Clasificación de música a través del análisis de señales de audio

Alberto Isaac Pico Lara, Diana Angélica Martínez Hernández

Repositorio en github del proyecto, notebook del proyecto

Resumen

Alguna vez has comparado tus gustos musicales con los de tu familia y amigos, te has puesto a pensar ¿por qué te gusta tal música y no otra? La respuesta está en tu construcción como persona basada en tus experiencias y contexto, pero también es resultado de un proceso psicológico y cognoscitivo particular. Los humanos son buenos para descubrir qué música suena similar. Al hacerlo, desarrollamos un gusto por la música que nos gusta. Las computadoras, por otro lado, son menos capaces de comprender la similitud musical. El ser humano puede determinar con facilidad la diferencia y la similaridad entre géneros musicales; sin embargo, existen patrones similares entre ellos que pueden resultar imperceptibles para el que la escucha. Por medio de análisis de señales se buscará similaridad entre ellos, así mismo averiguar si un modelo de aprendizaje automático es capaz de <<diferenciar>> (como tarea de clasificación o clusterización) entre géneros musicales según las características de la señal de audio de la música.

Introducción

Alguna vez has comparado tus gustos musicales con los de tu familia y amigos, te has puesto a pensar ¿por qué te gusta tal música y no otra? La respuesta está en tu construcción como persona basada en tus experiencias y contexto, pero también es resultado de un proceso psicológico y cognoscitivo particular. Desde la psicología esto se define como la percepción musical, que se refiere a "la integración de las variables físicas del sonido con procesos cognoscitivos como el aprendizaje, la memoria y la emotividad en conjunto con un contexto estético y sociocultural determinado", señala la maestra Concepción Morán, de la Facultad de Psicología de la UNAM.

Quiere decir que el significado que le das a la música es producto de una reconstrucción personal que se encuentra en constante cambio según las situaciones que te rodean, ocasionando que otorgues a cierto tipo de música significados particulares, cuyos efectos no siempre son los mismos en ti en distintos momentos de tu vida.

Es así como la percepción musical es diferente según tu historia personal, la cual puede determinar en cada persona ciertas reacciones emocionales, como la alegría o la tristeza, a distintas piezas musicales, razón por la cual elegimos un género u otro según nos proporcione bienestar o tranquilidad.

Prácticamente a diario utilizamos plataformas para escuchar nuestras canciones favoritas, y muchos de los sistemas de recomendación de música tienden a agrupar canciones por idioma, país del artista o popularidad.

Esto hace difícil descubrir nueva música en base a cómo suena. Por ejemplo, si a una persona le gusta la música rítmica y urbana, seguramente las recomendaciones entrarán en el bucle del Reggaetón si ya ha guardado algunas canciones de este género. Por este motivo, recomendar y agrupar canciones en base a sus características, podría permitir conocer nueva música, aunque no sea muy popular o cercana geográficamente.

Por otro lado, descubrir si realmente a las personas que dicen gustarle "un poco de todo", en realidad, son frecuencias y patrones con ciertas características en común lo que le hace atractivo para la persona.

Hipótesis

El ser humano puede determinar con facilidad la diferencia y la similaridad entre géneros musicales; sin embargo, existen patrones similares entre ellos que pueden resultar imperceptibles para el que la escucha. Por medio de análisis de señales se buscará similaridad entre ellos, así mismo averiguar si un modelo de aprendizaje automático es capaz de <<diferenciar>> (como tarea de clasificación o clusterización) entre géneros musicales según las características de la señal de audio de la música.

Estado del arte

Los humanos son buenos para descubrir qué música suena similar. Al hacerlo, desarrollamos un gusto por la música que nos gusta. Las computadoras, por otro lado, son menos capaces de comprender la similitud musical. Por ejemplo, sistemas de recomendación de música como el algoritmo Discover Weekly de Spotify pueden medir la similitud calculando la frecuencia con la que aparecen las canciones en listas de reproducción juntos en lugar de usar la música en sí.

El objetivo de dicho proyecto fue utilizar una técnica de aprendizaje profundo llamada Inferencia Adversarially Learned para aprender una incorporación de métrica en canciones y compararla con varias incorporaciones de métricas de referencia que no dependen del aprendizaje profundo, como PCA. En otras palabras, queremos aprender una función que genera una representación de menor dimensión de una canción de entrada sin hacer uso de etiquetas para esas canciones. Las representaciones de menor dimensión de las canciones hacen que determinar la similitud de las canciones sea trivial y, por lo tanto, tienen aplicaciones a una amplia variedad de análisis y recomendaciones musicales problemas. Nosotros, sin embargo, manejaremos una clasificación menos profunda y con sólo tres géneros y 390 canciones; usando un método de clusterización.

Implementación en Python de un sistema sobre una Raspberry Pi capaz de controlar una matriz 8x8 de LED de colores RGB en función de un sonido de entrada, a través de un cable de audio, en tiempo real. Se intenta simular un efecto de sinestesia entre el sonido y la imagen. Se introduce la base teórica del procesado digital de señales, con especial énfasis en su aplicación a las señales de audio. Se explican diferentes conceptos como el de los filtros mel, esenciales en el funcionamiento básico del sistema a implementar. Al ser un trabajo interdisciplinar, también aparecerán conceptos musicales. La señal de audio de entrada se va aventajando y se procesa como un stream en el dominio de la frecuencia por bandas mediante un filtrado mel, calculando la potencia por bandas y determinando el nivel de LEDs que se encienden en la matriz. Cada columna corresponde a una banda de frecuencia, correspondiéndole el nivel de potencia con las filas. Respecto a los colores, se dispone de un modo monofónico (modo pitch) basado en la detección de la frecuencia fundamental y un modo polifónico (modo default) basado en la tímbrica espectral, es decir, en las posiciones relativas de los centroides espectrales por banda.

Por nuestra parte, en la investigación tomaremos las características de la señal enfocadas para un análisis de audio específicamente.

Enlaces:

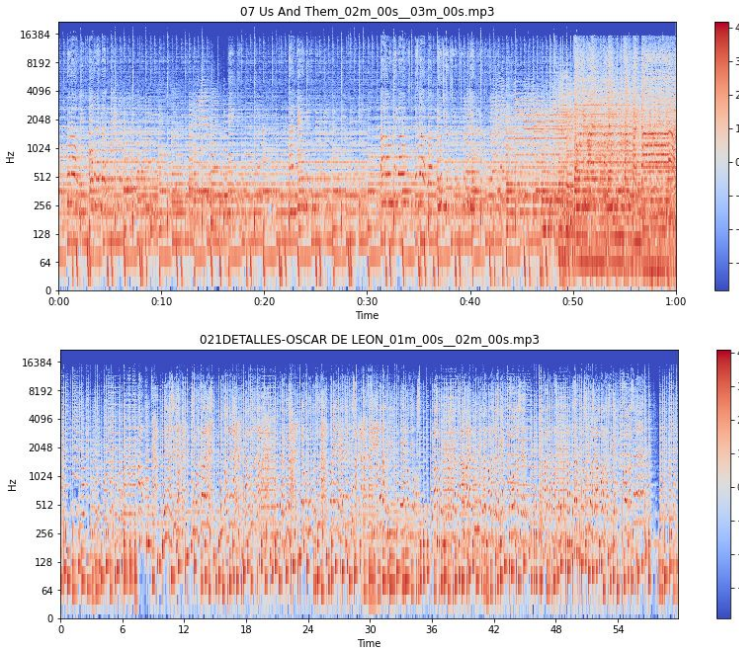
song2vec: Determining Song Similarity using Deep Unsupervised Learning

Audio Fingerprinting with Python and Numpy

Implementación de dispositivo de procesado de señal de audio y generación de excitación a LEDs individualmente direccionables

Tempo and beat analysis of acoustic musical signals

Metodología (objetivos)

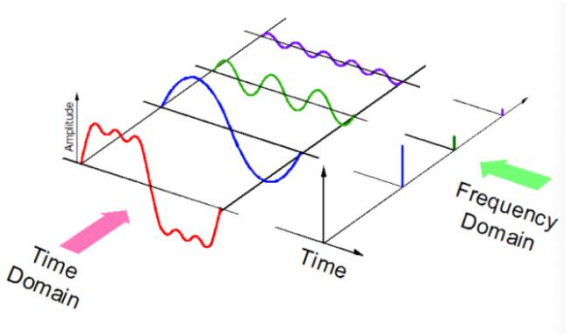


1.Entendimiento del caso de estudio

Como se mencionó anteriormente, el tema del gusto de la música no sólo es un tema psicológico de los humanos sino también un tema de análisis para las computadoras. Teniendo esto como antecedente se busca encontrar patrones en las señales de cada género creando así una justificación a las señales que nuestro cerebro recibe según el estado de ánimo, en el presente trabajo sólo se tocará la parte computacional, es decir, la parte analítica de las canciones y no del ser humano.

Para el presente trabajo se obtuvieron 130 canciones de tres géneros diferentes como el rock progresivo, la salsa y la música electrónica teniendo un total de 390 canciones en formato mp3.

El sonido se representa en forma de una señal de audio con parámetros como frecuencia, ancho de banda, decibelios, etc. Una señal de audio típica se puede expresar en función de la amplitud y el tiempo.



Estos sonidos están disponibles en muchos formatos, lo que hace posible que la computadora los lea y analice. Algunos ejemplos son: mp3, WMA (Windows Media Audio), wav (Waveform Audio File).

Por lo tanto, se realizará un análisis exhaustivo de las siguientes características de las 390 canciones que se reunieron para este trabajo.

Tasa de cruces con cero: es la tasa de cambios de signo a lo largo de una señal, es decir, la tasa a la que la señal cambia de positivo a cero a negativo o de negativo a cero a positivo. Esta función se ha utilizado mucho tanto en el reconocimiento de voz como en la recuperación de información musical, siendo una función clave para clasificar los sonidos de percusión.

ZCR es definido formalmente como:

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_{\mathbb{R}_{<0}}(s_t s_{t-1})$$

Donde s es una señal de longitud y T and $1_{\mathbb{R}_{<0}}$ es una función indicadora.

En algunos casos, solo se cuentan los cruces "positivos" o "negativos", en lugar de todos los cruces, ya que, lógicamente, entre un par de cruces por cero positivos adyacentes debe haber uno y solo un cruce por cero negativo.

Centroide Espectral: es una medida utilizada en el procesamiento de señales digitales para caracterizar un espectro. Indica dónde se encuentra el centro de masa del espectro. Perceptualmente, tiene una conexión robusta con la impresión de brillo de un sonido.

Se calcula como la media ponderada de las frecuencias presentes en la señal, determinada mediante una transformada de Fourier, con sus magnitudes como ponderaciones:

$$\text{Centroid} = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$$

donde x (n) representa el valor de frecuencia ponderado, o magnitud, del número de intervalo n , y f (n) representa la frecuencia central de ese intervalo.

Caída Espectral: Un extractor de características que extrae el punto de caída espectral. Esta es una medida de la cantidad de sesgo a la derecha del espectro de potencia.

El punto de caída espectral es la fracción de contenedores en el espectro de potencia en la que el 85% de la potencia se encuentra en frecuencias más bajas.

Cepstrum de Frecuencia Mel: el cepstrum de frecuencia mel (MFC) es una representación del espectro de potencia a corto plazo de un sonido, basado en una transformada de coseno lineal de un espectro de potencia logarítmica en una escala de frecuencia mel no lineal .

Los coeficientes cepstrales de frecuencia Mel (MFCC) son coeficientes que forman colectivamente un MFC. Se derivan de un tipo de representación cepstral del clip de audio (un "espectro de un espectro" no lineal). La diferencia entre el cepstrum y el cepstrum de frecuencia mel es que en el MFC, las bandas de frecuencia están igualmente espaciadas en la escala mel, lo que se aproxima más a la respuesta del sistema auditivo humano que las bandas de frecuencia espaciadas linealmente que se usan en el cepstrum normal. Esta deformación de frecuencia puede permitir una mejor representación del sonido, por ejemplo, en la compresión de audio.

Los MFCC se obtienen comúnmente de la siguiente manera:

- Tomar la transformada de Fourier de (un extracto en ventana de) una señal.
- Mapear las potencias del espectro obtenido anteriormente en la escala mel , utilizando ventanas triangulares superpuestas .
- Tomar los registros de las potencias en cada una de las frecuencias mel.
- Tomar la transformada de coseno discreta de la lista de potencias mel log, como si fuera una señal.
- Los MFCC son las amplitudes del espectro resultante.

Puede haber variaciones en este proceso, por ejemplo: diferencias en la forma o el espaciado de las ventanas utilizadas para mapear la escala, o adición de características dinámicas como "delta" y "delta-delta" (primera y segunda -orden de diferencia trama a trama) coeficientes.

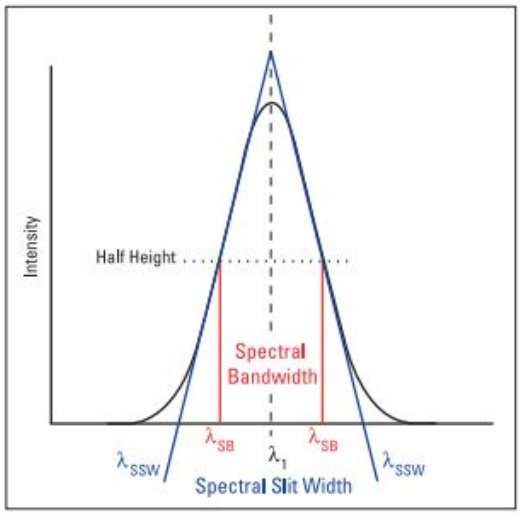
Los valores de MFCC no son muy robustos en presencia de ruido aditivo, por lo que es común normalizar sus valores en los sistemas de reconocimiento de voz para disminuir la influencia del ruido. Algunos investigadores proponen modificaciones al algoritmo básico de MFCC para mejorar la robustez, como elevar las amplitudes log-mel a una potencia adecuada (alrededor de 2 o 3) antes de tomar la DCT (Transformada de coseno discreta), que reduce la influencia de baja componentes energéticos.

Frecuencia cromática: se relaciona estrechamente con las doce clases de tonos diferentes. Las características basadas en croma, que también se conocen como " perfiles de clase de tono ", son una herramienta poderosa para analizar música cuyos tonos se pueden categorizar de manera significativa (a menudo en doce categorías) y cuya afinación se aproxima a la escala de temperamento igual. Una propiedad principal de las características cromáticas es que capturan las

características armónicas y melódicas de la música, al tiempo que son resistentes a los cambios en el timbre y la instrumentación.

Tempograma: se refiere a la velocidad de una pieza musical. Más precisamente, el tempo se refiere a la frecuencia del ritmo musical y viene dado por el recíproco del período de tiempo. El tempo a menudo se define en unidades de latidos por minuto (BPM).

Ancho de banda espectral: El ancho de banda espectral se define como el ancho de la banda de luz a la mitad del pico máximo (o ancho completo a la mitad del máximo [FWHM]) y se representa por las dos líneas rojas verticales y λSB en el eje de longitud de onda. Ancho de rendija espectral, un término relacionado al que se hace referencia con frecuencia en Regulaciones EP, se define como la extensión total de longitudes de onda representado por las líneas azules y λSSW. El ancho de banda espectral del instrumento siempre será más estrecho que el ancho de la rendija espectral.



El ancho de banda espectral de un espectrofotómetro está relacionado con el ancho de rendija física del monocromador y, por tanto, a las capacidades de resolución del instrumento.

Se sugiere revisar en el notebook de jupyter para complementar.

2.Entendimiento de los datos

Antes de poder explorar , tenemos que transformar los datos en crudo de las muestras de audio en representaciones más significativas. Es por ello que todas las pistas fueron cortadas para que tuvieran la misma duración de un minuto y la misma tasa de bits por segundo fijada en 128 kbps.

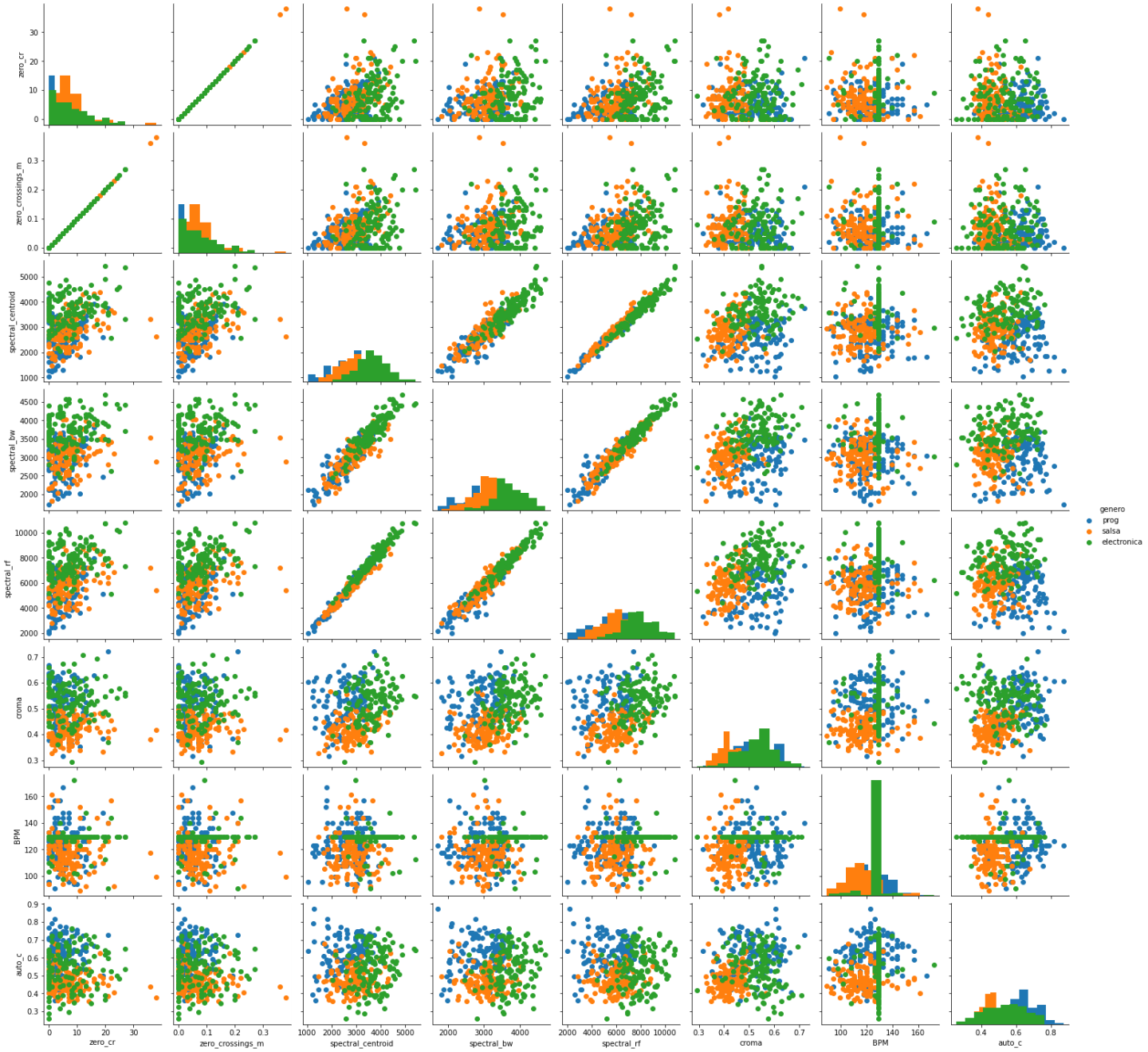
Para el análisis y extracción de características de las pistas de audio se utilizó la librería Librosa que proporciona un conjunto de herramientas para la recuperación de información espectral de señales de audio. Con la herramienta load de librosa se obtiene el vector de cada pista de audio, al cual se le extrajo las siguientes características:

- Tasa de Cruces con cero. El tipo de dato es un número entero
- Media del vector de centroides espectrales. El tipo de dato es un número decimal.
- Media del vector del ancho de banda espectral. El tipo de dato es un número decimal.
- Media del vector de caída espectral. El tipo de dato es un número decimal.
- Media del vector de las frecuencias cromáticas. El tipo de dato es un número decimal.
- Media de cada uno de los 20 vectores del Cepstrum de Frecuencia MEL. El tipo de dato es un número decimal.
- Auto Correlación del tiempo en la rítmica. El tipo de dato es un número decimal.
- El BPM estimado. El tipo de dato es un número decimal.

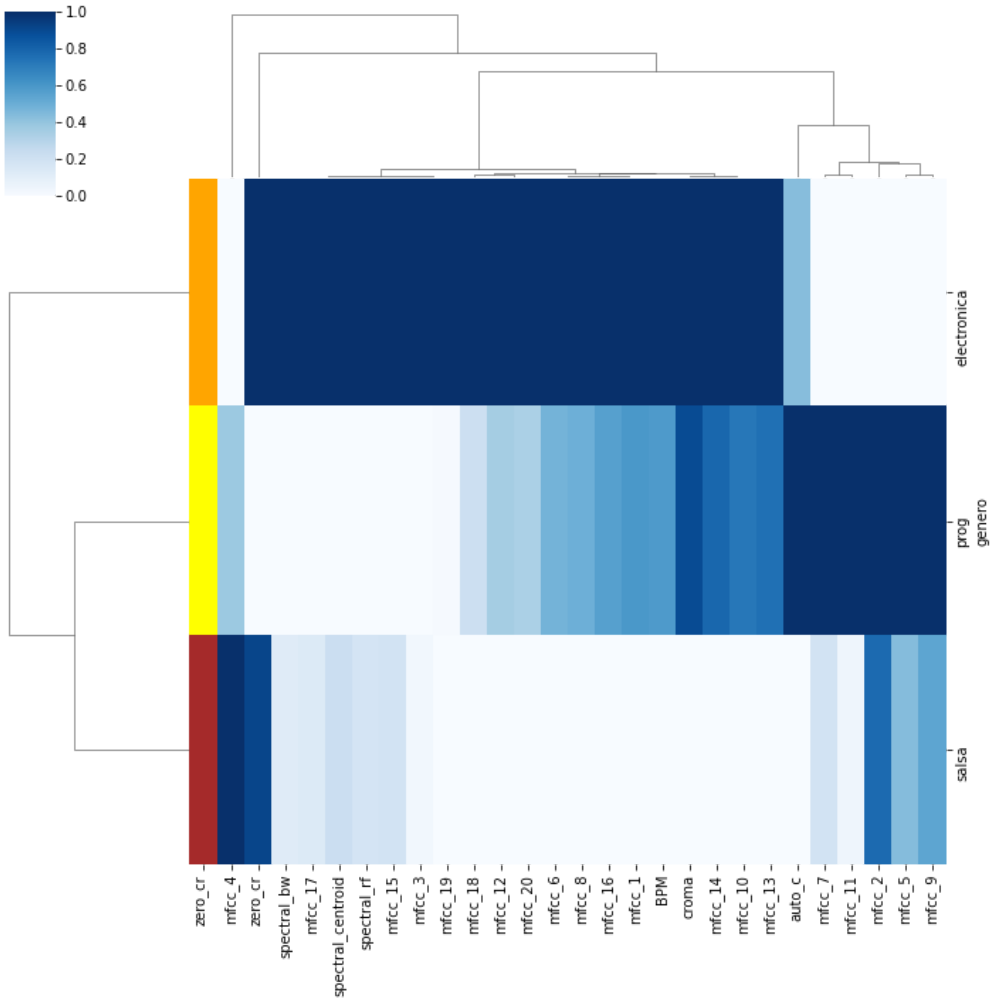
Dando como resultado 27 variables por cada una de las pistas, que, junto con el nombre de la pista y la etiqueta de género nos dan un total de 29 variables de las que el nombre y la etiqueta de género son variables cualitativas de tipo nominal y ordinal respectivamente mientras que las otras 27 variables son cuantitativas con la tasa de cruces con cero como variable discreta y el resto variables continuas, todas de intervalo.

Posterior a esta extracción se escribió el objeto contenedor de esta información a un archivo CSV para su posterior carga.

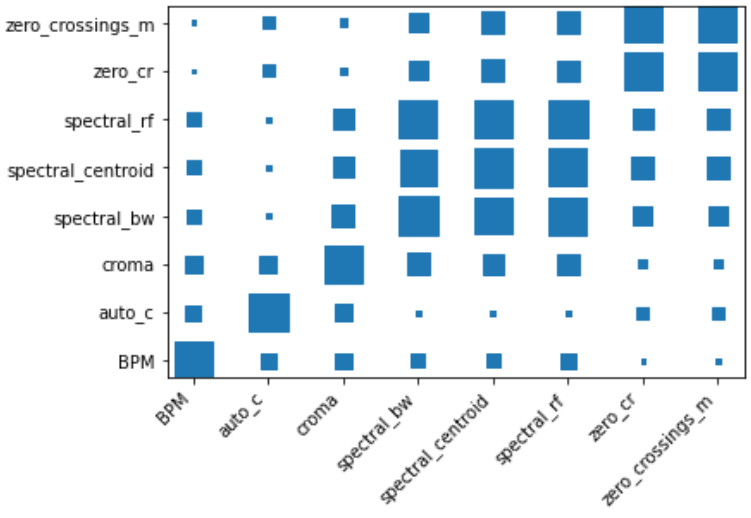
En el siguiente gráfico se muestran las características de los datos para las variables numéricas (dar clic sobre la imagen para verla con mayor detalle)



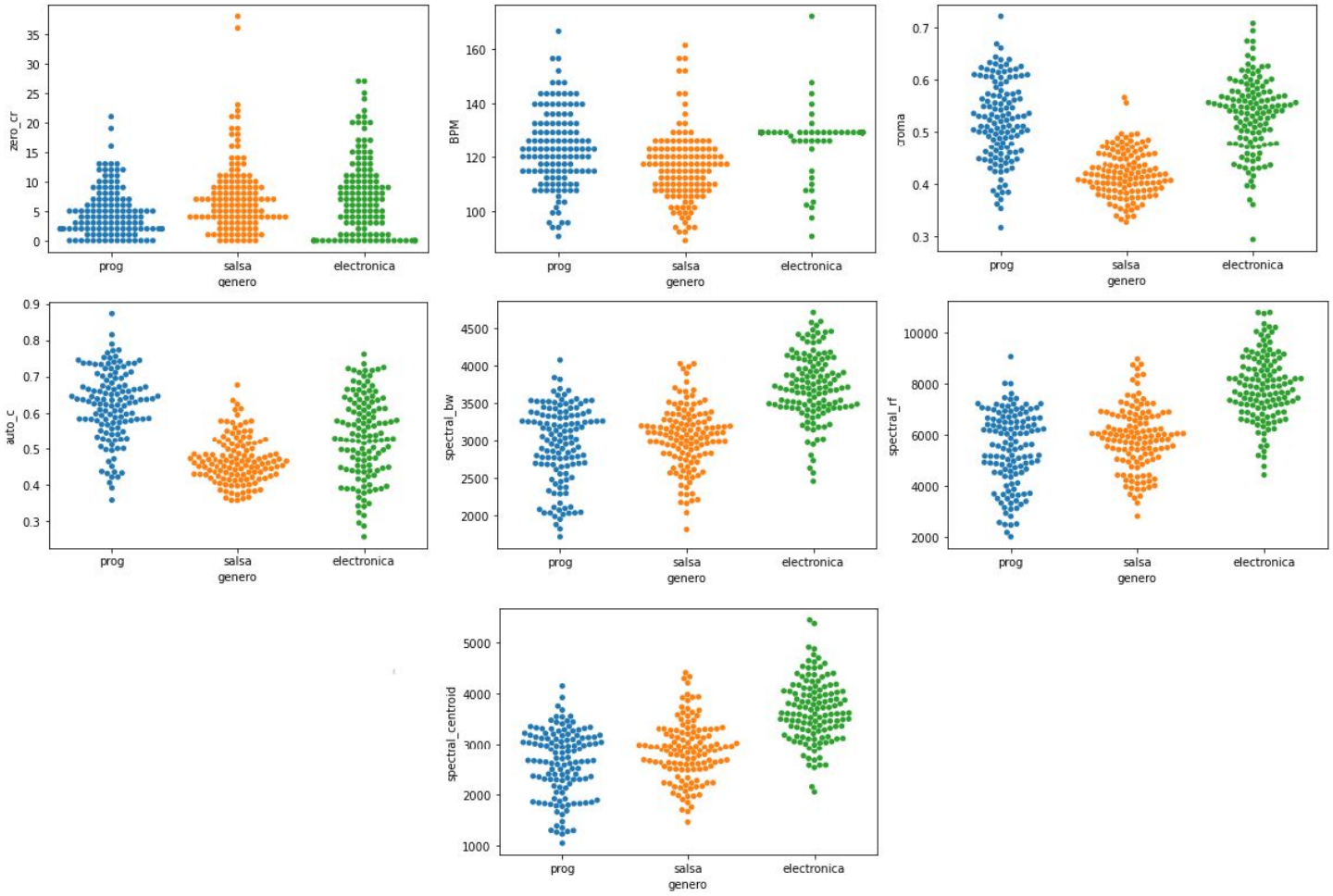
Así, mismo en el siguiente dendrograma se puede apreciar la correlación que existe entre las variables en cada uno de los géneros seleccionados



Y la correlación entre cada una de las variables de manera global



En el siguiente gráfico se puede observar la representación de la distribución de los valores para los valores espectrales y de rítmica por cada género.



Pre procesamiento de datos

Se crearon dos variables una variable que contenga todas las características y otra que sea el "target". Una vez separadas se procede a estar las columnas numéricas con MinMaxScaler.

Modelado

Se aplicaron tres metodologías diferentes para evaluar el mejor modelo 2 de clusterización (K-vecinos, K-means) y uno de clasificación (Red neuronal). Todos los entrenamientos se realizaron con 70% de entrenamiento y 30% prueba.

K vecinos

Con GridSearchCV se obtuvo un rango de 3- 66 n-neighbors con un peso uniforme siendo 16 n-neighbors el parámetro con mejor optimización para el modelo.

Kmeans

En este método se realizó la curva de Elbow para obtener la distancia media entre los puntos de datos y su centroid. Se tiene la siguiente curva la cual visiblemente es "suave", considerando a 3 como un buen número para K.

Red Neuronal

Con GridSearchCV se obtuvo lo siguiente como parámetro óptimo:

Función de activación: Tangente hiperbólica
Tamaños de las capas ocultas: (10,60,10)
Tasa de aprendizaje: adaptive

Evaluación

Se evaluaron y compararon los resultados de precisión de cada modelo teniendo los siguientes resultados:

Precisión	K vecinos	Red neuronal
Train	0.8352	0.9414
Test	0.8889	0.9402

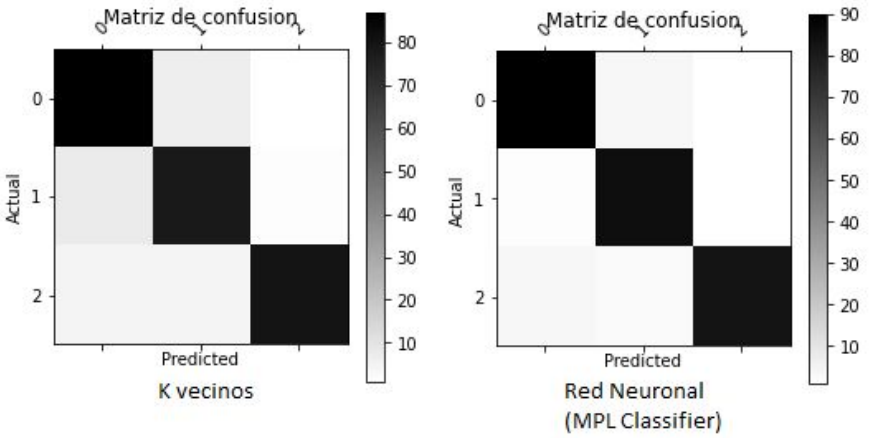
El modelo de K means no fue posible evaluarlo con precisión como los modelos anteriores por lo que se prosiguió a realizar una simulación para conocer los resultados de dicho modelo, proponiendo para los géneros progresivo, salsa y electrónica etiquetas como 0, 1 y 2 respectivamente.

	archivo	genero	label
0	01-TheMountain_02m_00s__03m_00s.mp3	prog	1
1	01ANightmareToRemember_02m_00s__03m_00s.mp3	prog	2
2	01AcidRain_02m_00s__03m_00s.mp3	prog	2
3	01BalrogBoogie_02m_00s__03m_00s.mp3	prog	1
4	01C'estlaVie_02m_00s__03m_00s.mp3	prog	2
...	...	...	...
385	WhenLoveTakesOver_01m_00s__02m_00s.mp3	electronica	1
386	WhenTheFunkDrops(OriginalMix)_01m_00s__02m_00s...	electronica	2
387	Wintermix_01m_00s__02m_00s.mp3	electronica	2
388	Yee_01m_00s__02m_00s.mp3	electronica	0
389	YouMakeMe_01m_00s__02m_00s.mp3	electronica	2

Validación

Se calcularon las matrices de confusión de los modelos de K vecinos y Red neuronal teniendo como resultado lo siguiente:

En tanto en K vecinos como en red neuronal se puede notar alto grado de sensibilidad y precisión lo cual nos indica que ambos modelos son buenos para la el conjunto de datos que tenemos, sin embargo, dadas las escalas la red neuronal muestra un mejor desempeño para el tratado, evaluación y predicción a nuestro conjunto de datos.



Implementación

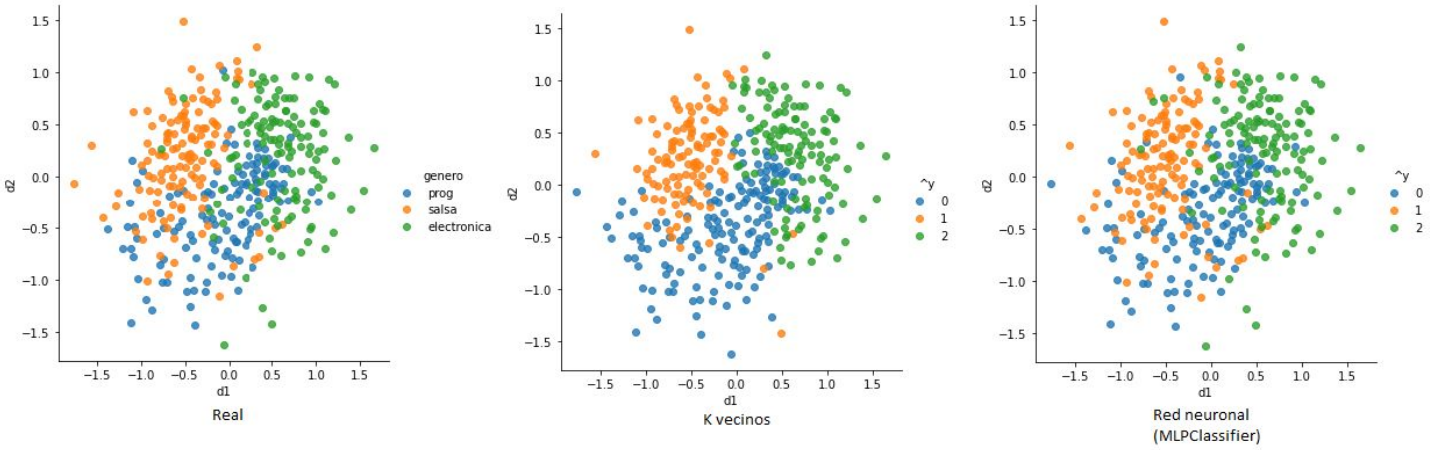
Para la implementación o comprobación práctica del modelo, se le introdujo un nuevo dataset con 3 nuevas canciones de cada género para saber si el modelo era capaz de analizar y clasificar el género al que pertenece cada canción nueva.

Como era de esperarse el modelo los clasificó con éxito, asignando cada género a cada canción desconocida por medio de sus propiedades de sonido que previamente fueron extraídas.

Resultados

Como resultados podemos presentar que el mejor modelo para la clasificación e identificación de señales de audio fue una Red Neuronal (MLP Classifier) ya que tiene mayor porcentaje de sensibilidad y precisión para la predicción de resultados. Sin embargo, K vecinos no deja de ser una buena opción para la clusterización de datos. Posteriormente,se trataron y

transformaron las dimensiones de modo que quedaran sólo dos para que estas fueran visualizadas y comparadas para la comprobación de aprendizaje.



Por otro lado, la red neuronal como se mencionó con anterioridad fue puesta a prueba con datos nuevos sin previa clasificación pero con selección de mismas características se notó una gran generación de conocimientos ya que fue capaz de identificar exitosamente el género al que pertenecía cada pista nueva.

Conclusión y discusión

El modelo cumplió su hipótesis de clasificar exitosamente una canción desconocida siguiendo los patrones de las características del sonido en tres géneros estudiados: Rock progresivo, Electrónica y Salsa. Sin embargo, consideramos que nos falta mucha experiencia y estudio para llevarlo a un análisis profundo de similaridad entre ambos géneros ya que se requiere de mayores conocimientos profundos sobre deep learning. Por otro lado, este trabajo abrió un mundo de oportunidades para el análisis de la señal del sonido, uno de ellos el análisis de imagen con el uso de los espectrogramas generados en este trabajo, dejando pendiente una gran investigación de encontrar similitudes entre géneros. Una aplicación posterior podría ser etiquetar las canciones por SÍ me gusta y NO me gusta, este podría ser un impulso bastante importante para los sistemas de recomendación basados en coincidencia de pistas con otras listas de reproducción ya que estas pistas siguen siendo hechas por personas en las que el contexto cultural, social y psicológico puede influir a la hora de elegir una canción de su agrado; empero del análisis de señales que puede encontrar esas características sui generis de los gustos de un usuario en particular. Por otro lado es un tema del cual se podrían recorrer muchos caminos llegando incluso a entrenar una inteligencia artificial para que haga canciones de cierto género dados los patrones de la señal de audio, pero ese será un trabajo posterior, cualquiera de estas dos salidas sería interesante y valdría la pena tratarlas a posteriori en una tesis.

Sólo queda agradecer al profesor por compartirnos el conocimiento, así como impulsarnos y motivarnos a realizar este trabajo, que sin duda tiene muchísimo potencial en un futuro.