Utiliza «Data Science» para encontrar tu próxima canción preferida

[iStock]/Getty Images.

¿Cómo es que aplicaciones como *Spotify, Deezer, Youtube* o *Apple Music* saben tanto lo que nos gusta? Además de recomendar contenido de personas con gustos similares a los nuestros, ¿qué otras técnicas utilizan estas plataformas para recomendar música?
Estas y otras preguntas las exploro en mi trabajo de Fin de Máster en «Data Science» y «Big Data» de Afi Escuela de Finanzas.

Ludwig Gerardo Rubio Jaime @LudwiGerardo | *Machine Learning Engineer en Omedena*

La aplicación de algoritmos de machine learning en la industria musical ha tenido un crecimiento considerable durante los últimos años, siendo integrados a distintos procesos de la industria: composición, mezcla, producción, remasterización, venta, recomendación, etc.

Las plataformas de *music streaming* son quizá el más claro ejemplo de la transformación de la industria. Ya no se trata solo de comprar música, sino de escucharla, compartirla y tener siempre disponibles nuevas opciones para escuchar.

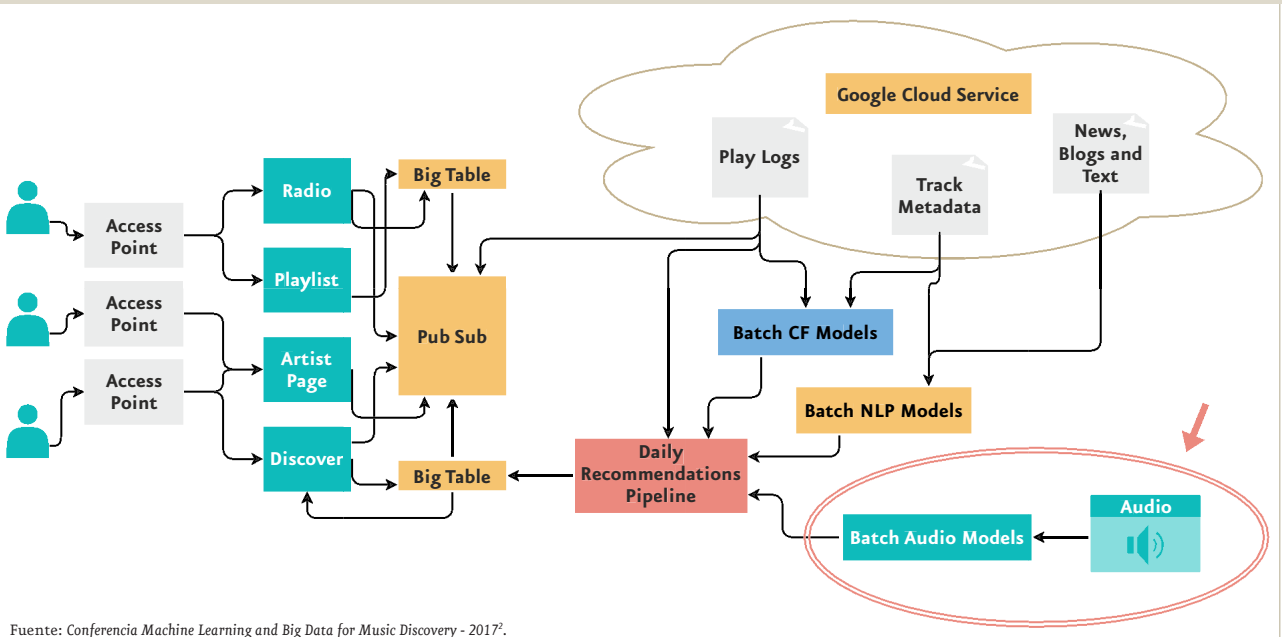
Los sistemas de recomendación basados en filtros colaborativos[1] son utilizados por distintas industrias con presencia digital debido a su comprobado éxito. Sin embargo, empresas como *Spotify* y *Youtube* integran en sus sistemas de recomendación modelos basados en contenido como los *Batch Audio Models* para el análisis de archivos de audio, utilizados en la búsqueda de patrones musicales que puedan mejorar las recomendaciones.

Análisis de los datos

Para explorar el uso de algoritmos de machine y deep learning en el reconocimiento de patrones musicales, hicimos uso del set de datos FMA[3] añadiendo un set de datos de música propia, generando un set de datos final de **104,343 archivos de música**.

Para poder medir la calidad de nuestros modelos entrenados, el objetivo es la clasificación de los archivos de música por género musical, para lo cual se

INTEGRACIÓN DE *BATCH AUDIO MODELS* EN SISTEMA DE RECOMENDACIÓN DE SPOTIFY



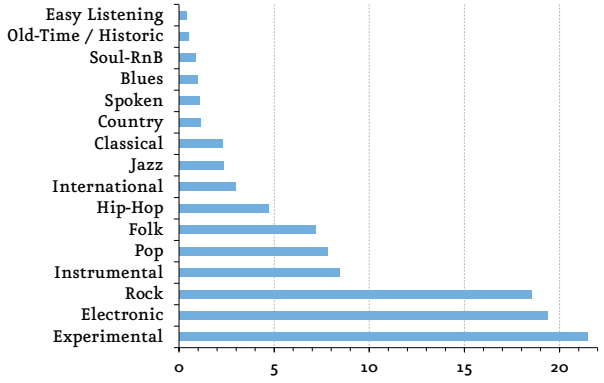
Fuente: *Conferencia Machine Learning and Big Data for Music Discovery - 2017*[2].

utilizaron **16 géneros de música** distintos: Experimental, Electronic, Rock, Instrumental, Pop, Folk, Hip-Hop, International, Jazz, Classic, Country, Spoken, Blues, Solu-RnB, Old-Time / Historic y Easy Listening.

Nuestro set de archivos de música se encuentra desequilibrado en cuanto a géneros musicales: mientras que la clase mayoritaria (Experimental) concentra un 21% del total de archivos, la minoritaria (Easy Listening) apenas representa el 0,4% del total.

Archivos por género musical

(%)



Fuente: *elaboración propia.*

Algunas otras observaciones interesantes que encontramos en el análisis de datos es que contamos con un total de 332,22 días de música, el 24% de los archivos contenían el origen de sus artistas siendo la gran mayoría de América del Norte y la UE y el 96% de las canciones con letra son en inglés.

Para el entrenamiento de nuestro modelo utilizamos 93,222 archivos de audio y 11,121 archivos como set de pruebas.

Modelización y Resultados

Nuestro proceso de modelización consta de dos ejes principales, en ambos casos utilizamos las métricas de *Accuracy* para interpretar con facilidad el resultado del modelo y *Kappa*[4] buscando maximizar una métrica funcional para clasificación multiclase que tome en cuenta la probabilidad por azar y el desbalanceo de clases.

Machine learning

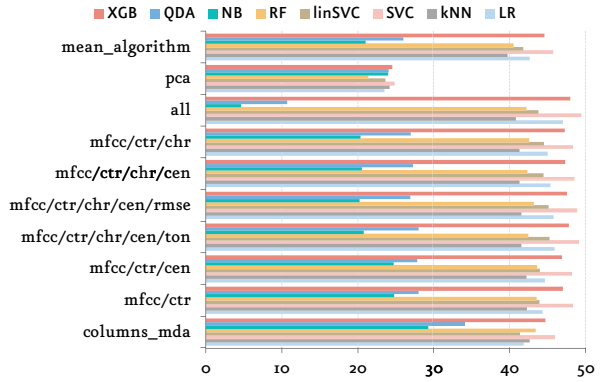
Para la utilización de algoritmos de clasificación con datos estructurados realizamos un pre-procesamiento de los archivos de audio, extrayendo patrones musicales a través de distintas técnicas de recuperación de información musical (Music Information Retrieval - MIR)[5] con la librería de python *Librosa*.[6]

Como resultado de las técnicas de recuperación de información musical se obtienen matrices numéricas que representan la señal en el dominio del tiempo o de la frecuencia, rescatando características como la velocidad, la potencia, la melodía, el timbre e incluso las armonías y acordes. Por cada matriz obtuvimos sus momentos matemáticos[7] buscando la reducción de dimensionalidad, obteniendo como resultado 518 variables.

Como métodos de selección de variables utilizamos técnicas como PCA[8], MDA[9] y diferentes combinaciones de técnicas MIR que dieron mejores

resultados. Algunas de estas técnicas son *MFCC: Mel Frequency Cepstral Coefficients, CTR: Spectral contrast, CHR: Chroma, CEN: Spectral Centroid, RMSE: Root Mean Square Energy* y *TON: Tonnetz*.

Accuracy – Selección de variables y modelos* (%)



* *Accuracy* obtenido por diferentes modelos y técnicas de reducción de dimensionalidad.
Fuente: *elaboración propia*.

Basado en el valor obtenido de *Kappa* y *Accuracy* de cada modelo, el tiempo de entrenamiento, y la diversidad de forma de construcción de cada algoritmo, se creó un *ensemble* por votación que incluye los algoritmos *Xtreme Gradient Boosting*, *Logistic Regression* y *Linear Support Vector Machine*.

El resultado final obtenido es un **46,32%** de aciertos con un Kappa de 33,22%. Pudimos observar que nuestro modelo es influenciado por las tres clases mayoritarias; experimental, rock y electrónica, debido al desbalance de clases; sin embargo, el género con mejor predicción es una clase minoritaria (Old-Time / Historic).

Deep learning

Spotify utiliza como parte de sus modelos basados en *Batch Audio Models*, arquitecturas basadas en Redes Neuronales Convolucionales (*CNN*[10]), por lo que exploramos por lo menos 3 arquitecturas distintas a las que llamaremos como:

- Modelo Spotify[11]
- Modelo Deep Sound[12]
- Modelo CNN - LSTM[13]

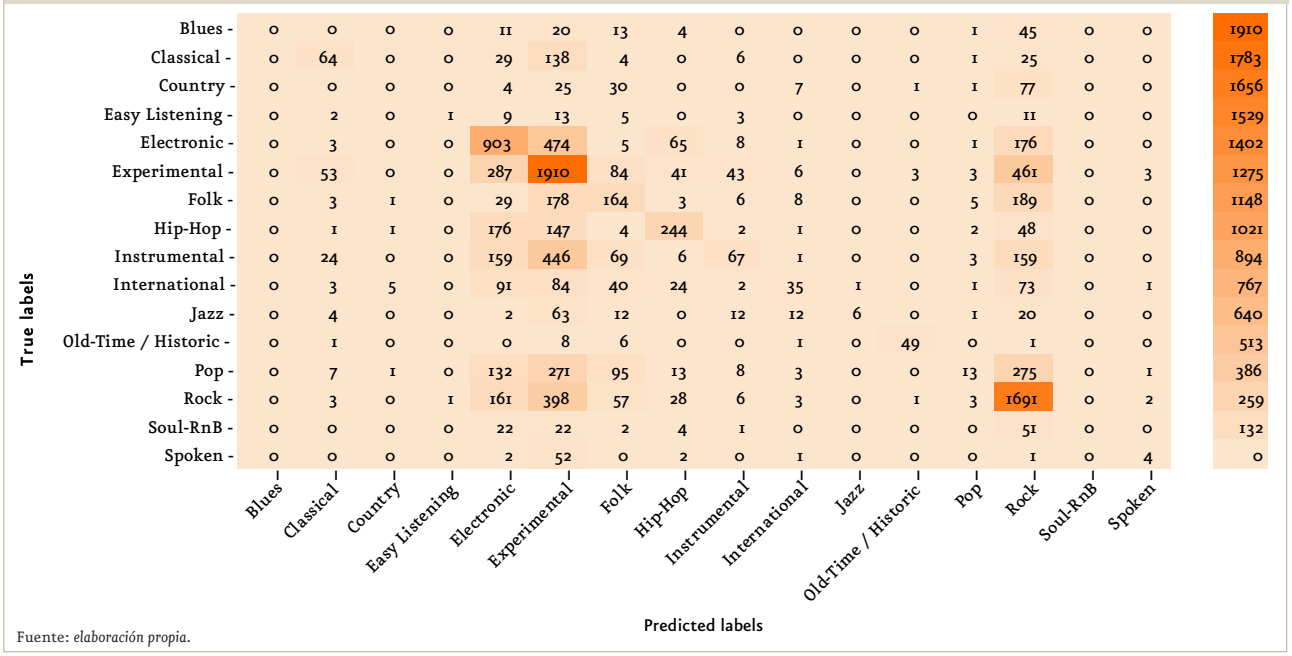
Para realizar el entrenamiento y prueba de estos algoritmos, se recortaron ventanas de audio de 30 segundos, convertimos los archivos de audio en matrices que representan un espectrograma[14] con dimensiones de 646 x 128, y se definió una estrategia *EarlyStopping*[15] donde al no existir mejoramiento durante por lo menos dos épocas en la función de pérdida, se detiene el entrenamiento.

La arquitectura con mejores resultados obtenidos es una arquitectura propuesta por nosotros que comprende el uso de Redes Convolucionales y *LSTM*, obteniendo una exactitud del **61,38%**.

Nuestra arquitectura CNN-LSTM fue utilizada para la implementación de una aplicación de visualización dinámica donde podemos observar la clasificación de archivos de audio en tiempo real[16] basada en aplicación *DeepSound*.

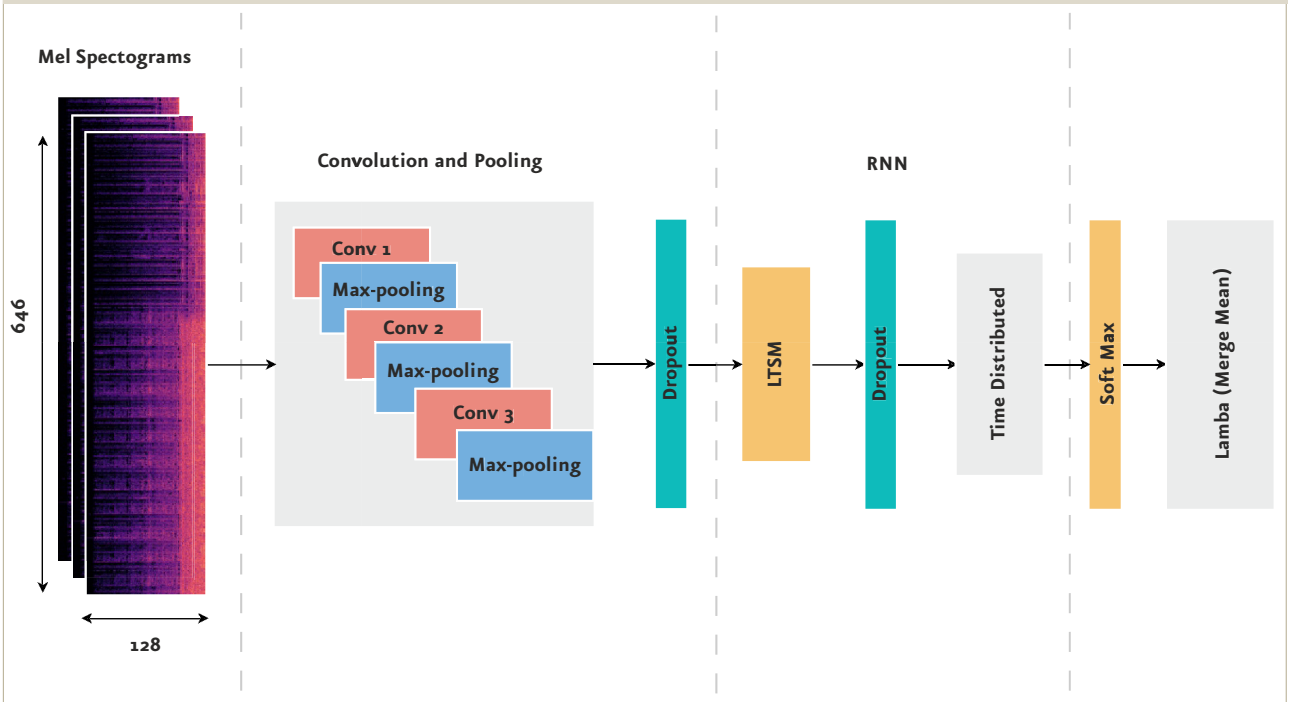
El proceso de modelización descrito requirió del uso de servicio en la nube como Google Cloud, y EC2 de Amazon Web Services (GPU y CPU).

MATRIZ DE CONFUSIÓN DE MODELO *ENSEMBLE*



Fuente: *elaboración propia*.

ARQUITECTURA RED CNN - LSTM PROPUESTA

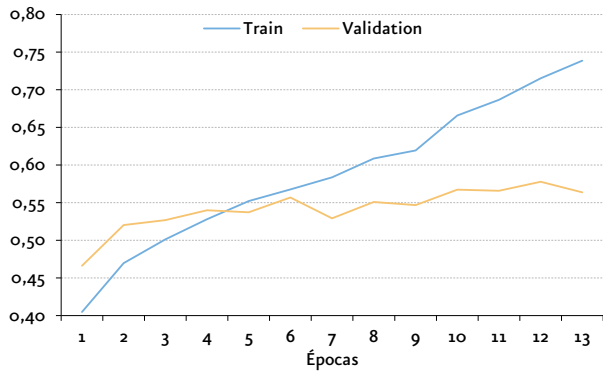


Fuente: *elaboración propia*.

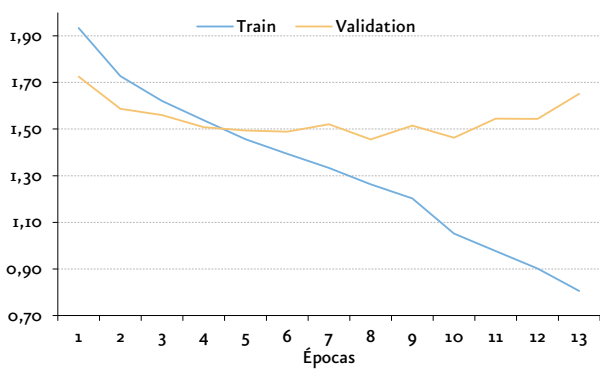
Comportamiento de la función de pérdida y el *accuracy* en fases de entrenamiento y validación de las arquitecturas propuestas

(%)

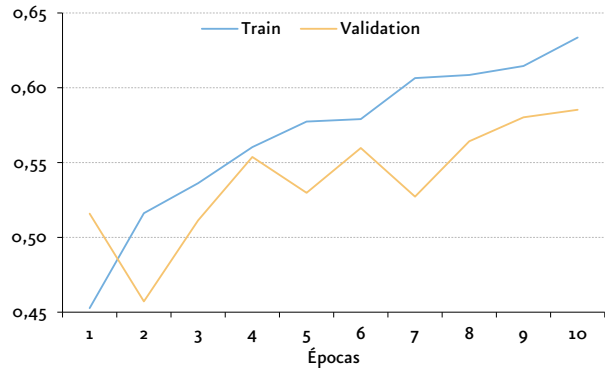
Spotify – Accuracy



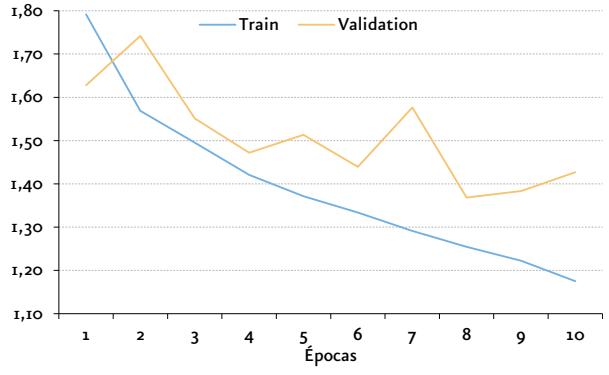
Spotify – Función de pérdida



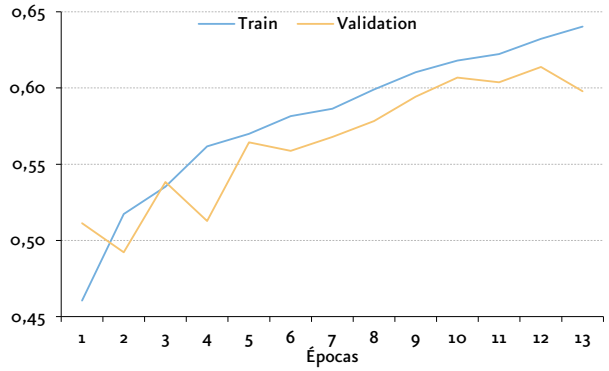
DeepSound – Accuracy



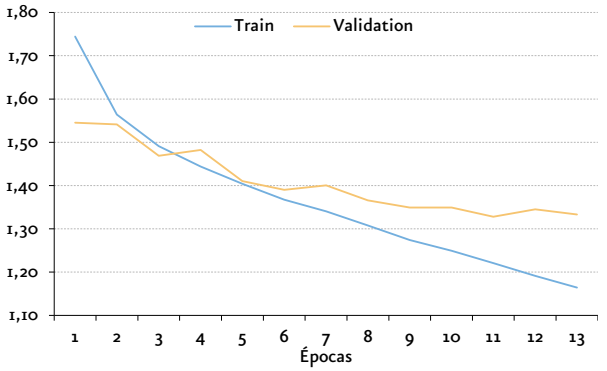
DeepSound – Función de pérdida



LTSM – Accuracy



LTSM– Función de pérdida



Fuente: *elaboración propia.*

Conclusión

Las técnicas de recuperación de información musical (MIR) demostraron eficacia en el reconocimiento de patrones musicales y aporte a la clasificación de archivos por género musical, sin embargo, con las arquitectura basadas en redes neuronales donde nuestro pre-procesamiento de datos consistió en convertir el audio a un formato de espectrograma, obtenemos mejores resultados sin el esfuerzo de entender y conocer técnicas especializadas de señales y audio.

Las arquitecturas basadas en redes neuronales nos permiten el aprovechamiento de modelos previamente entrenados, lo cual será útil para su escalabilidad, re-entrenamiento continuo e integración con modelos basados en filtros colaborativos.

Para conocer más detalles de las técnicas MIR, técnicas de selección de variables, pruebas de balanceo, y resultados detallados, puede acceder a la memoria del TFM[17] ::

[1] *Filtros colaborativos*

[2] *Conferencia Spotify*

[3] *Free Music Archive (FMA)*

[4] Kappa

[5] *Music Information Retrieval (MIR)*

[6] *Librosa*

[7] *Momentos matemáticos*

[8] *Principal Component Analysis (PCA)*

[9] *Mean Decrease in Accuracy (MDA)*

[10] *Convolutional Neural Network (CNN)*

[11] Modelo *Spotify*

[12] Modelo *Deep Sound*

[13] *Long Short-term Memory (LSTM)*

[14] *Mel Spectrogram*

[15] *EarlyStopping*

[16] Aplicación de clasificación de género musical en tiempo real

[17] Memoria TFM: Modelo de clasificación de géneros musicales basado en recuperación de información musical (MIR) y análisis de espectrogramas por Ludwig Rubio, Junio 2019.