

Utiliza «Data Science» para encontrar tu próxima canción preferida



¿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 colaborativos¹ son utilizados por distintas industrias con presencia digital debido a su comprobado éxito. Sin embargo, empresas como *Spotify* y *Youtube* inte-

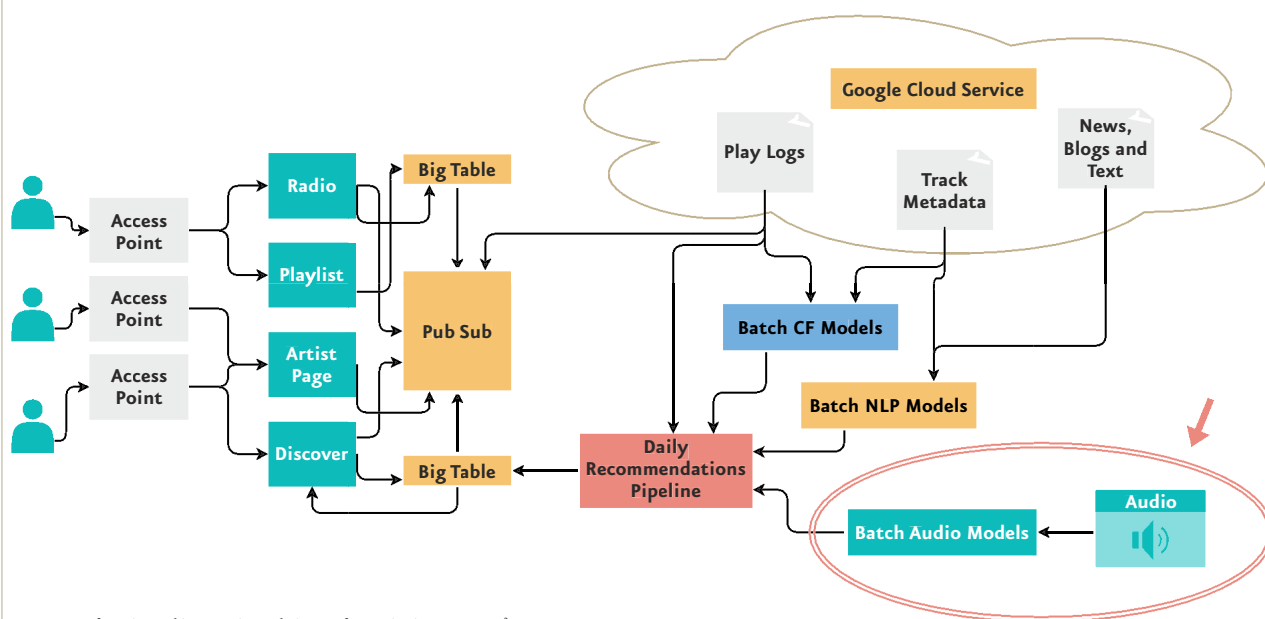
gran en sus sistemas de recomendación modelos basados en contenido como los *Batch Audio Models* para el análisis de archivos de audio, utilizados en la búsqueda de patrones musicales que puedan mejorar las recomendaciones.

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 FMA³ 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

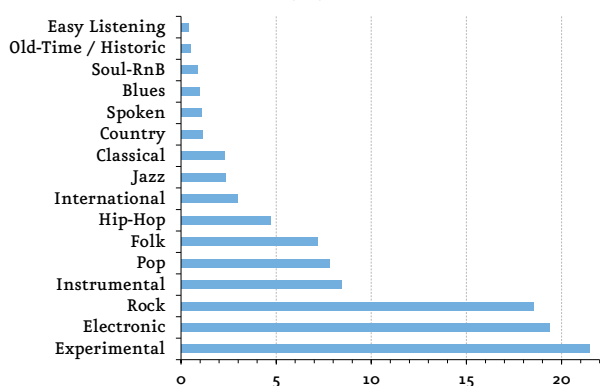


Fuente: Conferencia Machine Learning and Big Data for Music Discovery - 2017².

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

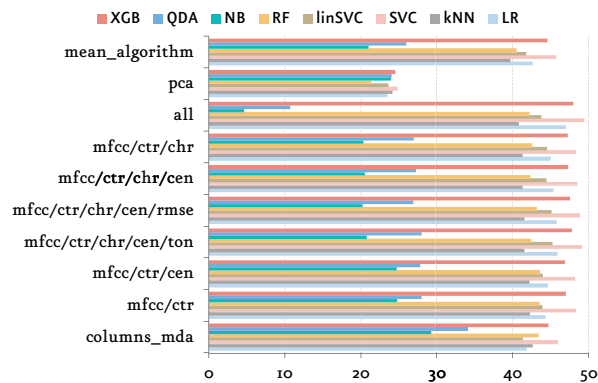
Para la utilización de algoritmos de clasificación con datos estructurados realizamos un pre-procesamiento de los archivos de audio, extrayendo patrones musicales a través de distintas técnicas de recuperación de información musical (Music Information Retrieval - MIR)⁵ con la librería de python *Librosa*.⁶

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 matemáticos⁷ buscando la reducción de dimensionalidad, obteniendo como resultado 518 variables.

Como métodos de selección de variables utilizamos técnicas como PCA⁸, MDA⁹ 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 (CNN¹⁰), por lo que exploremos por lo menos 3 arquitecturas distintas a las que llamaremos como:

- Modelo Spotify¹¹
- Modelo Deep Sound¹²
- Modelo CNN - LSTM¹³

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 espectrograma¹⁴ con dimensiones de 646 x 128, y se definió una estrategia EarlyStopping¹⁵ 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 real¹⁶ basada en aplicación DeepSound.

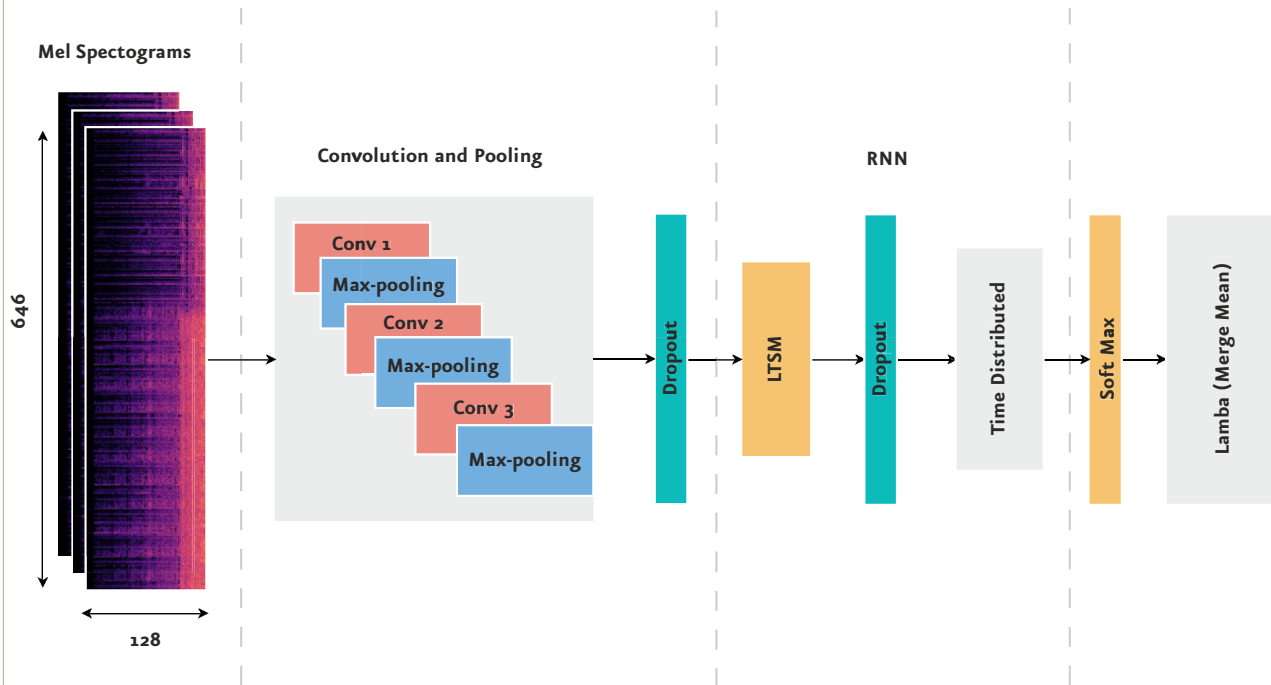
El proceso de modelización descrito requirió del uso de servicio en la nube como Google Cloud, y EC2 de Amazon Web Services (GPU y CPU).

MATRIZ DE CONFUSIÓN DE MODELO ENSEMBLE

True labels																		
	Blues	Classical	Country	Easy Listening	Electronic	Experimental	Folk	Hip-Hop	Instrumental	International	Jazz	Old-Time / Historic	Pop	Rock	Soul-RnB	Spoken		
Blues	0	0	0	0	11	20	13	4	0	0	0	0	1	45	0	0	1910	
Classical	0	64	0	0	29	138	4	0	6	0	0	0	1	25	0	0	1783	
Country	0	0	0	0	4	25	30	0	0	7	0	1	1	77	0	0	1656	
Easy Listening	0	2	0	1	9	13	5	0	3	0	0	0	0	11	0	0	1529	
Electronic	0	3	0	0	903	474	5	65	8	1	0	0	1	176	0	0	1402	
Experimental	0	53	0	0	287	1910	84	41	43	6	0	3	3	461	0	3	1275	
Folk	0	3	1	0	29	178	164	3	6	8	0	0	5	189	0	0	1148	
Hip-Hop	0	1	1	0	176	147	4	244	2	1	0	0	2	48	0	0	1021	
Instrumental	0	24	0	0	159	446	69	6	67	1	0	0	3	159	0	0	894	
International	0	3	5	0	91	84	40	24	2	35	1	0	1	73	0	1	767	
Jazz	0	4	0	0	2	63	12	0	12	12	6	0	1	20	0	0	640	
Old-Time / Historic	0	1	0	0	0	8	6	0	0	1	0	49	0	1	0	0	513	
Pop	0	7	1	0	132	271	95	13	8	3	0	0	13	275	0	1	386	
Rock	0	3	0	1	161	398	57	28	6	3	0	1	3	1691	0	2	259	
Soul-RnB	0	0	0	0	22	22	2	4	1	0	0	0	0	51	0	0	132	
Spoken	0	0	0	0	2	52	0	2	0	1	0	0	0	1	0	4	0	
	Blues	Classical	Country	Easy Listening	Electronic	Experimental	Folk	Hip-Hop	Instrumental	International	Jazz	Old-Time / Historic	Pop	Rock	Soul-RnB	Spoken		

Fuente: elaboración propia.

ARQUITECTURA RED CNN - LSTM PROPUESTA

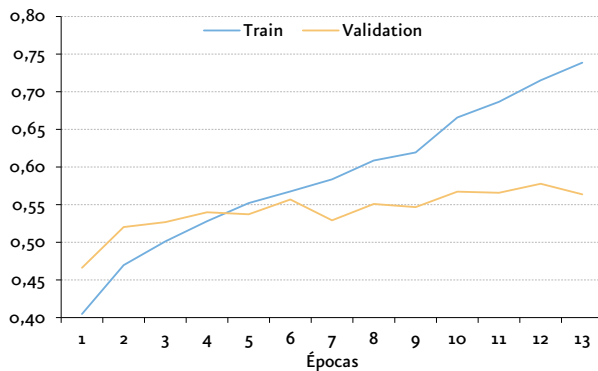


Fuente: elaboración propia.

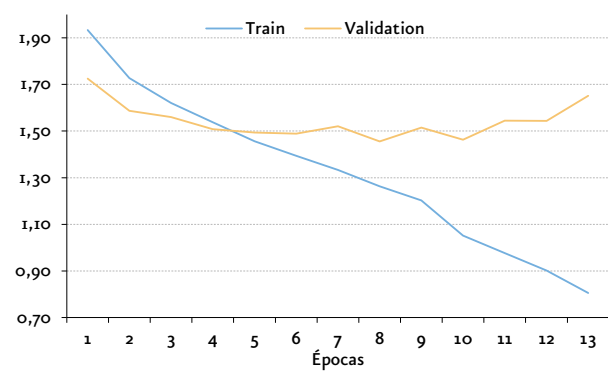
Comportamiento de la función de pérdida y el accuracy en fases de entrenamiento y validación de las arquitecturas propuestas

(%)

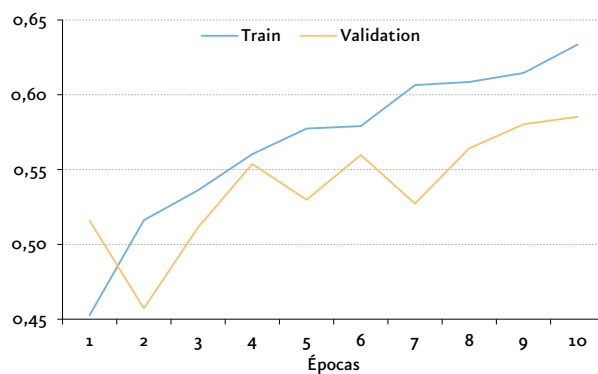
Spotify – Accuracy



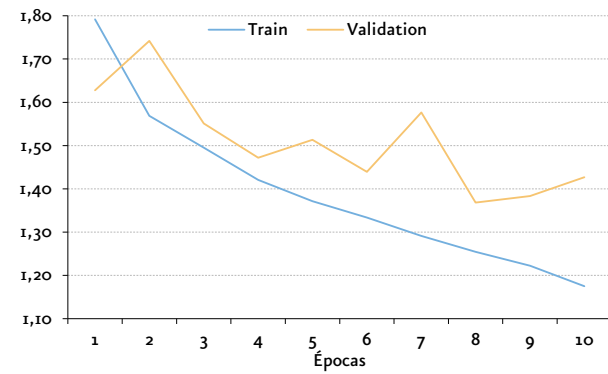
Spotify – Función de pérdida



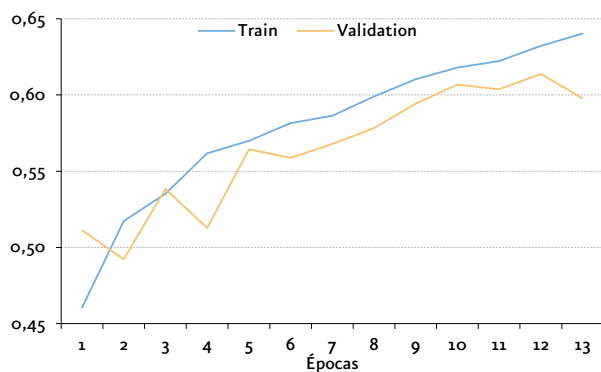
DeepSound – Accuracy



DeepSound – Función de pérdida



LTSM – Accuracy

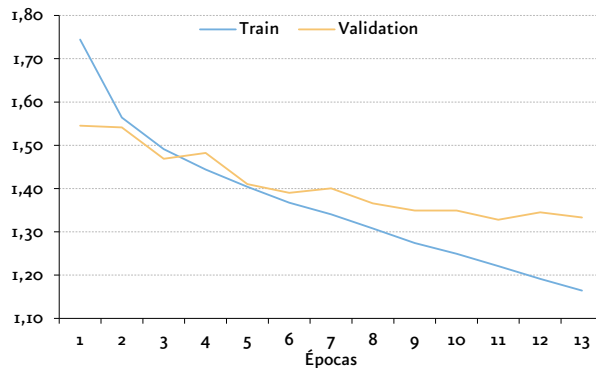


Fuente: elaboración propia.

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.

LTSM– Función de pérdida



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 TFM¹⁷ ::

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.