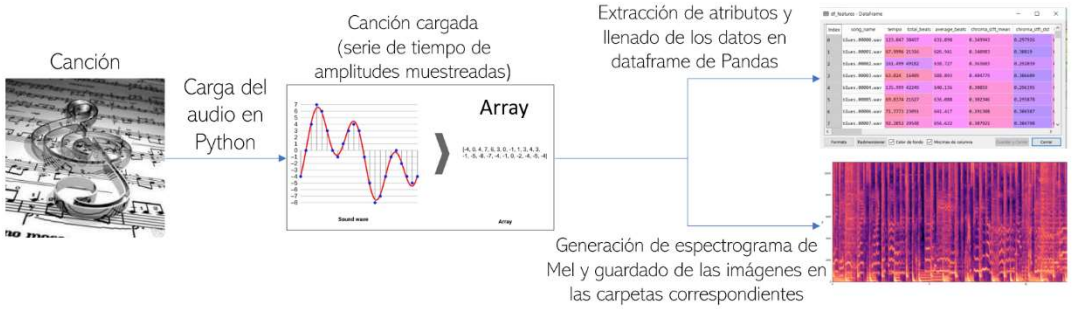
- Lectura de artículos y estudios acerca del tratamiento de audio con algoritmos de Machine Learning.
- Búsqueda y recopilación de posibles datasets de audio.
- Aprendizaje del procesamiento y análisis de audio con Python (librerías, código).

Dataset y librería de audio para el proyecto

- Se ha elegido para el proyecto el "GTZAN Genre Collection", el cual contiene 1000 canciones para 10 géneros y cada género con 100 muestras de música de 30 segundos.
- Para la extracción de atributos y espectrogramas, se ha utilizado la librería de Python denominada Librosa, la cual está diseñada para análisis de audio y música.

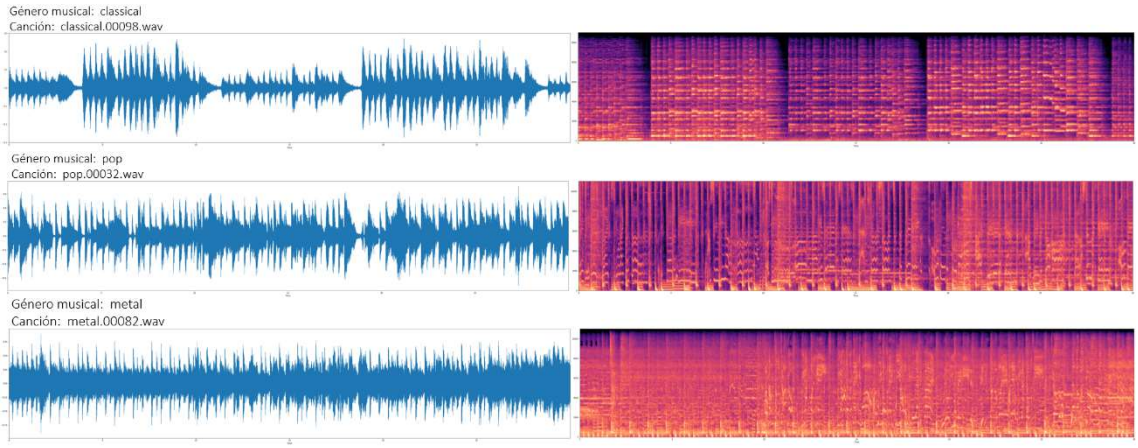
## Creación de los datasets para los modelos de Machine Learning

En esta fase del proyecto se crearon dos datasets. Uno de variables numéricas y un segundo dataset de imágenes, específicamente de espectrogramas de Mel, que modelan la distribución de energía espectral teniendo en cuenta la percepción del oído humano.




## Análisis de algunas canciones

Se pueden observar a continuación las diferencias entre 3 géneros musicales del dataset. En la forma de onda se pueden observar solo las variaciones de amplitud en el tiempo, mientras que el espectrograma de Mel correspondiente arroja mucha más información de la canción, revelando las variaciones del espectro de frecuencias en el tiempo y de su intensidad, reflejada esta última en la escala de colores.

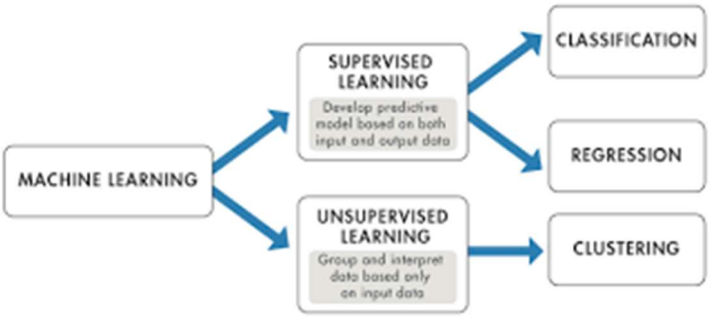



Formas de onda Espectrogramas de Mel

# Machine Learning

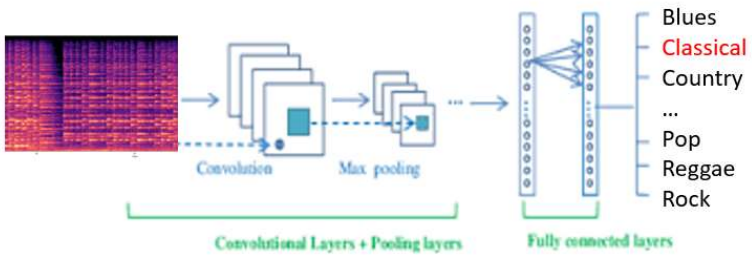
Algunas nociones del Aprendizaje Automático

- Los modelos pueden clasificarse en supervisados, donde el modelo se entrena con los datos de entrada y de salida. El algoritmo aprende con una parte de los datos y luego evaluamos el modelo entrenado con otra parte de los datos que no han sido tocados para el aprendizaje. Si el valor a predecir es continuo, se califica como de regresión, mientras que si es discreto sería de clasificación.
- Los modelos no supervisados agrupan e interpretan los datos solo con datos de entrada, estableciendo "clusters".




Machine Learning con el dataset de espectrogramas

Las Redes Neuronales Convolucionales, *CNN* por sus siglas en inglés, son muy utilizadas para la clasificación de imágenes. Para este proyecto, la idea es entrenar la red neuronal convolucional de tal forma que extraiga patrones comunes a cada género de los espectrogramas y permita diferenciar unos géneros de otros.




Design of Convolution Neural Network

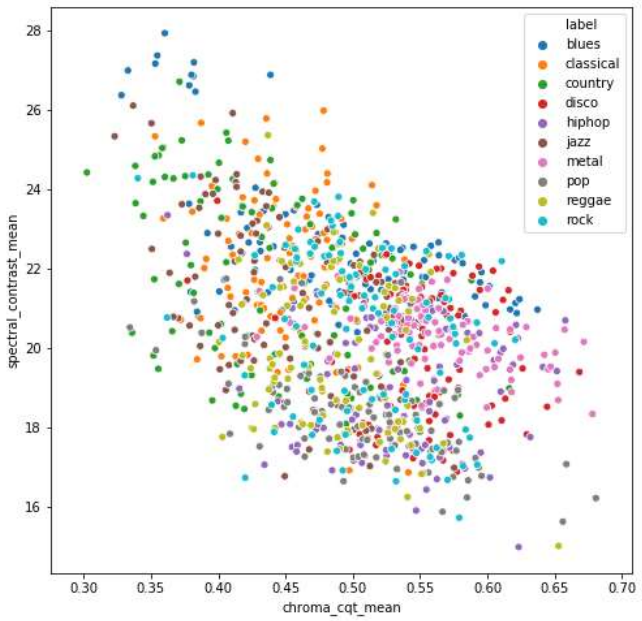
## Primeras pruebas con Redes Convolucionales para los 10 géneros

Debido a los pobres resultados de las predicciones de redes convolucionales con espectrogramas (precisión de la predicción de 45% a 53%) y teniendo en cuenta que solo hay 100 muestras por clase para 10 clases diferentes, se decidió agruparlas utilizando *K-Means*.


Parámetros y arquitectura de la Red Neuronal Convolucional

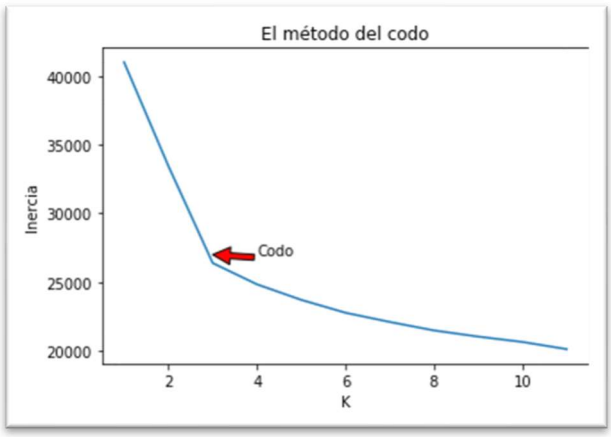
## *K-Means* para agrupar clases (Dataset de atributos numéricos)

En el "scatterplot" de dos de las variables extraídas de los audios se puede observar cómo es difícil discriminar clases. Géneros musicales muy diferentes como el clásico y el metal se diferencian bien, pero otros, como por ejemplo el *reggae*, tienen una dispersión mayor, haciendo más difícil su clasificación.



Agrupamiento con *K-Means*

Calculamos las inercias del *K-Means* para 12 "clusters" y se hizo el dibujo del codo para encontrar el número de "clusters" (clases) que mejor diferencia nuestros datos.



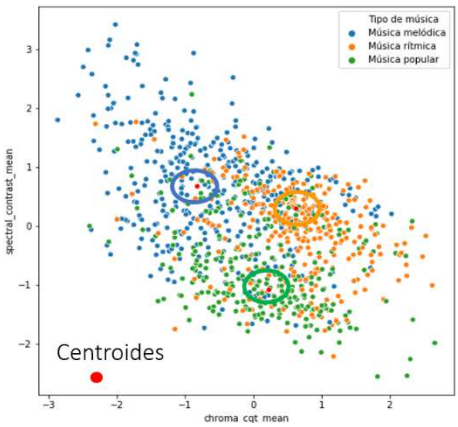

La pendiente de la inercia versus el número de "clusters" cambia drásticamente (disminuye) en k=3. Se utilizó entonces este número para definir, ajustar y predecir con *K-Means*.

Agrupamiento con *K-Means* – Resultados

El algoritmo de *K-Means* agrupa los 10 géneros musicales en 3 "clusters", en los cuales se pueden diferenciar mejor las clases.

| label | count | unique | top | freq |
|---|---|---|---|---|
| blues | 100 | 2 | 1 | 53 |
| classical | 100 | 3 | 1 | 98 |
| country | 100 | 3 | 1 | 57 |
| disco | 100 | 3 | 2 | 53 |
| hiphop | 100 | 3 | 0 | 54 |
| jazz | 100 | 3 | 1 | 73 |
| metal | 100 | 2 | 2 | 97 |
| pop | 100 | 2 | 0 | 92 |
| reggae | 100 | 3 | 0 | 46 |
| rock | 100 | 3 | 2 | 51 |

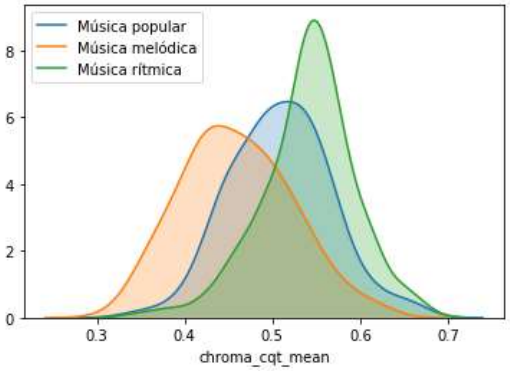
```
cluster0=["hiphop", "pop", "reggae"]
cluster1=["blues", "classical", "country", "jazz"]
cluster2=["disco", "metal", "rock"]
```

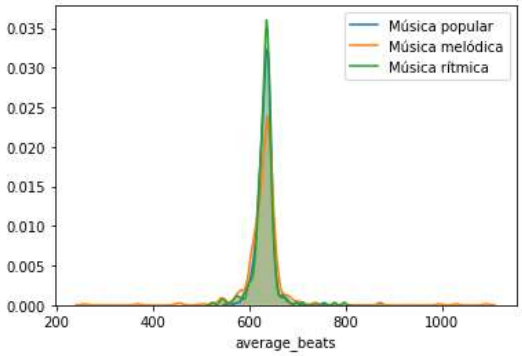


# Análisis y limpieza de los datos

Visualización

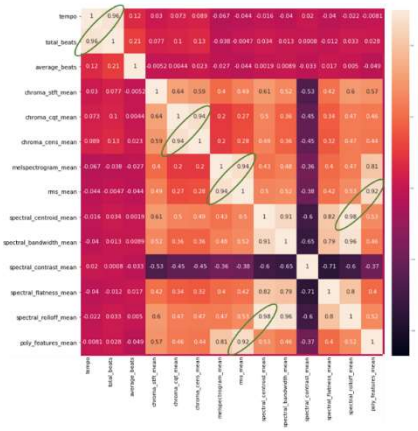
El "Tipo de Música" agrupa las 10 clases originales en 3 "clusters" del estudio del *K-Means*. Se puede hacer un mejor análisis de las variables y eliminar aquellas que peor diferencian las clases.

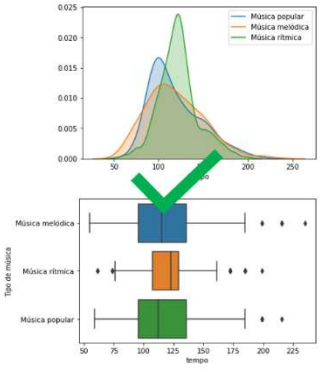


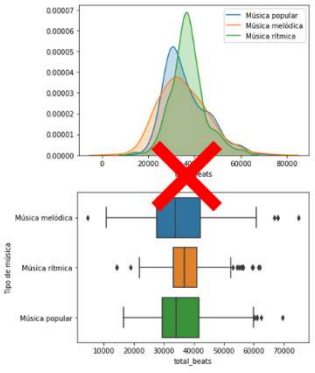




Limpieza de los datos

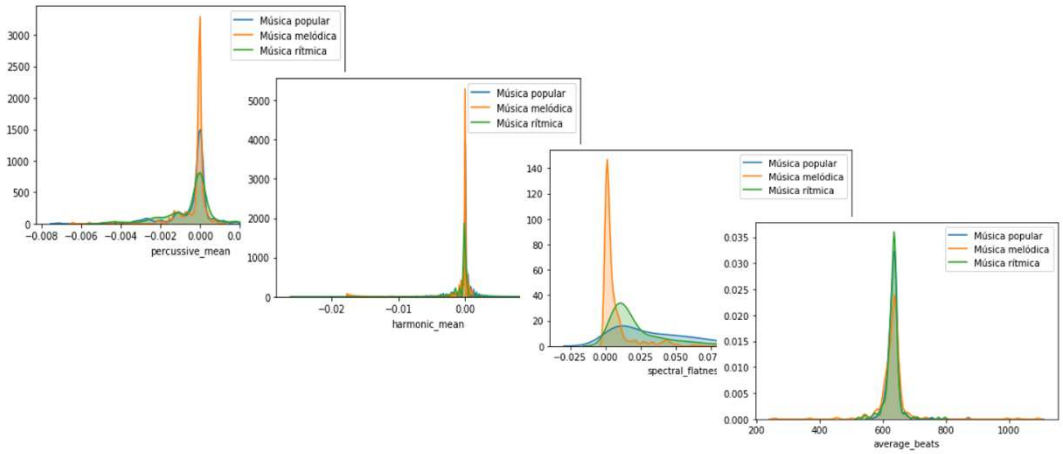
Se realizó la visualización conjunta de distribuciones y "boxplots" de las variables muy correlacionadas para analizar que variables eliminar en función de cómo separan los tipos de música y el número de "outliers". También se tuvo en consideración la importancia del atributo para describir la canción.





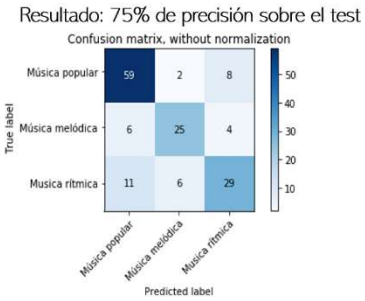
Se analizaron otras variables que podrían aportar poco en la diferenciación de tipos de música y se eliminaron del dataset.




# Machine Learning

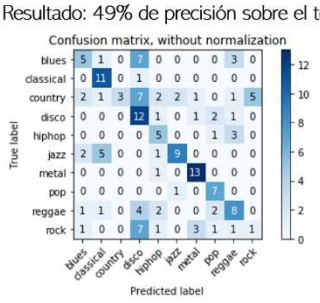
## Resultados de Redes Neuronales Convolucionales

Dataset: Espectrogramas
Número de clases: 3




El mayor número de errores se produce entre la música rítmica y la popular. Esto es coherente con lo que vimos en los espectrogramas y las distribuciones de los atributos

Dataset: Espectrogramas
Número de clases: 10

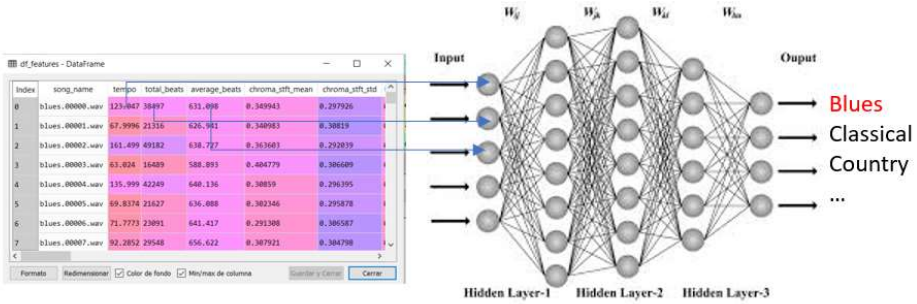



Para 10 clases el porcentaje de acierto disminuye en un 25%

El mayor número de errores se produce cuando se predice como música disco géneros como el blues, country y rock. Los espectrogramas de estos géneros tienen patrones parecidos dificultando a la red hacer su trabajo de clasificación

## Redes neuronales tipo *Multilayer Perceptron (MLP)* con el dataset de atributos

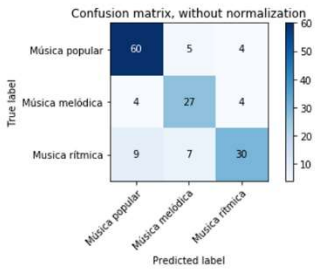
En las Redes Neuronales Multicapa los parámetros de entrada serán las variables de cada canción, las cuales se relacionan con las capas ocultas a través de funciones de activación y pesos. Estos últimos se van actualizando en las diferentes iteraciones con relaciones complejas entre las neuronas estableciendo al final unas probabilidades para cada categoría en la capa de salida.




## Resultados de Redes Neuronales Multicapa

Dataset: Atributos
Número de clases: 3

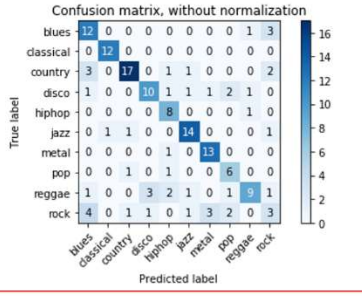
Resultado: 78% de precisión sobre el test




El mayor número de errores se produce entre la música rítmica y la popular. Esto es coherente con lo que vimos en los espectrogramas y las distribuciones de los atributos

Dataset: Atributos
Número de clases: 10

Resultado: 69% de precisión sobre el test




El mayor número de errores se produce entre rock, metal y blues que musicalmente tienen orígenes similares

# Conclusiones

- A pesar de ser un dataset pequeño (solo 100 muestras por género), las redes neuronales consiguen buenas predicciones, sobre todo para el dataset de atributos (78% de precisión para tres clases y 69% para los 10 géneros).
- El uso de espectrogramas para la predicción de géneros musicales presenta resultados pobres. Patrones tan específicos y sutiles son difíciles de entrenar para la red neuronal.
- *K-Means* funciona bien para agrupar géneros con características similares. Tendría utilidad para crear *playlists*, por ejemplo.

# Acciones futuras

- Combinar ambos datasets para obtener mejores predicciones.
- Probar estos algoritmos con más muestras de audio.
- Hacer *Data Engineering* para crear nuevas variables que diferencien mejor las clases.