Facultad de Informática
UNIVERSIDADE DA CORUÑA

TRABAJO APRENDIZAJE AUTOMÁTICO
GRADO EN INGENIERIA INFORMÁTICA
MENCIÓN EN COMPUTACIÓN

CATEGORIZACIÓN DE GÉNEROS MUSICALES A TRAVÉS DEL PROCESAMIENTO DE SEÑALES DE AUDIO

Estudiante: Victor Nathanael Badillo Aldama

Estudiante: Juan Domínguez Rodríguez

Estudiante: Diego Dopazo García

Estudiante: Pablo Legide Vidal

Estudiante: Pablo López Martínez

Estudiante: Javier Rodríguez Rodríguez

A Coruña, noviembre de 2024.

Índice general

\

Capítulo 1

Introducción

En la escena musical actual, existe una gran diversidad de géneros musicales que no solo muestran su gran complejidad, sino que también reflejan la diversidad cultural y las múltiples perspectivas de la sociedad. Con el surgimiento de plataformas de streaming como Spotify o Apple music, se ha producido una diversificación y proliferación de distintos géneros musicales que desafían las clasificaciones tradicionales dando lugar a nuevas formas de etiquetar la música. Desde el hip-hop hasta la música electrónica, pasando por el indie y el reguetón, la música abarca una amplia gama de estilos y sonidos que reflejan las diversas culturas y experiencias humanas, así como sus gustos cambiantes. En este contexto, la clasificación de géneros musicales se ha convertido en un campo complejo y en constante evolución, donde cada día surgen nuevas etiquetas y subgéneros que demuestran esta gran innovación dentro de la producción musical.

Existen una gran serie de soluciones que pretenden ayudar a resolver el complejo problema que presenta esta industria, destacando entre ellos la inteligencia artificial. El uso de las nuevas tecnologías y descubrimientos relacionados con este campo, brinda la oportunidad de facilitar esta tarea de clasificar la amplia diversidad musical, permitiendo a los usuarios explorar y descubrir nuevos géneros de manera más eficiente y personalizada. A través de algoritmos avanzados, la IA podría llegar a generar composiciones originales y pronosticar éxitos consiguiendo adaptar la experiencia auditiva del usuario y alterando significativamente los procesos de producción, distribución y consumo de música establecidos hasta la fecha.

Estos algoritmos se emplean concretamente en el campo del aprendizaje automático y permiten a las máquinas aprender patrones a partir de datos y realizar tareas específicas sin intervención humana directa. Esta rama de la inteligencia artificial se basa en la capacidad de los sistemas informáticos para mejorar automáticamente su rendimiento a través de la experiencia, sin necesidad de programación explícita para cada tarea. El aprendizaje automático

abarca una amplia variedad de técnicas, y se utiliza en una variedad de aplicaciones, incluyendo reconocimiento de voz, procesamiento de imágenes, análisis de texto, y mucho más. Su uso ha crecido significativamente en los últimos años debido a los avances en el poder computacional, el acceso a grandes conjuntos de datos y el desarrollo de algoritmos más sofisticados.

En este contexto, este proyecto se enfoca en el desafío del reconocimiento de géneros musicales a través del aprendizaje automático basado en el análisis de la onda de sonido generada. Esta iniciativa proporcionará la capacidad de distinguir entre una amplia variedad de géneros y ofrecerá una ayuda sencilla y eficaz para cualquier usuario amante de la música. Para ello algoritmos de aprendizaje automático tratarán de reconocer correlaciones entre diferentes características de la música, como el ritmo, la instrumentación y la estructura armónica, para poder clasificar de manera precisa los distintos géneros musicales. Esta capacidad de análisis avanzado no solo ofrece una herramienta eficiente para etiquetar y organizar la música, lo que mejorará significativamente la experiencia del usuario al identificar y acceder rápidamente a los géneros de su preferencia, sino que también puede abrir nuevas oportunidades creativas al ayudar a los artistas a explorar y experimentar con diferentes estilos y tendencias musicales.

Se comenzará describiendo el problema detalladamente (2. Descripción del problema). Posteriormente se realizará un análisis de la bibliografía utilizada (3. Análisis Bibliográfico) la cual se podrá consultar (7. Bibliografía). Terminando por las conclusiones finales y la solución al problema propuesto (5. Conclusiones y 6. Trabajo Futuro).

Capítulo 2

Descripción del problema

Dado que el problema a resolver se centra en distinguir géneros musicales a través de audio, las principales restricciones son la duración de las pistas, que será la misma para todas, y las condiciones de grabación del audio, que serán diferentes para introducir diversidad en los resultados.

El dataset a utilizar será GTZAN Vulpe (2020), que es la base de datos pública más utilizada en investigación para el reconocimiento de géneros musicales. Esta consiste en una recopilación de 1000 pistas grabadas entre el 2000 y el 2001 que se pueden dividir en 10 géneros, correspondiéndole 100 pistas a cada uno de los géneros. Blues, música clásica, country, disco, hip hop, jazz, metal, pop, reggae y rock, son los géneros que se van a contemplar para poder etiquetar cada pista.

Las pistas son fragmentos de 30 segundos extraídos de canciones completas y están en formato wav codificadas con 16 bits. Además, provienen de diversas fuentes como CDs, emisiones radiofónicas y grabaciones de micrófono, introduciendo condiciones de grabación variadas. Al ser guardadas, todas las pistas han sido estandarizadas a una frecuencia de muestreo de 22050Hz y formato mono. Su fecha de grabación fue entre 2000 y 2001.

Respecto al tamaño de ventana utilizado en este problema para hacer la clasificación, se hará uso de la fórmula 2.1

$$WindowSize = Fragment - Overlap \tag{2.1}$$

Obteniendo un tamaño de ventana de 49152 muestras.

Como para la base de datos las clases están balanceadas, teniendo para cada género 100 pistas que pertenecen únicamente a ese género, la mejor métrica considerada para clasificar es la precisión. Esto es, debido al interés por evaluar el clasificador, para poder comprobar si el

sistema etiqueta las pistas de audio correctamente.

Otra medida considerada fue el F1-score debido a que las pistas de audio provienen de diversas fuentes y condiciones de grabación por lo que pueden contener ruido, dificultando la tarea del clasificador. Sin embargo, al no ser más importante la clasificación correcta de positivos que de negativos, se prioriza la precisión.

Además, dado que el problema tiene una naturaleza multiclase, utilizar la precisión para comparar no requiere del uso de ninguna estrategia como la macro y permite que cada género musical tenga el mismo peso a la hora de clasificar.

Capítulo 3

Análisis Bibliográfico

Existen numerosos trabajos en el campo del reconocimiento de géneros musicales que siguen una aproximación similar, entre ellos, el realizado por Lucas Rodríguez (2021), una alumna de la UPM que realizó el mismo trabajo que el propuesto por nuestro equipo como trabajo de fin de grado. En este trabajo se utilizan fragmentos de 30 segundos para distinguir entre géneros musicales distintos como: pop, música clásica, rock, metal y reggae.

Por otro lado existen otras aproximaciones en el mismo campo como la creada por Pimenta-Zanon, Bressan, y Lopes (2021), proponen un enfoque basado en redes complejas para la extracción de características y la clasificación de géneros musicales. Este método innovador convierte inicialmente las piezas musicales en secuencias de notas musicales para extraer medidas topológicas para caracterizar la topología de la red, lo que conforma un vector de características aplicable a la clasificación de géneros musicales.

Otros enfoques como el de Silla, Koerich, y Kaestner (2008) utilizan múltiples vectores de características a modo de clasificadores binarios, cuyos resultados se fusionan para asignarle a cada canción la etiqueta del género musical al que pertenece .Para poder llegar a este resultado final se lleva a cabo un procedimiento de combinación usando algoritmos de aprendizaje automático clásicos como Naïve-Bayes, Árboles de Decisión, k Vecinos Más Cercanos, Máquinas de Vectores de Soporte y Redes Neuronales de Perceptrón Multicapa. Para este trabajo se escoge una Base de Datos de Música Latina, que contiene 3,160 piezas musicales categorizadas en 10 géneros musicales.

Dentro del contexto internacional de este tipo de investigaciones, destaca el trabajo realizado por unos ingenieros del MIT los cuales usan un modelo de aprendizaje automático que replica el proceso sensorial de reconocimiento de géneros musicales, Trafton (2018). Lo interesante de este trabajo reside en la aproximación humana que realiza la inteligencia artificial imitando

la forma que tenemos de procesar los estímulos auditivos. El estudio tuvo un gran impacto en el sector y salió en la revista Neuron el 19 de abril de 2018.

En otra línea de investigación de este campo, el trabajo realizado por Liu, DeMori, y Abayomi (2022) destaca por su enfoque en el reconocimiento de conjuntos abiertos. Exploran la segmentación de clases de géneros conocidos y desconocidos utilizando los conjuntos de datos GTZAN y FMA. En cada caso, comienzan con una clasificación de géneros de conjunto cerrado y luego aplican métodos de reconocimiento de conjuntos abiertos. Presentan un algoritmo para la tarea de clasificación de géneros musicales utilizando OSR. Además, demuestran la capacidad de recuperar géneros conocidos, así como la identificación de patrones auditivos para géneros nuevos. Emplean una búsqueda en cuadrícula tanto en OpenMax como en softmax para determinar la precisión total óptima de clasificación para cada configuración experimental, y muestran la interacción entre el etiquetado de géneros y la precisión de reconocimiento de conjuntos abiertos.

Un artículo muy interesante dentro de este contexto es el publicado por Dutt (2022) en el cual hace una guía de como extraer características de audios para clasificar música en géneros musicales. Para ello indica entre que géneros clasifica, la base de datos que utiliza (GTZAN) y que características extraídas de los audios para poder clasificar los mismos y para ello utiliza el espectrograma, mel-espectrograma y MFCC.

Uno de los trabajos más recientes que encontramos en este campo fue el realizado por Aguilar Sánchez (2023). El cual se enfoca en analizar y comparar la utilización de una red neuronal convolucional y una red neuronal perceptrón multicapa aplicadas a la clasificación de audio por género musical.

Otro artículo creado por Guo, Gu, y Liu (s.f.). se basa en el conjunto de datos FMA (Free Music Archive). Utilizando características extraídas de estas pistas de música, se aplican diversas técnicas de aprendizaje automático para mejorar la precisión de la clasificación en más del 30% en comparación con el modelo base propuesto anteriormente.
Dicho informe comienza con una introducción al conjunto de datos utilizado y el proceso de preprocesamiento de los datos. Luego, se describen las técnicas de aprendizaje automático empleadas, que incluyen máquinas de vectores de soporte (SVM), regresión logística, vecinos más cercanos (KNN) y redes neuronales. Cada técnica se evalúa y se discuten los resultados obtenidos.
Se destaca que la selección de características y la regularización son aspectos críticos para mejorar el rendimiento de los modelos, especialmente en un conjunto de datos con una gran

cantidad de características. Además, se discuten los desafíos específicos enfrentados por algunas técnicas, como la tendencia al sobreajuste en SVM con kernel RBF y la maldición de la dimensionalidad en KNN.
Finalmente, se plantean futuras líneas de trabajo, que incluyen la exploración de técnicas para abordar el sobreajuste en SVM-RBF, la recopilación de más datos para equilibrar la distribución de géneros musicales y la aplicación de modelos de aprendizaje automático en escenarios reales. Se reconoce la contribución equitativa de todos los miembros del equipo en el proyecto, así como las referencias a estudios previos relevantes en el campo de la clasificación de géneros musicales.

Capítulo 4
Desarrollo

4.1 Aproximación I

4.1.1 Descripción

Para esta primera aproximación, se simplificará el problema original de clasificación de pistas en solo dos géneros musicales, que además tienen características muy diferentes: el metal y la música clásica. Para poder resolverlo, se van a utilizar varios modelos de aprendizaje automático que serán árboles de decisión, kNN, SVM y redes de neuronas artificiales.

Para ello se aplicará una Transformada Rápida de Fourier (FFT) a los segmentos de las pistas iniciales. Para poder utilizar la FFT, se requiere un número de muestras del audio que sea potencia de 2, en nuestro caso, usaremos 65536.

Para lograr esto, los datos originales de entrada pertenecientes a la base de datos, tienen que ser modificados para que las pistas originales, que tenían una duración de 30 segundos, se ajusten ahora a una duración de 2,97 segundos, por lo que el fragmento de audio inicial será dividido en segmentos de dicha duración con un *overlap* de 16384 muestras, consiguiendo así 14 segmentos por audio. Además al acotar el problema en dos géneros, se dispondrá en total de 200 pistas ya que en la base de datos original a cada género le corresponden 100 de las 1000 pistas en total, lo cual conllevará en 2800 segmentos de audio para esta primera aproximación.

Para obtener el tiempo que debe ocupar cada segmento se utiliza la fórmula 4.1

$$d = f_{FFT}/fs \tag{4.1}$$

Donde d es la duración, fs la frecuencia muestral de la pista y fFFT la frecuencia necesaria para la FFT.

Debido a que todas las pistas de la base de datos tienen la misma frecuencia muestral (22050Hz) y trabajaremos siempre con la misma fFFT, la duración de los segmentos será siempre la misma.

Para cada uno de estos segmentos, se calculará la media y desviación típica de la FFT, y dichos datos se guardarán en el archivo **genres.data** junto con el género al que pertenecen.

Para esta primera aproximación se utilizará la media y la desviación típica como parámetros para entrenar los diferentes modelos de aprendizaje automático ya mencionados. Cabe recalcar que se utilizarán únicamente los dos géneros anteriormente citados.

En primer lugar se cargan los datos de entrada que están almacenados en genres.data y para realizar la clasifición de las salidas de la red neuronal se utilizará la técnica de *one-hot encoding*, dado que los géneros están representados como texto y se requiere una representación binaria para las salidas.

A continuación, el conjunto de datos se somete a validación cruzada utilizando la técnica de *k-folds*. En este proceso, el conjunto de datos se divide en 10 particiones de tamaño igual. En cada iteración, 9 de estas particiones se utilizan para el entrenamiento del modelo, mientras que la partición restante se utiliza para la validación. Este proceso se repite 10 veces, de manera que cada partición se utiliza una vez como conjunto de validación. Esto garantiza una evaluación más robusta del modelo respecto al tradicional *hold-out* al promediar los resultados de las 10 iteraciones.

Seguidamente, se escogen y posteriormente se normalizan los datos a través de la técnica de normalización *Min-Max*. En este caso se decide normalizar la media y la desviación típica de la TTF para obtener consistencia en los datos(las canciones de metal tienen picos altos y bajos de frecuencias), que sea más fácil trabajar con estos mismos y para obtener mejores resultados.

Una vez escogida la arquitectura de los diferentes modelos de aprendizaje, el siguiente paso será entrenarlos. Como cada arquitectura consta de diferentes características, su entrenamiento será explicado en mayor detalle en la sección **4.1.2**

Con los diferentes modelos se pone a prueba con el conjunto de datos de test para ver su nivel de generalización y su precisión para clasificar.

4.1.2 Resultados

RNA

Durante el entrenamiento de la RNA se prueban configuraciones en las que se varía entre diferentes topologías, tratando de optimizar el rendimiento del modelo. Entre los hiperparámetros que serán utilizados se incluyen la tasa de aprendizaje (en nuestro caso 0.01), el número máximo de epochs de entrenamiento (decidimos que un número adecuado serían 1000), un tamaño de lote de 20, y se realiza el entrenamiento 50 veces para cada fold.

Tras realizar pruebas con 8 arquitecturas diferentes para las que se cambia el número de neuronas en cada capa oculta y variando el número de capas ocultas entre 1 y 2, a las que hay que sumarles las capas de entrada y salida, obtuvimos los resultados que se muestran en la tabla 4.1. Cabe destacar que la función de transferencia que usamos ha sido la función de activación sigmoidal(σ) será la *softmax*.

Cuadro 4.1: Resultados del modelo RNA en la aproximación 1

Topología	**Precisión (%)**	**Desviación estándar (%)**
[2, 1]	85.72	7.44
[7, 2]	92.86	4.04
[4, 4]	88.75	6.45
[5, 3]	92.82	3.11
[3]	93.625	1.01
[4, 6]	**95.53**	**1.29**
[2, 5]	88.08	6.63
[5]	94.14	0.86

De las 8 configuraciones probadas para esta iteración, la topología con la que obtuvimos mejores resultados es [4,6], con una precisión del 95.53%. Esto nos demuestra que es la arquitectura que más capacidad de generalización tiene, lo cual es positivo. Aunque en esta aproximación haya sido la topología con mejores resultados hay que tener presente que esto puede cambiar en futuras iteraciones.

kNN

Durante el entrenamiento del modelo de kNN, se exploran diferentes configuraciones cambiando un único hiperparámetro hasta encontrar la configuración óptima para este modelo. El hiperparámetro en este caso será el número de vecinos, k, que se cambiará en un rango de 4 a 10, como se puede apreciar en la tabla 4.2.

Cuadro 4.2: Resultados del modelo kNN en la aproximación 1

K	Precisión (%)	Desviación estándar (%)
4	95.46	0.91
5	95.36	0.86
6	95.54	0.74
7	95.50	1.11
8	95.21	0.83
9	**95.64**	**1.10**
10	95.54	0.93

La arquitectura con la que finalmente se obtuvieron los mejores resultados fue con k=9. Este valor de k proporcionó una precisión del 95.64%, con una desviación estándar de 1.10%.

De las 6 arquitecturas probadas esta es con la que se obtiene un mayor valor para la precisión, lo que se puede considerar un buen resultado. Para esta arquitectura también se obtiene la segunda desviación típica más alta, pero al ser la arquitectura con mayor precisión se demuestra como este modelo tiene una buena capacidad para generalizar y adaptarse a nuevos datos.

Sin embargo, esta arquitectura podría no ser la óptima para futuras iteraciones, sobre todo cuando se amplíe el número de clases en las que se puede clasificar el resultado o se usen nuevas métricas.

SVM

Para entrenar las Máquinas de Vectores de Soporte (SVM), se probaron diferentes configuraciones de hiperparámetros. Estos son, el tipo de kernel (como lineal, polinómico o radial)

y el parámetro de regularización C. Los resultados se pueden observar en la tabla 4.3.

Cuadro 4.3: Resultados del modelo SVM en la aproximación 1

Kernel	**C**	**Precisión (%)**	**Desviación estándar (%)**
rbf	1	94.82	0.91
linear	1	94.14%	0.86
poly	1	91.96	0.74
sigmoid	1	84.03	1.11
rbf	**200**	**94.92**	**0.83**
linear	200	93.96	1.10
poly	200	92.64	0.93
sigmoid	200	89.57	0.93

La configuración con la que se consiguieron los mejores resultados en esta aproximación fue con un kernel de tipo rbf (Función de base radial) y un valor de C particularmente alto como es 200, que a pesar de que podría llevar a un ajuste demasiado elevado.

En este caso la precisión obtenida es del 94.92% con una desviación típica de 0.93%. Realizando la distintas pruebas relativas a la aplicación de modelos de SVM a este problema, se apreció una mejora significa en cualquiera de los kernels a medida que se aumentaba el valor de C. Esto podría llevar a problemas debido a un posible sobreentrenamiento, pero gracias a que es utilizada la función de crossvalidation nos aseguramos que esto no suceda.

Es importante destacar que esta configuración tendrá que ser ajustada en futuras iteraciones del modelo, especialmente a medida que se añadan instancias y clases a la base de datos, ya que puede variar el rendimiento de esta configuración.

Árboles de decisión

Para poder entrenar el modelo de árbol de decisión, se utilizaron varias configuraciones de este mismo cambiando el hiperparámetro de profundidad máxima y dejando el hiperparámetro de estado aleatorio igual a 1 para todas las configuraciones para poder obtener un

resultado que dependa de la profundidad máxima y trabajar todo el rato con el mismo valor. Se utilizarán valores de profundidad máxima entre 1 y 6 en este caso, como se ve en la tabla 4.4.

Cuadro 4.4: Resultados del modelo de árboles de decisión en la aproximación 1

maxDepth	**Precisión (%)**	**Desviación estándar (%)**
1	93.17	1.15
2	93.17	1.15
3	94.50	0.94
4	94.92	1.13
5	**95.25**	**0.98**
6	94.92	1.33

Al usar esta arquitectura, se obtuvo una precisión del 95,25% para una profundidad máxima de 5 niveles, obteniendo una desviación estándar de 0,98%. Esta es la configuración que proporciona el mejor resultado. Con estos datos se deriva que una profundidad de 1 es más simple que el resto ya que al tener menos profundidad podría no ser suficiente para clasificar de la mejor manera. Al ir aumentando la profundidad máxima la precisión aumenta hasta cierto punto que empieza a decrementar que en este caso es para una profundidad de 6.

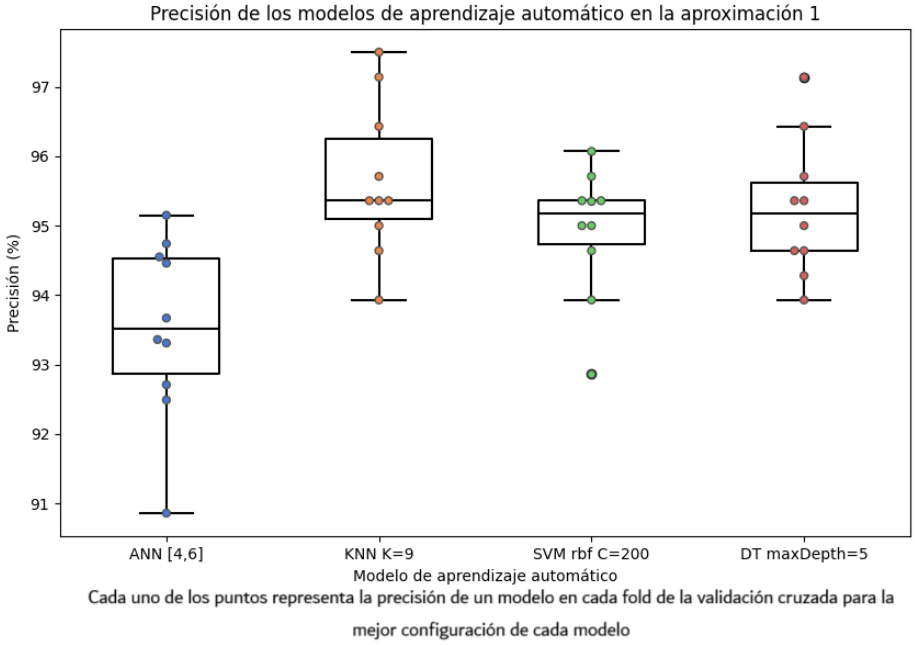
4.1.3 Discusión

La matriz de confusión representada en la tabla 4.5 se realiza en base al mejor fold del mejor modelo de esta aproximación. Como se puede observar el modelo se equivoca en pocas ocasiones, esto es debido a que la métrica elegida en esta aproximación tiene valores dispares entre los dos géneros a clasificar.

Género Real	Género Predicho	
	Classical	Metal
Classical	131	5
Metal	2	142

Cuadro 4.5: Matriz de confusión del kNN con k = 9

Figura 4.1: Comparación de la precisión de los distintos modelos de aprendizaje automático



Respecto al modelo para el cual se obtienen los mejores resultados en este caso es para el de kNN, que como se puede apreciar en la gráfica 4.1, tiene la mayor precisión respecto al conjunto de test, lo que significa que es el modelo que mejor clasifica en este caso. Concretamente la mejor precisión obtenida para este modelo es de, 95.64% a la vez que una desviación típica de 1.10%, que a pesar de ser más alta que el mejor resultado del resto de modelos, sigue considerándose un valor bajo. También cabe destacar que este modelo es el que proporciona los mejores resultados independientemente de la configuración, ya que todos sus resultados rondan el 95% de precisión, para cualquiera de los valores probados de k.

En futuras aproximaciones, a medida que se extraigan más características de los segmentos y se añadan nuevos géneros a la clasificación, la complejidad aumentará. Entre las métricas que se podrían estudiar destaca el Root Mean Square o RMS, que mide la magnitud promedio de una señal, y permitirá distinguir los géneros mas ruidosos de los más armónicos.

4.2 Aproximación II

4.2.1 Descripción

Para esta segunda aproximación, se añadirá un nuevo atributo extraído de los audios para clasificar entre los mismos. El atributo escogido será el RMS, siglas de Root Mean Square, que mide de la magnitud promedio de una señal, concretamente la intensidad o amplitud de una onda sonora. Por ejemplo, el rock y el metal tienden a tener RMS más altos debido a su naturaleza más ruidosa.

Se incluirá el género jazz entre los géneros a clasificar para esta nueva aproximación. Respecto a la anterior aproximación se hallan diferencias significativas como pasar de clasificación binaria a clasificación multiclase. Al haber otro atributo, la topología de la ANN será diferente al tener en cuenta el número de atributos, el número de neuronas en la capa de salida también sera diferente, en específico, será el mismo número de neuronas de salida que de atributos. Esto se debe al método para representar la salida en el caso de clasificación multiclase, utilizando la técnica de oneHotEncoding, en la que se representa la salida con un 1 para la clase perteneciente y un 0 para el resto de clases.

Los modelos de aprendizaje automático a utilizar serán los mismos que en la primera aproximación, es decir, árboles de decisión, kNN, SVM y redes de neuronas artificiales. También se realizará una prueba con diferentes parámetros para cada modelo diferente y se compararán los resultados de estos para esta aproximación y los resultados en comparación con la primera aproximación.

Para esta aproximación, los datos originales de entrada pertenecientes a la base de datos, se modificarán igual que para la anterior y las pistas de 30 segundos tendrán una duración de 2,97 segundos obtenida con la fórmula 4.1, dividiéndolo en 14 segmentos por audio. Sin embargo, al tratar ahora con 3 géneros en vez de dos, se dispondrá de 300 pistas divididas en 4200 segmentos aproximadamente con los que se trabajará en esta segunda aproximación.

4.2.2 Resultados

RNA

Durante el entrenamiento del modelo de la RNA, se realizaron pruebas mediante diferentes topologías, variando el número de neuronas por capa oculta y el número de capas ocultas entre 1 y 2. Después de las pruebas obtuvimos los siguientes resultados, que se pueden observar en la tabla 4.6. En esta aproximación se incorporó un nuevo atributo, el RMS, y una nueva posible salida, el género de jazz.

Cuadro 4.6: Resultados del modelo RNA en la aproximación 2

Topología	**Precisión (%)**	**Desviación estándar (%)**
[2, 1]	75.27	4.09
[7, 2]	81.75	2.14
[4, 4]	80.32	3.34
[5, 3]	81.88	2.87
[3]	82.86	1.41
[4, 6]	81.79	2.37
[2, 5]	79.24	2.64
[5]	**83.08**	**1.34**

La topología con la que mejor resultados se obtuvieron fue con 1 capa oculta, con 5 neuronas. Esta topología resultó en una precisión del 83.08%, con una desviación estándar de 1.34%.

En relación con la primera aproximación, es interesante notar que, aunque se logró una precisión ligeramente menor para la misma arquitectura, la desviación estándar también se redujo significativamente, dando lugar a resultados más consistentes entre sí. Esta disminución en la precisión se debe en gran medida a la introducción de un nuevo género a clasificar, el jazz.

kNN

Los valores obtenidos para esta aproximación con la incorporación de un nuevo atributo, en este caso el RMS, y una nueva posible salida, el género jazz, se pueden ver en la tabla 4.7.

Cuadro 4.7: Resultados del modelo kNN en la aproximación 2

K	Precisión (%)	Desviación estándar (%)
4	73.93	2.11
5	75.15	2.11
6	74.74	1.76
7	75.34	1.75
8	74.98	2.42
9	**75.44**	**1.92**
10	74.79	2.20

La arquitectura con la que finalmente se obtuvieron los mejores resultados fue con k=9. Este valor de k proporcionó una precisión del 75.44%, con una desviación estándar de 1.92%.

De las 6 arquitecturas probadas esta es con la que se obtiene un mayor valor para la precisión, lo que se puede considerar un buen resultado. Otro valor interesante es cuando el hiperparámetro k se asigna con un valor de 7. Para este caso se obtiene una precisión de 75.34% y una desviación estándar de 1.75%

En esta aproximación los valores de precisión son mucho mas bajos que en la anterior aproximación ya que al tener que clasificar entre un género más, se necesitan más datos para clasificar los audios correctamente.

SVM

Para entrenar las Máquinas de Vectores de Soporte (SVM), se probaron diferentes configuraciones de hiperparámetros. En este caso se varió el tipo de kernel probando el rbf, linear, poly y sigmoid a la vez que se fue modificando el parámetro de regularización C, hasta obtener los siguientes resultados, observables en la tabla 4.8.

Cuadro 4.8: Resultados del modelo SVM en la aproximación 2

Kernel	C	Precisión (%)	Desviación estándar (%)
rbf	**1**	**75.39**	**1.91**
linear	1	72.72	2.44
poly	1	64.02	1.97
sigmoid	1	50.14	1.24
rbf	200	75.37	1.86
linear	200	73.29	2.15
poly	200	66.43	1.59
sigmoid	200	40.49	1.15

La configuración con la que se consiguieron los mejores resultados en esta aproximación fue con un kernel de tipo rbf (Función de base radial) y un valor de C 1. Sin embargo, con un valor de C de 200 que podría llevar a un ajuste demasiado elevado, se obtiene un resultado parecido con una precisión ligeramente inferior, reduciendo tambien la desviación típica.

En este caso la precisión obtenida es del 75.39% con una desviación típica de 1.91%. En esta nueva aproximación, a diferencia de la anterior, subir el valor de C no significa necesariamente mejorar la precisión de todos los kernels, ya que como se puede apreciar para el kernel sigmoid esto no pasa, dando en ambos casos un resultado bastante bajo para la precisión. Además se puede ver como aumentar la complejidad del problema los valores para la precisión han disminuido notablemente a la vez que aumentan los de la desviación típica.

Es importante destacar que esta configuración tendrá que ser ajustada en futuras iteraciones del modelo, tratando de mejorar la precisión para las siguientes aproximaciones.

Árboles de decisión

Al igual que en la aproximación uno, se probarán los diferentes valores para la profundidad máxima desde 1 hasta 6 y se mostrarán los resultados en la tabla 4.9.

Cuadro 4.9: Resultados del modelo de árboles de decisión en la aproximación 2

maxDepth	**Precisión (%)**	**Desviación estándar (%)**
1	61.70	2.80
2	72.09	2.32
3	71.49	2.59
4	73.50	2.15
5	74.05	1.70
6	**74.39**	**1.39**

Al usar esta arquitectura, se obtuvo una precisión del 74.39% para una profundidad máxima de 6 niveles, obteniendo una desviación estándar de 1.39%. Esta es la configuración que proporciona el mejor resultado.

Con estos resultados se puede ver que la precisión ha disminuido considerablemente con la anterior aproximación. Al haber más atributos, se necesita una mayor profundidad, por eso una profundidad de 1 da un valor tan bajo para la precisión.

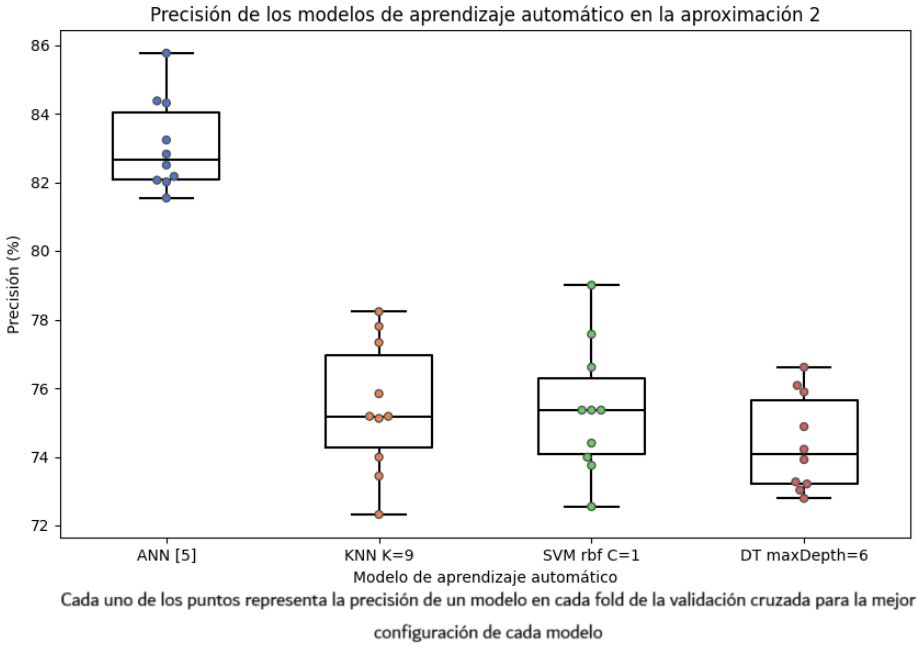
4.2.3 Discusión

La matriz de confusión representada en la tabla 4.10 se realiza en base al mejor fold del mejor modelo de esta aproximación, que es las RNAs. En esta aproximación ya se detectan más errores a la hora de clasificar entre los distintos géneros. En concreto los géneros que más confunde son el metal y el jazz.

Género Real	Género Predicho		
	Classical	Jazz	Metal
Classical	141	2	16
Jazz	1	96	26
Metal	9	33	95

Cuadro 4.10: Matriz de confusión de la RNA [5]

Figura 4.2: Comparación de la precisión de los distintos modelos de aprendizaje automático



Respecto a los resultados obtenidos se puede observar que la media de la precisión de los 10 folds del modelo RNA con la arquitectura [5] es la mayor, por lo que este modelo es el que

obtiene mejores resultados, como se aprecia en la gráfica 4.2. Concretamente la mejor precisión obtenida durante esta aproximación para este modelo es de 83.08%, con una desviación típica de 1.34%.

Para esta aproximación obtuvimos unos resultados inesperados debido a la inclusión de una nueva clase. El rendimiento de los modelos, se ve afectado notablemente, como se puede observar, reduciendo su precisión en aproximadamente un 20% a excepción de en las RNAs en las que no hay un descenso tan grande. En este caso, haber añadido una nueva métrica al conjunto de datos, parece no ser suficiente para conseguir clasificar correctamente los distintos géneros.

En la siguiente aproximación se añadirán nuevas medidas para intentar mejorar el porcentaje de la precisión sin aumentar excesivamente la complejidad de la tarea, es decir, añadiremos un solo género musical más, muy diferenciado de los actuales.

4.3 Aproximación III

4.3.1 Descripción

Para la tercera aproximación, se implementan dos nuevas métricas y se añade un único género más, buscando así mejorar la precisión respecto a la anterior aproximación.

Estas métricas son, en primer lugar, el MFCC, que mide las características espectrales representando la distribución de energía en diferentes bandas de frecuencia, según una escala de frecuencia *Mel* que se ajusta a la percepción auditiva humana. Esta métrica es útil para géneros como el jazz, ya que permite diferenciar las características tonales distintivas de los instrumentos como el saxofón o el piano, como se puede observar según la tabla 4.11.

La segunda métrica es, el Zero Crossing Rate indica la cantidad de veces que la señal de audio cruza el eje de tiempo y cambia de polaridad en un determinado intervalo de tiempo. Para géneros menos enérgicos como la música clásica, el ZCR será más bajo, mientras que en el metal este será más alto debido a la densidad y la intensidad de la señal de audio.Como se puede observar en la tabla 4.12 los valores para esta métrica, difieren de forma significativa entre géneros, por lo que se puede intuir que será útil a la hora de clasificar entre géneros.

Género	Métrica: MFCC			
	Media	Desviación Típica	Valor mínimo	Valor máximo
Classical	-3.12	2.78	-12.02	5.08
Jazz	0.33	2.41	-10.12	10.80
Hip-hop	3.18	2.44	-7.81	9.17
Metal	2.57	1.49	-4.77	7.37

Cuadro 4.11: Comparativa entre géneros para la métrica MFCC

Género	Métrica: Zero Crossing Rate			
	Media	Desviación Típica	Valor mínimo	Valor máximo
Classical	15.77	3.24	6.34	28.72
Jazz	17.69	2.43	7.52	24.09
Hip-Hop	20.83	1.65	13.94	25.76
Metal	21.96	1.56	13.51	26.14

Cuadro 4.12: Comparativa entre géneros para la métrica Zero Crossing Rate

Como en la anterior aproximación al añadir un género más, se produjo una bajada notable en la precisión, en este caso decidimos añadir un único género, asegurando que sus características fueran muy diferentes de los que ya se usaban.

En esta aproximación, al añadir un único género, se añade una clase más entre las que clasificar, por lo que sigue siendo un problema de clasificación multiclase. Se realizarán pruebas con diferentes hiperparámetros para cada uno de los modelos de aprendizaje automático y se comparará si los resultados han mejorado respecto a la segunda aproximación.

Los datos originales de entrada pertenecientes a la base de datos, se modificarán igual que para la anterior y las pistas de 30 segundos tendrán una duración de 2,97 segundos, obtenida gracias a la fórmula 4.1, dividiéndolo en 14 segmentos por audio. Al tratar con 4 géneros se dispondrá de 400 pistas divididas en 5600 segmentos.

4.3.2 Resultados

RNA

En esta nueva aproximación se volvieron a probar los mismo hiperparámetros y las misma topologías que probamos en las aproximaciones anteriores, obteniendo los resultados que se detallan en la tabla 4.13.

Cuadro 4.13: Resultados del modelo RNA en la aproximación 3

Topología	Precisión (%)	Desviación estándar (%)
[2, 1]	75.38	2.60
[7, 2]	85.33	1.76
[4, 4]	86.02	1.06
[5, 3]	85.98	2.51
[3]	86.44	1.06
[4, 6]	85.64	1.76
[2, 5]	83.42	2.20
[5]	**86.62**	**0.88**

En este caso podemos observar que la topología con la que se obtuvieron los mejores resultados consta de una capa oculta con 5 neuronas. Esta topología vuelve a ser la configuración con los mejores resultados respecto a la aproximación anterior, mejorando su precisión y su desviación estándar, siendo los valores de estas un 86.62% para la precisión y un 0.88% para la desviación estándar.

Esta mejoría en la precisión puede ser debida al aumento de parámetros de entrada, siendo estos además distintivos entre los distintos géneros a clasificar en esta iteración.

kNN

Los valores obtenidos para esta aproximación al introducir al problema el valor de la media y desviación típica para las dos nuevas métricas se puede observar en la tabla 4.14.

Cuadro 4.14: Resultados del modelo kNN en la aproximación 3

K	Precisión (%)	Desviación estándar (%)
4	**83.54**	**1.10**
5	83.39	0.80
6	82.80	1.63
7	83.48	1.04
8	82.91	1.29
9	82.53	1.49
10	82.68	1.54

La arquitectura con la que finalmente se obtuvieron los mejores resultados fue con k=4. Este valor de k proporcionó una precisión del 83.54%, con una desviación estándar de 1.10%.

De las 6 arquitecturas probadas esta es con la que se obtiene un mayor valor para la precisión, lo que se puede considerar un buen resultado. Otro valor interesante es cuando el hiperparámetro k se asigna con un valor de 7. Para este caso se obtiene una precisión de 83.48% y una desviación estándar de 1.04%

En esta aproximación los valores de precisión son notablemente mejores que en la anterior aproximación debido a las nuevas métricas añadidas, a pesar de haber incluído un nuevo género a clasificar.

SVM

Para entrenar este modelo, se modifica el tipo de kernel alternando rbf, linear, poly y sigmoid y se varía el hiperparámetro de regularización C. De esta forma se puede ver cual de estas arquitecturas proporciona un mejor resultado, siguiendo la tabla 4.15.

Cuadro 4.15: Resultados del modelo SVM en la aproximación 3

Kernel	**C**	**Precisión (%)**	**Desviación estándar (%)**
rbf	1	77.27	1.46
linear	1	72.82	1.49
poly	1	77.21	1.43
sigmoid	1	2.37	0.77
rbf	**200**	**81.83**	**1.50**
linear	200	73.57	1.75
poly	200	79.64	1.04
sigmoid	200	2.33	0.75

La configuración con la que se consiguieron los mejores resultados en esta aproximación fue con un kernel de tipo rbf (Función de base radial) y un valor de C 200. Es conveniente observar que la segunda opción con mayor precisión sería con un kernel de tipo poly y que además aportaría una desviación típica mejor. Adicionalmente se ha notado una gran disminución de precisión en el caso del tipo de kernel sigmoide (llegando a valores cercanos a 0), que por sus características no realiza una buena modelización del problema.

En este caso la precisión obtenida es del 81.83% con una desviación típica de 1.50%. En esta nueva aproximación, volvemos a obtener unos mejores valores de precisión a medida que se eleva el valor de C, esto se puede explicar debido a que la complejidad del problema ha aumentado de nuevo con la adición de nuevos géneros musicales.

Es importante destacar que esta configuración tendrá que ser ajustada en futuras iteraciones del modelo, tratando de mejorar la precisión para las siguientes aproximaciones.

Árboles de decisión

Para entrenar este modelo solo es necesario modificar el valor de la profundidad máxima desde 1 hasta 6, al igual que en las aproximaciones previas, con la modificación de este hiperparámetro se obtienen los resultados que se pueden ver en la tabla 4.16.

Cuadro 4.16: Resultados del modelo de árboles de decisión en la aproximación 3

maxDepth	Precisión (%)	Desviación estándar (%)
1	46.44	2.70
2	65.91	2.12
3	67.76	2.25
4	70.00	1.72
5	72.24	1.68
6	**75.31**	**2.35**

Entrenando este modelo, se puede observar como con la arquitectura con valor profundidad máxima 6, que es la única con la que los resultados han mejorado, se obtiene una precisión de 75.31% siendo esta la más alta de todas las distintas configuraciones de hiperparámetros, además de una desviación típica de 2.35%.

Observando los distintos resultados se puede ver que la precisión ha disminuido para todos los casos en los que el valor de profundidad máxima va del 1 al 5. Esta bajada en la precisión viene acompañada también de una pequeña disminución del valor de la desviación típica en todos estos casos. Esto es porque al existir más atributos, se necesita una mayor profundidad, y valores pequeños cada vez son menos efectivos, siendo especialmente notable la bajada de la precisión del valor de profundidad 1.

4.3.3 Discusión

Como se puede ver en la matriz de confusión 4.17 realizada en base al mejor fold de la arquitectura [5] de la RNA, que es el mejor modelo de esta aproximación, el género que mejor clasifica es el HipHop ya que tiene el mayor número de segmentos bien clasificados y el menor número de segmentos mal clasificados, generando únicamente confusión con el género Metal. El resto de géneros producen resultados similares.

Observando los resultados de la gráfica 4.3, el modelo de las RNA sigue siendo el que obtiene los mejores resultados y que por lo tanto clasifica mejor las nuevas entradas del problema. El valor obtenido con la mejor arquitectura ([5]) en este modelo, para la precisión es de 86,62%,

Género Real	Género Predicho			
	Classical	HipHop	Jazz	Metal
Classical	109	0	26	5
HipHop	0	120	0	26
Jazz	27	2	95	11
Metal	5	23	6	102

Cuadro 4.17: Matriz de confusión de la RNA [5]

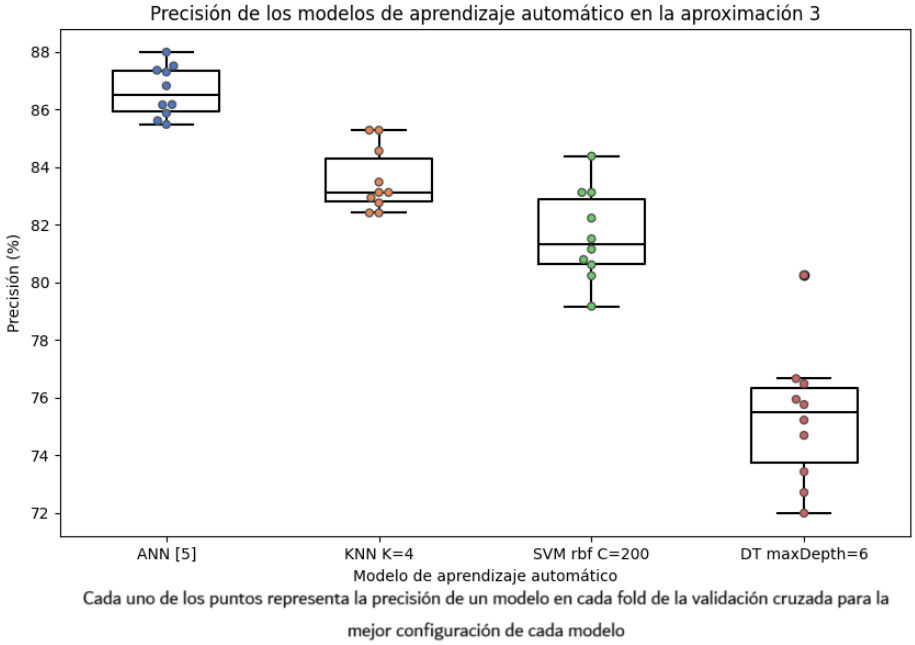
con una desviación típica de 0.88%. En este caso se obtiene un aumento en la precisión que no llega a ser el de la primera aproximación, ya que el problema ahora es más complejo. Sin embargo, con este modelo el valor que se obtiene para la precisión es alto y viene acompañado además, de una reducción en la desviación típica.

En comparación con la aproximación anterior, se puede ver como para todos los modelos, se obtiene una mejora en los resultados respecto a la precisión. Sin embargo, para alguno de estos se puede apreciar como existen arquitecturas que a medida que aumenta la complejidad del problema dejan de ser útiles ya que proporcionan unos resultados demasiado bajos. A pesar de esto, añadir nuevas clases entre las que clasificar sigue siendo un problema, ya que se puede apreciar como la precisión disminuye. Sin embargo, a diferencia de la aproximación anterior para todos los modelos a excepción de las RNAs, esta disminución es mucho menor, aproximadamente en un 10% respecto a la primera aproximación e incluso mostrando una mejora a través del aumento de entre un 5% y 8% en la precisión respecto a la segunda aproximación.

Esta mejora, incluso aumentando la complejidad del problema, viene dada porque en este caso se añade un género muy diferente de los anteriores, el hiphop. En la aproximación anterior, con las métricas disponibles, la tarea de diferenciar entre los géneros jazz y clásico aumentaba excesivamente la complejidad. Esto no sigue ocurriendo en esta aproximación, ya que la mayor cantidad de métricas y sus propiedades, permite realizar una clasificación mejor entre dichos géneros y además, como se comenta anteriormente, el nuevo género tiene unas características muy distintas a los ya escogidos en aproximaciones previas.

Para una futura aproximación se tratará de aumentar el número de géneros de manera que se pueda comprobar si sigue siendo un problema, que disminuye la precisión notablemente. En ese caso, se añadirán nuevas métricas para compensar este posible aumento en la complejidad.

Figura 4.3: Comparación de la precisión de los distintos modelos de aprendizaje automático



4.4 Aproximación IV

4.4.1 Descripción

En esta aproximación, se implementan dos nuevas métricas y se añaden dos géneros más, buscando así mejorar la precisión respecto a la anterior aproximación a la vez que se aumenta la complejidad del problema.

En este caso, la primera métrica es el Spectral centroid que indica la frecuencia promedio ponderada por la energía espectral y será útil, a la hora de distinguir entre géneros con características espectrales diferentes, como el brillo del pop y la densidad de un género como el metal. Se puede observar en la tabla 4.18 como los géneros metal y disco contemplan una media que los distingue del resto de géneros que tienen valores similares.

La otra métrica, es el Spectral flatness, y se calcula haciendo la proporción entre la magnitud armónica y la magnitud promedio del la potencia espectral. Por lo tanto y como se aprecia en la tabla 4.19, es útil para diferenciar entre sonidos que presentan características armónicas que

suelen provenir de instrumentos como el piano presente en géneros como la música clásica y aquellos más ruidosos o inarmónicos, de géneros como el disco o el metal.

Género	Métrica Spectral centroid			
	Media	Desviación Típica	Valor mínimo	Valor máximo
Classical	0.07	0.03	0.01	0.21
Jazz	0.07	0.04	0.01	0.31
Hip-hop	0.10	0.03	0.02	0.27
Metal	0.14	0.03	0.02	0.27
Reggae	0.09	0.04	0.02	0.32
Disco	0.12	0.03	0.03	0.28

Cuadro 4.18: Comparativa entre géneros para la métrica Spectral centroid

Género	Métrica: Spectral flatness			
	Media	Desviación Típica	Valor mínimo	Valor máximo
Classical	1524.76	459.35	538.23	3359.47
Jazz	1929.01	739.77	561.65	5802.80
Hip-hop	2617.58	500.48	951.94	3949.71
Metal	2705.47	385.19	863.37	3880.49
Reggae	2464.85	679.04	1021.96	5443.04
Disco	2722.86	461.53	1511.87	4292.21

Cuadro 4.19: Comparativa entre géneros para la métrica Spectral flatness

Al conseguir mejorar la precisión en la anterior aproximación, a pesar de añadir un nuevo género, para esta se añadirán dos más, el reggae y el disco. Sin embargo, hay que tener en cuenta que estos nuevos géneros, aunque son diferentes, pueden tener características que para alguna de las métricas den valores parecidos a los de los géneros ya existentes, por lo que la precisión puede disminuir de nuevo.

Los modelos de aprendizaje automático serán los mismos que en aproximaciones previas y modificando los hiperparámetros para probar diferentes arquitecturas en cada uno de estos, se comparará si los resultados han mejorado respecto a la tercera aproximación.

Los datos originales de entrada pertenecientes a la base de datos, se modificarán igual que en aproximaciones pasadas y las pistas de 30 segundos tendrán una duración de 2,97 segundos, obtenida gracias a la fórmula 4.1, dividiéndolo en 14 segmentos por audio. Al tratar con 6 géneros, para esta aproximación se dispondrá de 600 pistas divididas en 8400 segmentos.

4.4.2 Resultados

RNA

En esta nueva aproximación se volvieron a probar los mismo hiperparámetros y las misma topologías que probamos en las aproximaciones anteriores, obteniendo los resultados que se detallan en la tabla 4.20.

Cuadro 4.20: Resultados del modelo RNA en la aproximación 4

Topología	Precisión (%)	Desviación estándar (%)
[2, 1]	79.69	1.95
[7, 2]	85.36	1.23
[4, 4]	86.00	0.82
[5, 3]	85.77	1.01
[3]	85.67	0.62
[4, 6]	85.18	0.88
[2, 5]	83.16	0.77
[5]	**86.47**	**0.63**

En esta aproximación podemos observar que la topología con la que se obtuvieron los

mejores resultados consta de una capa oculta de cinco neuronas, la cual también tuvo los mejores resultados en la aproximación anterior. El resultado que obtenemos en este caso es muy similar al de la iteración previa, con un 86.47% de precisión y un 0.63% de desviación estándar.

En todas las topologías probadas, a pesar del aumento de parámetros de entrada, que incluyeron dos métricas adicionales para esta iteración, no se observa una mejora significativa en la precisión en ninguna de las topologías respecto a su homónima en la aproximación anterior. Esto puede ser debido a que en esta aproximación también se añadieron dos géneros a mayores, lo cual aumento la complejidad del problema. Sin embargo, se observó una mejora en la desviación estándar.

kNN

Los valores obtenidos para esta aproximación al introducir al problema el valor de la media y desviación típica para las dos nuevas métricas se puede observar en la tabla 4.21.

Cuadro 4.21: Resultados del modelo kNN en la aproximación 4

K	Precisión (%)	Desviación estándar (%)
4	**78.99**	**0.83**
5	78.68	1.49
6	77.86	1.20
7	77.49	1.57
8	77.20	1.29
9	76.77	1.44
10	76.83	1.19

La arquitectura con la que finalmente se obtuvieron los mejores resultados fue con k=4. Este valor de k proporcionó una precisión del 78.99%, con una desviación estándar de 0.83%.

Con la incorporación de dos nuevos atributos para la clasificación de los audios en los distintos géneros y los dos nuevos géneros entre los que clasificar se obtuvo una disminución de la precisión en torno a un 5% y 6% con respecto a la aproximación anterior para este modelo. Al introducir estos nuevos datos la clasificación se vuelve más compleja y la precisión en la clasificación disminuye.

SVM

Para entrenar este modelo, se modifica el tipo de kernel alternando rbf, linear, poly y sigmoid y se varía el hiperparámetro de regularización C. De esta forma se puede ver cual de estas arquitecturas proporciona un mejor resultado, siguiendo la tabla 4.22.

Cuadro 4.22: Resultados del modelo SVM en la aproximación 4

Kernel	**C**	**Precisión (%)**	**Desviación estándar (%)**
rbf	1	66.96	1.54
linear	1	60.0	1.83
poly	1	66.95	1.18
sigmoid	1	1.30	0.50
rbf	**200**	**75.20**	**1.90**
linear	200	61.98	2.04
poly	200	71.00	1.43
sigmoid	200	3.32	0.55

El modelo resultante para esta aproximación proporciona una precisión del 75.20% como la mayor precisión obtenido con los diferentes hiperparámetros escogidos, con una desviación típica de 1.90%. De nuevo el mejor resultado se obtuvo con un kernel de tipo rbf y un valor de 200 para C. Se puede observar que la siguiente opción más precisa se obtiene con un kernel de tipo rbf y con un valor de 1 para C. Esta precisión es de 66.96% con una desviación típica de 1.54%. Parece más interesante la tercera precisión más alta del modelo que consta de una precisión de 66.95%, un 0.01% inferior a la anterior. Sin embargo, esta misma tiene una desviación típica de un 1.18% por lo que es más interesante esta configuración.Como en anteriores aproximaciones se obtiene una precisión superior con cada tipo de kernel cuando el hiperparámetro C aumenta.

Respecto a la anterior aproximación ha disminuido la precisión para cada una de las diferentes configuraciones para este modelo, obteniendo una precisión un 6.63% inferior para el mejor modelo.

Árboles de decisión

Para entrenar este modelo solo es necesario modificar el valor de la profundidad máxima desde 1 hasta 6, al igual que en las aproximaciones previas, con la modificación de este hiperparámetro se obtienen los resultados que se pueden ver en la tabla 4.23

Cuadro 4.23: Resultados del modelo de árboles de decisión en la aproximación 4

maxDepth	Precisión (%)	Desviación estándar (%)
1	29.79	1.64
2	41.67	1.72
3	47.57	1.67
4	54.44	2.22
5	55.89	2.67
6	**60.07**	**2.00**

Entrenando este modelo, se puede observar como con la arquitectura con valor de profundidad máxima 6 sigue siendo con la que se obtienen mejores resultados. Cabe destacar que en esta iteración se empeorara el rendimiento significativamente respecto a la iteración anterior en todas las configuraciones. Siendo la mejor precisión obtenida de 60.07% teniendo una desviación típica del 2.00 %.

Observando los distintos resultados se puede ver que la precisión ha disminuido para todos los casos. Esta bajada en la precisión viene acompañada también de una pequeña disminución del valor de la desviación típica en todos estos casos. Esto es porque al existir más atributos, se necesita una mayor profundidad, y valores pequeños cada vez son menos efectivos, siendo especialmente notable la bajada de la precisión del valor de profundidad 1.

4.4.3 Discusión

Como se puede ver en la matriz de confusión 4.24 realizada en base al mejor fold de la arquitectura [5] de la RNA, que es el mejor modelo de esta aproximación, el género que mejor se clasifica es el Disco, ya que tiene el mayor número de segmentos bien clasificados y el menor número de segmentos mal clasificados. Este está seguido muy de cerca por el Reggae. También se puede observar que en contraposición a la aproximación anterior ahora el género que peor se clasifica es el HipHop, el cual antes era el que mejor se clasificaba.

En cuanto a los resultados reflejados en la gráfica 4.4, en esta aproximación, la RNA obtuvo los mejores resultados y se adapta de una forma correcta a los nuevos incrementos de nuestro

Género Real	Género Predicho					
	Classical	Disco	HipHop	Jazz	Metal	Reggae
Classical	77	1	11	13	13	6
Disco	1	109	2	0	19	1
HipHop	19	1	63	27	6	19
Jazz	20	1	14	80	5	22
Metal	29	19	12	5	87	1
Reggae	4	1	31	12	1	105

Cuadro 4.24: Matriz de confusión de la RNA [5]

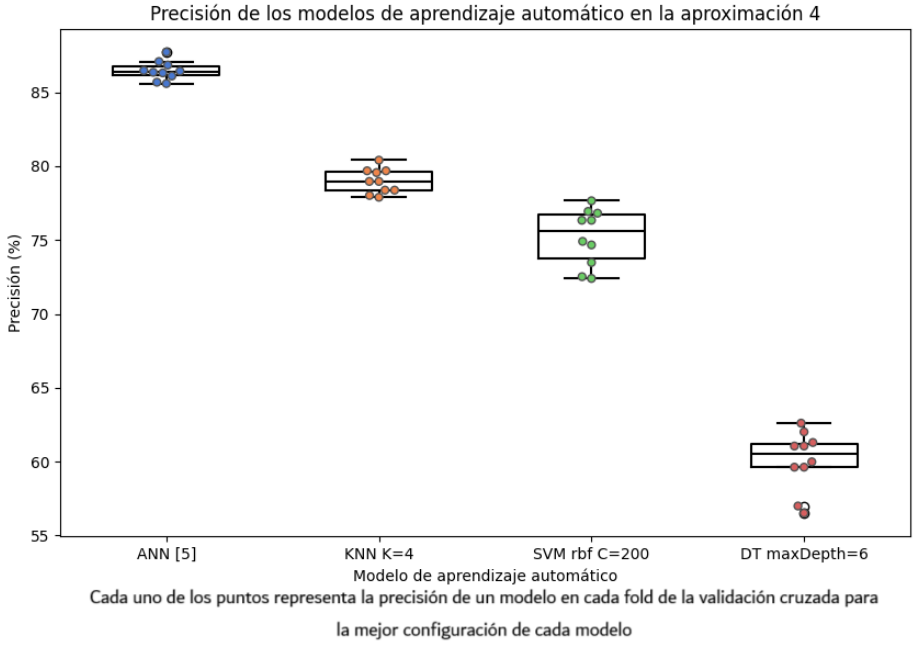
modelo. El valor obtenido para la precisión es de 86.47%, con una desviación típica de 0.63%. Comparando estos valores con los valores del modelo KNN de la anterior aproximación, que es con el que se obtenían los mejores resultados, se puede apreciar como la precisión ha bajado ligeramente, menos de un 1%, mientras que el valor de la desviación típica es mucho menor.

Entre las diferencias con la aproximación anterior, se puede ver como para todos los modelos, a excepción de las RNAs, se produce un decremento respecto a la precisión. Para el resto de modelos por lo tanto, se aprecia una disminución notable para la mayoría de arquitecturas que esta sobre un 5% en la mayoría de casos, como por ejemplo para las mejores arquitecturas de los modelos SVM y kNN.

La razón de estos resultados se debe a la adición de 2 géneros nuevos, disco y reggae en contraposición a la anterior aproximación que solo se añadió uno. Estos dos géneros son lo suficientemente distintos del resto, sin embargo nos encontramos en un punto en el que cada vez es más difícil obtener buenos resultados debido a que contamos con 5 géneros en total. Esto puede provocar casos en los que ciertas pistas de audio de distintos géneros se puedan llegar a parecer excesivamente. Es necesario señalar en adición que los nuevos atributos utilizados en esta aproximación ayudan a mitigar estos problemas.

Para una futura aproximación se seguirá con la línea de trabajo propuesta en la anterior iteración, de aumentar el número de géneros de manera que se pueda comprobar si se sigue apreciando una notable disminución de la precisión. Posteriormente se continuará con la adición de nuevos atributos para contrarrestar este aumento de complejidad. También se tratará de variar las arquitecturas escogidas para algunos modelos en los que se obtienen valores muy bajos para la precisión, y así, poder comprobar si esta aumenta.

Figura 4.4: Comparación de la precisión de los distintos modelos de aprendizaje automático



4.5 Aproximación V

4.5.1 Descripción

Para la quinta aproximación, tratando de mejorar la precisión obtenida en la anterior, se añadirán dos nuevas métricas que ayuden en la clasificación entre los géneros ya existentes en el problema, manteniendo así la complejidad. Además se modificarán algunas de las arquitecturas escogidas para los distintos modelos en previas aproximaciones, ya que al aumentar la complejidad del problema en aproximaciones pasadas, producían resultados por debajo de lo esperado.

La primera de las métricas que se añaden será el Myriad, que se calcula mediante los Weighted Myriad Filters, ponderando la energía espectral. Esta métrica permite capturar detalles en las pistas como la textura sonora y la distribución de frecuencias y por lo tanto, es útil para diferenciar géneros con características espectrales distintas, como la brillantez del hiphop y la densidad característica del metal como se puede observar en la tabla 4.25.

La segunda métrica es la permutation entropy, que muestra la complejidad de una señal temporal mediante su estructura y organización en el dominio temporal. Por ejemplo puede ser util para diferenciar entre la suavidad del jazz y la complejidad rítmica de géneros como el hiphop. Se pueden apreciar sus características en la tabla 4.26.

Género	Métrica Myriad			
	Media	Desviación Típica	Valor mínimo	Valor máximo
Classical	-251184.19	567264.02	-863212.53	2178673.64
Disco	-205872.65	195620.82	-528671.76	2245009.21
Hip-hop	-30768.34	602028.65	-596732.83	2217394.86
Jazz	-120514.53	610456.09	-682452.56	2238273.12
Metal	113627.99	725143.70	-515760.03	2224134.41
Reggae	-118689.00	561490.34	-735832.27	2212852.63

Cuadro 4.25: Comparativa entre géneros para la métrica Myriad

Género	Métrica Permutation Entropy			
	Media	Desviación Típica	Valor mínimo	Valor máximo
Classical	0.36	0.12	0.16	0.86
Disco	0.74	0.11	0.35	0.98
Hip-hop	0.76	0.11	0.24	0.98
Jazz	0.49	0.17	0.17	0.98
Metal	0.73	0.07	0.33	0.94
Reggae	0.70	0.13	0.27	0.99

Cuadro 4.26: Comparativa entre géneros para la Permutation Entropy

Debido a la disminución de la precisión, se espera que añadiendo estas dos nuevas métricas y sin incluir nuevos géneros entre los que clasificar, los resultados de los diferentes modelos muestren un aumento en la precisión. Al mantener el número de géneros entre los que clasificar, el problema sigue siendo de clasificación multiclase y al igual que en en aproximaciones previas se usarán los mismos modelos de aprendizaje automático. Sin embargo, se modificarán algunas de las arquitecturas escogidas en previas aproximaciones para modelos como los árboles de decisión, tratando de aumentar su complejidad y así conseguir mejores

resultados de precisión.

Los datos originales de entrada pertenecientes a la base de datos, se modificarán igual que para las anteriores aproximaciones, por lo que las pistas de 30 segundos tendrán una duración de 2,97 segundos, obtenida gracias a la fórmula 4.1, dividiéndolo en 14 segmentos por audio. Al seguir teniendo 6 géneros diferentes, se dispondrá por lo tanto de 600 pistas, divididas en 8400 segmentos.

4.5.2 Resultados

RNA

En esta nueva aproximación se volvieron a probar los mismo hiperparámetros, pero cambiando las topologías que probamos en las aproximaciones anteriores, obteniendo los resultados que se detallan en la tabla 4.27.

Cuadro 4.27: Resultados del modelo RNA en la aproximación 5

Topología	Precisión (%)	Desviación estándar (%)
[4, 4]	86.9	0.76
[4, 6]	87.10	0.79
[2, 8]	85.26	1.16
[4, 8]	87.52	0.59
[2]	85.83	0.79
[8]	88.34	0.49
[6, 6]	88.34	0.49
[8,8]	**89.18**	**0.4**

Tras haber introducido 2 nuevos atributos de los fragmentos de audio para caracterizar estos mismos y poder clasificar los audios entre los géneros escogidos, la precisión para la RNA ha aumentado respecto a la aproximación 4. Esto se explica debido a que la adición de nuevos atributos permite abordar de mejor forma el problema lo cual se percibe como un notable aumento de la precisión en todas las topologías.

La mayor precisión obtenida ha sido de 89.18% con una desviación estándar de un 0.4% para una topología con dos capas ocultas. Esta topología que proporciona el mejor resultado de

clasificación consta de 8 neuronas en la primera capa oculta y de 8 neuronas en la segunda capa oculta.

kNN

Los valores obtenidos para esta aproximación al introducir al problema el valor de la media y desviación típica para las dos nuevas métricas se puede observar en la tabla 4.28.

Cuadro 4.28: Resultados del modelo kNN en la aproximación 5

K	Precisión (%)	Desviación estándar (%)
1	**85.94**	**0.81**
4	81.86	1.45
5	82.13	1.45
6	81.64	1.42
7	81.08	1.55
8	80.36	1.60
9	79.88	1.52
10	79.27	1.39

La arquitectura con la que finalmente se obtuvieron los mejores resultados fue con k=1. Este valor de k proporcionó una precisión del 85.94%, con una desviación estándar de 0.81%.

Con la incorporación de dos nuevos atributos para la clasificación de los audios en los distintos géneros y el no añadir nuevos géneros se obtuvo un incremento de la precisión en torno a un 6% y 7% con respecto a la aproximación anterior para este modelo. Al introducir estos nuevos atributos se produce una mitigación de la complejidad del sistema.

SVM

Para entrenar este modelo, se modifica el tipo de kernel alternando rbf, linear, poly y sigmoid y se varía el hiperparámetro de regularización C. De esta forma se puede ver cual de estas arquitecturas proporciona un mejor resultado, siguiendo la tabla 4.29.

Cuadro 4.29: Resultados del modelo SVM en la aproximación 5

Kernel	**C**	**Precisión (%)**	**Desviación estándar (%)**
rbf	1	71.40	1.42
linear	1	63.45	1.43
poly	1	72.96	1.57
sigmoid	1	1.41	0.59
rbf	**200**	**82.49**	**1.46**
linear	200	65.83	1.62
poly	200	78.39	1.45
sigmoid	200	2.79	0.62

El modelo resultante para esta aproximación proporciona una precisión del 82.49% como la mayor precisión obtenido con los diferentes hiperparámetros escogidos, con una desviación típica de 1.46%. De nuevo el mejor resultado se obtuvo con un kernel de tipo rbf y un valor de 200 para C.Como en anteriores aproximaciones se obtiene una precisión superior con cada tipo de kernel cuando el hiperparámetro C aumenta.

La mayor parte de los resultados de esta aproximación con respecto a las máquinas de soporte vectorial han mejorado considerablemente para cada una de los topologías. Este aumento de la precisión se explica por la adición de nuevos atributos y que en esta iteración no se han añadido nuevos géneros que aumentaran la complejidad del problema.

Árboles de decisión

Para entrenar este modelo solo es necesario modificar el valor de la profundidad máxima desde 6 hasta 50, con la modificación de este hiperparámetro se obtienen los resultados que se pueden ver en la tabla 4.30

Entrenando este modelo podemos ver una clara mejora respecto a la iteración anterior. Esto puede verse reflejado en el caso en el que la profundidad máxima es 6, el cuál fue el que mejor se comporto en las iteraciones anteriores y aun así mejoro significativamente su rendimiento, lo que nos lleva a concluir que el añadir nuevas métricas a mayores conllevó una mejora en los resultado. Cabe destacar, que exceptuando la profundidad máxima de 6, el resto se cambiaron respecto a la iteración anterior, ya que al observar que siempre obtenían

Cuadro 4.30: Resultados del modelo de árboles de decisión en la aproximación 5

maxDepth	**Precisión (%)**	**Desviación estándar (%)**
6	62.78	1.22
10	71.19	1.34
20	**74.85**	**1.44**
30	74.80	1.00
40	74.80	1.00
50	74.80	1.00

mejores resultados con el árbol de mayor profundidad decidimos incrementar este valor para ver que tal se comportaba.

En este punto del proyecto, la profundidad máxima que mejores valores nos proporcionó fue 20, que obtuvo un 74.85% de precisión y un 1.44% de desviación estándar. Como dato interesante se puede observar que al usar profundidades máximas superiores a este punto se empeora el resultado levemente, dando igual la profundidad máxima utilizada ya que nos dan el mismo resultado tanto en precisión como desviación estándar.

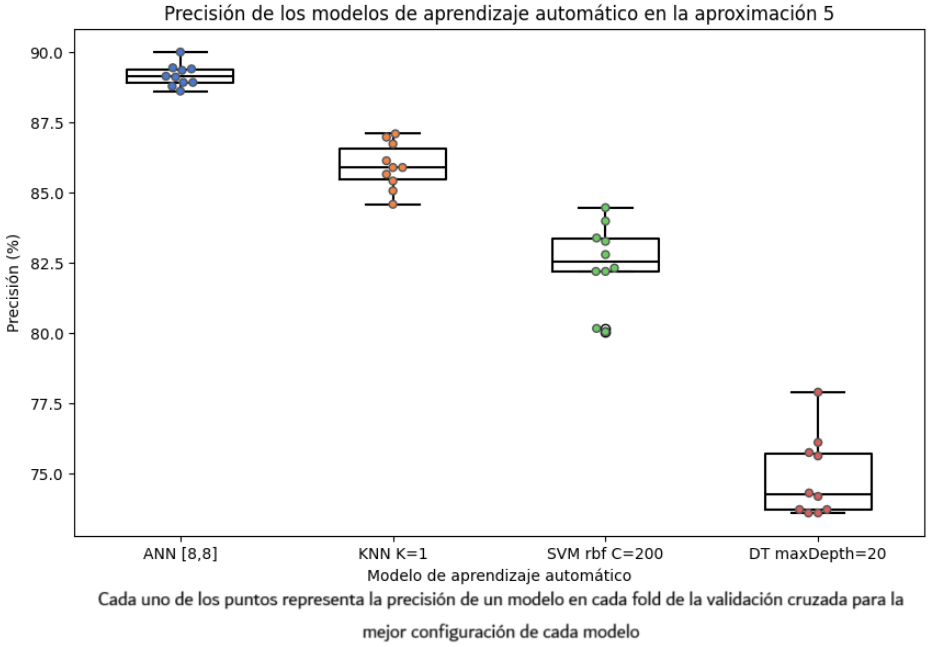
4.5.3 Discusión

La matriz de confusión representada en la tabla 4.31 se realiza en base al mejor fold del mejor modelo de esta aproximación. En esta aproximación se puede observar que ya existen géneros que nunca se confunden entre ellos y se obtienen valores muy altos de casos en los que el sistema clasifica bien, como es el caso de Disco y Reggae.

Género Real	Género Predicho					
	Classical	Disco	HipHop	Jazz	Metal	Reggae
Classical	81	1	17	17	11	1
Disco	1	109	1	1	17	0
HipHop	17	0	84	35	5	6
Jazz	22	0	9	72	2	10
Metal	26	22	3	2	94	4
Reggae	3	0	19	10	2	133

Cuadro 4.31: Matriz de confusión de RNA con [8,8] de topología

Figura 4.5: Comparación de la precisión de los distintos modelos de aprendizaje automático



Como se puede observar en la gráfica 4.5, para esta aproximación en la que se introdujeron dos métricas nuevas, el modelo de RNA, con una topología de [8,8] consiguió el mejor resultado, con una precisión del 89.18% y una desviación típica del 0.4%. Respecto al mejor resultado de la anterior aproximación se mejora la precisión en aproximadamente un 4% y disminuyendo ligeramente la desviación típica.

En general, respecto a la aproximación anterior, se pueden apreciar varias diferencias en cada uno de los modelos, produciéndose aumentos y disminuciones en función de las distintas arquitecturas. Por ejemplo, para el modelo de árboles de decisión se puede ver un aumento de en torno al 15% dependiendo de la arquitectura, ya que al aumentar la complejidad del modelo incrementando la profundidad máxima, se han conseguido mejores resultados. Para kNN en este caso, el valor de k = 1 produce el mejor resultado, y el resto aumenta ligeramente su precisión.Con las RNAs, al probar nuevas arquitecturas se obtienen valores de precisión mayores a la anterior aproximación y finalmente, en los SVM los kernel de tipo sigmoide, independientemente del valor de C siguen produciendo resultados muy bajos para la precisión, mientras que el resto de kernels y valores de C aumentan la precisión en mayor o menor porcentaje en función del kernel, mejorando la arquitectura que producía la mayor precisión en la anterior

aproximación en un 7% aproximadamente.

La razón de estos resultados se debe en mayor medida a la introducción de dos nuevas métricas, ya que estas nuevas características ayudan a diferenciar algunas de las pistas que podían producir confusión. Además al no aumentar la complejidad del problema en esta aproximación, dejando el mismo número de géneros entre los que es posible etiquetar cada pista, se consigue que por lo general los resultados respecto a la precisión sean mejores. Otra de las medidas tomadas para esta aproximación es la modificación de alguno de los hiperparámetros que producían resultados demasiado bajos en algunos modelos como los árboles, en los que se puede apreciar como mediante este cambio se ha conseguido aumentar notablemente la precisión de los resultados.

En la siguiente aproximación se aumentará el número de géneros entre los que clasificar evitando que la precisión disminuya notablemente. Esto puede llegar a ser problemático ya que los géneros restantes tienen características muy similares para las distintas pistas pertenecientes a la base de datos, por lo que los resultados podrán no ser satisfactorios. Otro objetivo será revisar las arquitecturas que producen mejores resultados para mantenerlas, y a la vez, se tratará de remplazar aquellas que den lugar a precisiones muy bajas.

4.6 Aproximacion VI

4.6.1 Descripción

En esta aproximación y a raíz de los resultados obtenidos en la aproximación anterior, se aumentará la complejidad del problema, tratando de no disminuir la precisión. Además se probarán los distintos modelos ya modificados en previas aproximaciones con los que se obtenían unos mejores resultados, para comprobar si consiguen este objetivo añadiendo un nuevo género al problema entre el que clasificar

El género que se decide añadir es el Rock, ya que presenta características similares al Metal, que era uno de los géneros junto con la música clásica, con los que se obtenían los valores más diferenciados para las métricas estudiadas en este sistema. Con esto, se busca observar los resultados producidos por los diferentes modelos a falta de un género muy diferenciado del resto. Como se puede observar en la tabla 4.32, para la métrica del spectral centroid se obtienen una media muy similar para todos los géneros presentes en el problema.

Al igual que en en aproximaciones previas se usarán los mismos modelos de aprendizaje

Género	Métrica spectral centroid			
	Media	Desviación Típica	Valor mínimo	Valor máximo
classical	0.08	0.03	0.01	0.21
disco	0.13	0.04	0.04	0.28
hiphop	0.11	0.04	0.02	0.27
jazz	0.08	0.04	0.02	0.31
metal	0.15	0.04	0.02	0.28
reggae	0.09	0.05	0.02	0.32
rock	0.11	0.04	0.02	0.22

Cuadro 4.32: Segunda comparativa entre géneros para la métrica spectral centroid

automático, con las mismas arquitecturas ya modificadas en la aproximación anterior, con las que se obtuvieron mejores resultados, a la espera de obtener una precisión similar.

De la misma forma, las pistas de la base de datos se dividirán en segmentos de 2,97 segundos. En este caso, teniendo 7 géneros diferentes, para esta aproximación se contará con 700 pistas divididas en 9800 segmentos.

4.6.2 Resultados

RNA

En esta nueva aproximación se volvieron a probar los mismos hiperparámetros y la misma topología que probamos en la aproximaciones anterior, obteniendo los resultados que se detallan en la tabla 4.33.

En este caso se obtuvo un resultado inesperado ya que a pesar de aumentar la complejidad del problema añadiendo un género nuevo, la precisión para todas las arquitecturas de las RNA ha aumentado respecto a la aproximación 5.

La mayor precisión obtenida, como se aprecia en la tabla 4.33, ha sido de 88.71% con una desviación estándar de un 0.56% para una topología [6,6].

En el resto de topologías, se puede apreciar como se producen resultados muy similares apenas mejorando los resultados de la aproximación pasada.

Cuadro 4.33: Resultados del modelo RNA en la aproximación 6

Topología	Precisión (%)	Desviación estándar (%)
[4, 4]	86.42	0.59
[4, 6]	86.77	0.59
[2, 8]	85.17	0.97
[4, 8]	87.08	0.7
[2]	85.67	0.57
[8]	87.84	0.53
[6, 6]	**88.71**	**0.56**
[8,8]	87.85	0.49

kNN

Los valores obtenidos para esta aproximación al introducir al problema el nuevo género se pueden apreciar en la tabla 4.34.

En este caso la arquitectura que obtiene mejores resultados se produjo con K=1, con una

Cuadro 4.34: Resultados del modelo kNN en la aproximación 5

K	Precisión (%)	Desviación estándar (%)
1	**81.05**	**1.50**
4	76.75	1.59
5	76.62	1.85
6	76.02	1.61
7	75.71	1.81
8	74.87	1.77
9	74.39	1.75
10	73.61	1.78

precisión del 81.05% y una desviación estándar de un 1.50%

Como era esperado, la precisión disminuye debido a la adición de un nuevo género y la ausencia de nuevas características, sin embargo, esta disminución no es tan notable debido a la

correcta generalización del modelo

SVM

Para entrenar este modelo, se modifica el tipo de kernel alternando rbf, linear, poly y sigmoid y se varía el hiperparámetro de regularización C. De esta forma se puede ver cual de estas arquitecturas proporciona un mejor resultado, siguiendo la tabla 4.35.

Cuadro 4.35: Resultados del modelo SVM en la aproximación 6

Kernel	**C**	**Precisión (%)**	**Desviación estándar (%)**
rbf	1	64.77	2.09
linear	1	55.45	2.18
poly	1	66.56	2.02
sigmoid	1	1.11	0.44
rbf	**200**	**76.82**	**2.03**
linear	200	58.43	1.83
poly	200	72.49	1.73
sigmoid	200	2.40	0.53

El modelo resultante para esta aproximación proporciona una precisión del 76.82% como la mayor precisión obtenido con los diferentes hiperparámetros escogidos, con una desviación típica de 2.03%. De nuevo el mejor resultado se obtuvo con un kernel de tipo rbf y un valor de 200 para C.Como en anteriores aproximaciones se obtiene una precisión superior con cada tipo de kernel cuando el hiperparámetro C aumenta.

La mayor parte de los resultados de esta aproximación con respecto a las máquinas de soporte vectorial han empeorado considerablemente para cada una de los topologías. Esta disminución de la precisión se explica por la adición de un nuevo género lo cual aumentó la complejidad del problema.

Árboles de decisión

Para entrenar este modelo solo es necesario modificar el valor de la profundidad máxima desde 6 hasta 50, con la modificación de este hiperparámetro se obtienen los resultados que se pueden ver en la tabla 4.36

Cuadro 4.36: Resultados del modelo de árboles de decisión en la aproximación 6

maxDepth	Precisión (%)	Desviación estándar (%)
6	55.46	2.74
10	65.43	2.15
20	**69.12**	**1.53**
30	68.78	1.21
40	68.78	1.21
50	68.78	1.21

Al haber introducido un género más en esta aproximación, los resultados mostrados en la tabla 4.36 muestran como se ha reducido la precisión obtenida para este modelo de árboles de decisión. Igual que anteriormente, a partir de 30 para el valor de profundidad, se obtiene una precisión y desviación estándar monótona.

Las desviaciones estándar para 6 y 10 en valor de profundidad son considerablemente superiores al resto de desviaciones estándar para los demás valores y se puede observar en la tabla 4.36 que la desviación estándar está entre los dos extremos de valores para la desviación pero esta se acerca más en torno a la desviación estándar obtenida para valores de profundidad superiores.

4.6.3 Discusión

Una vez más la matriz de confusión, la cual se puede ver en la tabla 4.37, muestra el mejor fold de la mejor arquitectura de las RNAs. Como se puede apreciar en estas los géneros que presentan menos problemas para diferenciarse del resto es el Disco. Mientras que el resto de géneros dan un resultado similar en el que la mayoría de los segmentos se clasifican correctamente y los que no se reparten entre el resto de géneros en una proporción similar. El género que presenta el peor resultado es el HipHop ya que el número de segmentos bien clasificados es menor que el número de segmentos mal clasificados.

Como se puede observar en la gráfica 4.6, en esta aproximación en la que se cuenta con un género más entre el que clasificar, son las RNAs las que consiguen un mejor resultado.

Género Real	Género Predicho						
	Classical	Disco	HipHop	Jazz	Metal	Reggae	Rock
Classical	58	2	10	5	5	10	10
Disco	0	132	3	0	13	0	1
HipHop	22	1	69	22	10	5	15
Jazz	10	0	17	78	2	10	21
Metal	15	8	4	7	87	0	23
Reggae	13	2	19	13	3	117	0
Rock	19	0	17	12	11	0	76

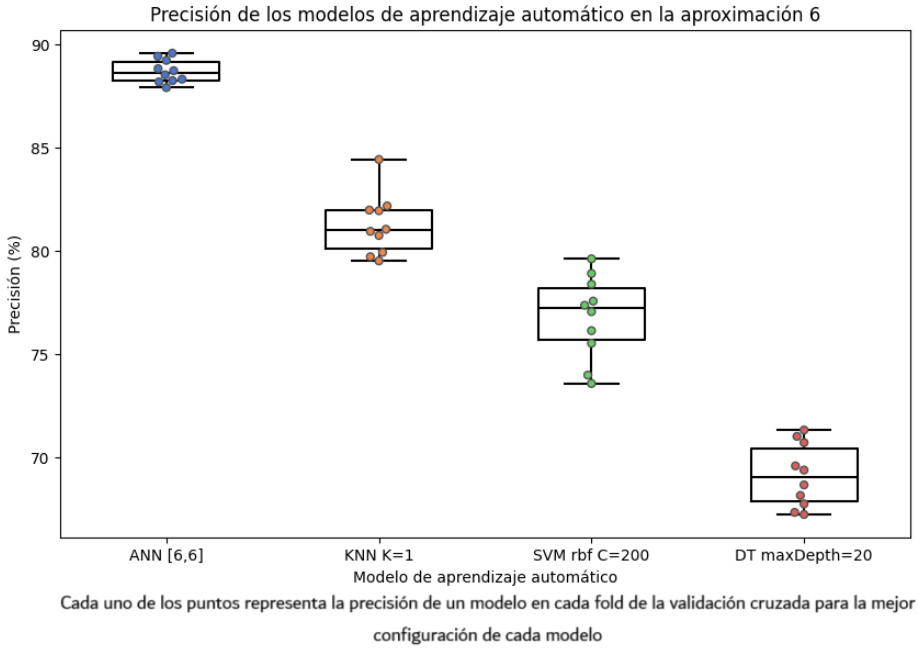
Cuadro 4.37: Matriz de confusión de la RNA [6,6]

Concretamente, aquella cuya arquitectura es [6,6], con una precisión del 88.71% y una desviación típica del 0.56%. Al igual que en la mayoría de aproximaciones anteriores, el modelo de RNA es el que produce el mejor resultado. En este caso además, apenas se reduce la precisión respecto al mejor modelo de la aproximación pasada ya que este descenso se sitúa en torno al 0.4%, proporcionando un buen resultado.

Para el resto de modelos se aprecia como se produce una disminución en el valor de la precisión en la mayoría de las configuraciones de hiperparámetros y arquitecturas. Esto se debe a que se añade el género rock a los géneros entre los que se pueden etiquetar, ya que al tener cada vez más categorías que se parecen entre si, las distintas métricas proporcionan resultados similares y la tarea de clasificación se vuelve más compleja

En la siguiente aproximación se cambiará el procedimiento respecto a todas las aproximaciones hasta ahora y se pasará a desarrollar una aproximación basada en deep learning. Este enfoque se presenta como una alternativa especialmente útil para problemas como el que se esta tratando, que consisten en el reconocimiento de patrones en señales de audio.

Figura 4.6: Comparación de la precisión de los distintos modelos de aprendizaje automático



4.7 Aproximacion VII

4.7.1 Descripción

Para esta aproximación ya no se usarán los modelos vistos en las anteriores aproximaciones, en su lugar esta aproximación estará basada en Deep Learning, usando redes convolucionales, que son un tipo de arquitectura de redes de neuronas artificiales profundas, compuestas por varias capas de nodos interconectados que procesan los datos de entrada en varias etapas, dando lugar a la identificación de características más complejas y que son especialmente útiles para obtener un procesamiento eficiente de datos como imágenes o señales.

Como las bases de datos usadas para este tipo de problemas suelen ser muy grandes, los datos se dividen en subconjuntos denominados batches. En este problema en específico al no disponer de una base de datos tan grande solo se usará un único batch. También hay que tener en cuenta que lo normal sería hacer validación cruzada, pero este proceso llevaría demasiado tiempo, y por tanto no se realizará durante esta aproximación, en su lugar se utilizará validación hold-out.

Por lo tanto, para esta aproximación se usarán los propios segmentos de audio provenientes de la base de datos, en concreto habrá 7 géneros diferentes, los mismos que en la aproximación pasada, de 4 canciones divididas en 14 segmentos cada una.

4.7.2 Resultados

Tamaño del filtro de convolución	Función de Transferencia	Precisión de Test (%)	Ciclo
3	tanh	10.96	23
3	ReLU	20.55	16
3	Sigmoid	13.69	1
4	**ReLU**	**23.29**	**12**
5	tanh	19.18	8
5	ReLU	21.92	24
5	Sigmoid	6.85	1
6	tanh	20.55	31
6	ReLU	9.59	9
6	Sigmoid	13.70	1

Cuadro 4.38: Comparación de las arquitecturas de deep learning

Se utiliza una red neuronal con tres capas de convolución, cada una seguida de una capa de max pooling para reducir progresivamente las dimensiones de la entrada. Después de las operaciones de convolución y max pooling, se aplica una capa densa para mapear las características extraídas a las clases de salida. Finalmente, se utiliza una función softmax para producir la distribución de probabilidad sobre las clases de salida.

Como se puede apreciar en la tabla 4.38 se fueron variando son tanto el tamaño del filtro de convolución, como la función de transferencia que utilizan. Y para cada una de estas arquitecturas se apuntó el valor de la precisión de test para el ciclo indicado, que es el último en el que mejoró la precisión de entrenamiento.

Se evaluaron filtros de tamaño 3, 4, 5 y 6 que hacen que varíe el rendimiento del modelo, siendo 4 el tamaño que da la mejor precisión. Por otro lado, las diferentes funciones de transferencia utilizadas fueron tanh, ReLU y Sigmoid, siendo ReLU la que proporciona un mejor resultado respecto a la precisión en la mayoría de los casos.

4.7.3 Discusión

Los resultados de esta aproximación fueron en general muy bajos para todas las arquitecturas que probamos. El mejor de ellos se obtuvo con una función de transferencia ReLU para un tamaño de filtro de convolución de 4, en la que en el ciclo 12 se consigue una precisión de test de 23.29% Además, debido a que los resultados tardaban demasiado en ejecutarse tuvo que reducirse el número de segmentos respecto a aproximaciones pasadas.

Comparado con aproximaciones pasadas los resultados para la precisión son mucho menores, disminuyendo entre un 60%-70% para la mejor arquitectura entre todos los modelos. Debido a esto creemos que nuestro problema no es adecuado para tratar con técnicas de deep learning. Este descenso en precisión se debe principalmente a la necesidad de disminuir las entradas, ya que, como se mencionó en la descripción de esta aproximación, si se añaden más entradas el proceso de deep learning llevaría demasiado tiempo.

Relacionado con estas dificultades, para una futura aproximación debería explorarse maneras de procesar los datos de entrada para tratar de reducir la complejidad del problema y así obtener mejores resultados.

Capítulo 5

Conclusiones

Finalmente y una vez desarrolladas todas las aproximaciones que se planearon inicialmente, se pueden comparar todos los resultados obtenidos, concluyendo que el sistema clasifica correctamente en la mayoría de los casos, independientemente del modelo de aprendizaje automático elegido. A pesar de esto, existen arquitecturas para alguno de los modelos que como se ha comentado en las aproximaciones correspondientes, no son útiles para este problema ya que no proporcionan buenos resultados. Sin embargo, a excepción de estas, se han obtenido resultados en el rango de 70%-80% de precisión para todos los modelos a excepción de las RNA cuyos resultados varían en torno al rango de 80%-90%. Por lo tanto, el sistema cumple el objetivo inicial que consistía en distinguir géneros musicales a través de audio, indicado en la descripción del problema.

A la hora de decidir el mejor modelo, es importante tener en cuenta que a medida que se avanzaba en las aproximaciones también aumentaba la complejidad, por lo que los mejores resultados para la precisión fueron obtenidos en la primera aproximación en la que la complejidad del problema era muy baja, ya que solo se clasificaba entre dos géneros con características muy diferentes. Entre el resto de aproximaciones los mejores resultados se obtuvieron para la aproximación 5, ya que el plan de trabajo buscaba mejorar el resultado en cada una de las aproximaciones, aumentando a su vez la complejidad del problema. En esta aproximación no se añadió ningún nuevo género por lo que partiendo de la anterior, se mejoraron los resultados en todos los modelos, obteniendo la precisión mas alta para la clasificación entre 6 géneros. Entre los distintos modelos, no se puede concluir cual es el mejor. Sin embargo, los resultados del modelo de RNA muestran unos valores promedio superiores al resto en 5 de las 6 aproximaciones (sin tener en cuenta la de deep learning). Para la aproximación 1 los mejores resultados se obtuvieron con la Knn, esto contrasta con el resto de aproximaciones en las que por alguna razón con la arquitectura que se obtuvo mejores resultados fue con la ANN. En adición este valor de precisión de Knn fue la mejor precisión obtenida en todas las

aproximaciones, esto debido a la poca complejidad con la que contaba el problema en estos estadíos del proyecto, lo cual hace más extraño los malos resultados obtenidos para la ANN.

Como dificultades encontradas a lo largo del trabajo, hay que destacar alguna como la búsqueda de características que ayudasen a la hora de clasificar las pistas entre los distintos géneros, ya que sin un amplio conocimiento sobre señales de audio, entender como estos parámetros pueden mostrar diferentes resultados en función de las características del audio para cada género, puede llegar a ser complejo. Además, entre los distintos modelos el de las RNA resultó ser el más costoso de ajustar correctamente. Esto provocó que muchos de los resultados fueran inesperados en cada una de las aproximaciones. haciendo que tuviéramos que revisar su implementación, mientras que el resto de modelos se comportaban de manera similar a medida que aumentaba la complejidad del problema o se añadían nuevas métricas, con el objetivo de mejorar los resultados de la clasificación.

Capítulo 6

Trabajo futuro

Relacionadas con esta temática de trabajo, en un futuro se podría tratar de explorar nuevas líneas como, por ejemplo, el diagnóstico médico con diferentes tipos de señales o el reconocimiento de partes de audio concretas como la voz en entornos ruidosos ya que son alternativas centradas en el procesamiento de señales de audio.

En el mundo real, este sistema podría implementarse en algunas de las plataformas de reproducción de música en streaming para recomendar canciones a los usuarios según sus preferencias musicales, como pueden ser Spotify o Apple music. Además profundizar en la aproximación relacionada con deep learning, mediante el uso de redes neuronales convolucionales para procesar las pistas o analizar también la letra podría mejorar notablemente la precisión del sistema.

Dentro de estas líneas, se puede tratar alguna de mayor complejidad, por ejemplo usando este sistema para identificar emociones en pistas de audio, lo que sería de gran interés para las plataformas mencionadas anteriormente ya que podrían basar sus recomendaciones en función del estado de ánimo del usuario. Otro ejemplo interesante sería el reconocimiento de grupos de música o canciones de una forma eficiente que tendría un impacto enorme en la sociedad, que cada día consume más música.

En un futuro se podrían usar nuevos sistemas de aprendizaje automático, como por ejemplo modelos basados en Transformers, los cuales están demostrando muy buenos resultados a la hora de procesar el lenguaje natural, lo cual es interesante para la línea de trabajo mencionada anteriormente de analizar la letra de una canción.

Bibliografía

Aguilar Sánchez, L. L. (2023). *Análisis comparativo de las técnicas de deep learning perceptrón multicapa y redes neuronales convolucionales aplicadas a la clasificación musical* (Tesis Doctoral no publicada). Universidad de San Carlos de Guatemala.

Clasificar fácilmente una cancion según su estilo musical. (s.f.). *bambinos*. Descargado de https://bambinosweb.es/como-saber-el-genero-de-una-cancion/?expand_article=1

Dutt, N. (2022). Music genre classification using cnn: Part 1- feature extraction. *Medium*. Descargado de https://medium.com/mlearning-ai/music-genre-classification-using-cnn-part-1-feature-extraction-b417547b8981

Ghosh, A. (2020). Uso del chroma_stft para reconocer el estilo musical. *Research Gate*. Descargado de https://www.researchgate.net/figure/Mel-Spectrogram-3-Chroma-STFT-The-Chroma-value-of-an-audio-basically-represent-the_fig4_346659500

Guo, L., Gu, Z., y Liu, T. (s.f.). *Music genre classification via machine learning*. CS.

Liu, K., DeMori, J., y Abayomi, K. (2022). Open set recognition for music genre classification. *arXiv preprint arXiv:2209.07548*.

Lucas Rodríguez, M. (2021). Detección automática de géneros musicales.

Luis Micó, J. (2018). Un programa detecta géneros musicales. *La Vanguardia*. Descargado de https://www.lavanguardia.com/tecnologia/20180601/443965097833/programa-detecta-generos-musicales-letras-canciones.html

Pico Lara, A. (2020). Clasificacion de la música a través de señales. *Medium*. Descargado de https://medium.com/clasificaci%C3%B3n-de-m%C3%BAsica-a-trav%C3%A9s-del-an%C3%A1lisis-de/clasificaci%C3%B3n-de-m%C3%BAsica-a-trav%C3%A9s-del-an%C3%A1lisis-de-se%C3%B1ales-de-audio-1da23481b47c

Pimenta-Zanon, M. H., Bressan, G. M., y Lopes, F. M. (2021). Complex network-based ap-

proach for feature extraction and classification of musical genres. *arXiv preprint arXiv:2110.04654*.

Que es el mfcc y sus aplicaciones. (2022). *wikipedia*. Descargado de https://es.wikipedia.org/wiki/MFCC

Que es el rms, como se calcula y para que se utiliza. (2024). *Polaridad*. Descargado de https://polaridad.es/que-es-rms-descubre-el-valor-eficaz-de-tus-senales/?expand_article=1

Que es el tempo y porque sirve para reconocer el estilo musical. (s.f.). *Escribir canciones*. Descargado de https://www.escribircanciones.com.ar/icomo-componer-musica/217-ique-es-el-tempo-bpm-y-como-afecta-la-musica.html

Silla, C. N., Koerich, A. L., y Kaestner, C. A. (2008). A machine learning approach to automatic music genre classification. *Journal of the Brazilian Computer Society*, *14*, 7–18.

Trafton, A. (2018). Machine-learning system processes sounds like humans do. *MIT News*. Descargado de https://news.mit.edu/2018/machine-learning-system-processes-sounds-humans-do-0419

Training data for music ml models. (2023). *Shaip*. Descargado de https://es.shaip.com/blog/training-data-for-music-ml-models/

Uso del spectral_centroid para reconocer el estilo musical. (2023). *wikipedia*. Descargado de https://en.wikipedia.org/wiki/Spectral_centroid

Vulpe, A. (2020). *Gtzan dataset - music genre classification*. Descargado de https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification

Work w/ audio data: Visualise, classify, recommend. (s.f.). *kaggle*. Descargado de https://www.kaggle.com/code/andradaolteanu/work-w-audio-data-visualise-classify-recommend

"Work w/ Audio Data: Visualise, Classify, Recommend" (s.f.) Pico Lara (2020) "Training data for music ML models" (2023) "Que es el tempo y porque sirve para reconocer el estilo musical" (s.f.) "Uso del spectral_centroid para reconocer el estilo musical" (2023) "Que es el RMS, como se calcula y para que se utiliza" (2024) Luis Micó (2018) "Que es el MFCC y sus aplicaciones" (2022) Vulpe (2020) "Work w/ Audio Data: Visualise, Classify, Recommend" (s.f.) Ghosh (2020) "Clasificar fácilmente una cancion según su estilo musical" (s.f.)