SISTEMA RECOMENDADOR DE CANCIONES DE ARTISTAS EMERGENTES BASADO EN PLAYLIST DE CANCIONES POPULARES

Beycker Alexis Ágredo Mosquera

Universidad Icesi
Facultad de Ingeniería
Maestría en Ciencia de Datos
2024

SISTEMA RECOMENDADOR DE CANCIONES DE ARTISTAS EMERGENTES BASADO EN PLAYLIST DE CANCIONES POPULARES

Beycker Alexis Ágredo Mosquera

Trabajo de grado

Yesid Ospitia Medina, PhD

Universidad Icesi
Facultad de Ingeniería
Maestría en Ciencia de Datos
2024

Tabla de contenido

Lista de acrónimos

AI	**A**rtificial **I**ntelligent
MER	**M**usic **E**motion **R**ecognition
PAC	**P**rincipal **C**omponents **A**nalysis

Introducción

La música ha sido reconocida a lo largo de la historia como un medio poderoso para influir en las emociones humanas. Desde tiempos antiguos, las culturas han utilizado la música como una herramienta para expresar sentimientos, promover el bienestar emocional y mejorar la calidad de vida. Actualmente, con el elevado crecimiento de las plataformas de streaming de audio como Spotify, Deezer y Amazon Prime Music, la relación entre la música y las emociones ha tomado mayor relevancia debido a la capacidad que tienen estas plataformas para llegar a millones de usuarios. Uno de los avances más destacados en el campo de la música es el desarrollo de sistemas de recomendación, que permiten personalizar la experiencia de escucha de cada usuario, ajustándose a sus preferencias musicales y estados emocionales.

Los sistemas de recomendación de canciones no solo sugieren música basada en géneros o artistas favoritos, sino que también pueden detectar patrones en las preferencias emocionales de los usuarios. Estudios en el campo de la ciencia de datos han demostrado que la música puede influir directamente en los niveles de arousal y valence, dimensiones que se utilizan para medir la activación emocional y el tono afectivo, respectivamente. Diversas investigaciones en este campo han permitido a los científicos de datos desarrollar modelos avanzados que predicen las canciones más adecuadas para diferentes momentos y estados emocionales, mejorando la satisfacción del usuario y optimizando la retención en las plataformas de streaming.

Gracias a la inteligencia artificial, el machine learning y el análisis de grandes volúmenes de datos, la ciencia de datos ha tenido importantes avances en la personalización de estos sistemas. Ya por hoy, existen algoritmos sofisticados que pueden analizar no solo el comportamiento de los usuarios, sino también las características emocionales de las canciones, lo que ha abierto la puerta a nuevas oportunidades en la manera en que se interactúa con la música. En este contexto, es importante comprender cómo las emociones y la música están interrelacionadas y cómo los sistemas de recomendación se han convertido en herramientas clave para potenciar esta conexión, permitiendo una experiencia musical más rica y significativa para los oyentes.

El objetivo de esta investigación es desarrollar un sistema de recomendación de canciones de artistas emergentes, para lo cual se empleará la base de datos MediaEval, una de las bases de datos más ampliamente utilizadas en la disciplina del reconocimiento de emociones en la música. La base de datos contiene 1802 extractos de canciones y canciones completas anotadas con valores de arousal y valence tanto de forma continua (por segundo) como durante toda la canción.

¿Cuál es la situación de interés?:

Clasificar un conjunto de canciones basados en el modelo **arousal-valence** para describir y medir las emociones que las personas experimentan al escuchar música representándolo en un espacio bidimensional.

¿Cuál es el problema?

El margen de precisión de los modelos recomendadores de música aún no alcanza valores óptimos en cuanto al nivel de precisión que requieren los sistemas MER.

¿Para quién es un problema?

Para la industria de plataformas de streaming de audio y sus diferentes usuarios, en particular, artistas y oyentes.

¿Por qué es un problema?

Porque el campo de recomendación musical basado en emociones es un campo de mucho potencial el cual falta por explorar, ya que en la actualidad los sistemas recomendadores existentes no alcanzan niveles de precisión óptimos.

¿Cuáles son las consecuencias del problema?

Pérdida de oportunidad al no proporcionar a los consumidores de contenido auditivo canciones o artistas basadas en sus emociones. Los nuevos artistas también se ven afectados, ya que sus canciones no suelen ser recomendadas por las plataformas de streaming de audio.

Antecedentes

Contexto

Uno de los primeros estudios importantes en MER fue el trabajo de Kate Hevner, quien desarrolló el "Adjective Circle" en 1936. Hevner clasificó las emociones musicales en ocho grupos de adjetivos. Este modelo fue uno de los primeros intentos de categorizar las emociones que la música puede despertar en los oyentes.

Algunos de los primeros intentos de predicción de emociones musicales implicaban enfoques basados en reglas y marcos jerárquicos. Por ejemplo, Feng, et al. [1] utilizaron la Estética de Medios Computacionales (CMA) para analizar el tempo y la articulación, mapeándolos en cuatro categorías de estado de ánimo: felicidad, enojo, tristeza y miedo, la precisión de su algoritmo alcanzó un 67%. Posteriormente Yang, et al. Utilizó un clasificador difuso KNN y un clasificador difuso de media más cercana para detectar cuatro clases de emoción musical a partir de 243 piezas de música pop moderna [3]. La mejor precisión de este nuevo método alcanzó un 78,33%.

Por otro lado, en el trabajo de Lu et al. [4] desarrollaron un marco jerárquico para extraer automáticamente la emoción musical de los datos acústicos. Se empleó la intensidad musical para representar la dimensión energética del modelo de Thayer, mientras que se utilizaba el timbre y el ritmo para capturar la dimensión del estrés. Con este modelo se alcanzó una precisión promedio en la detección del estado de ánimo de hasta el 86,3 %

Los sistemas MER requieren de un nivel de precisión muy alta, de la literatura revisada se observa que, aunque hay precisiones que ya alcanzan una precisión de casi el 87%, aún queda mucho margen de mejora.

Descripción del problema

El reconocimiento de emociones musicales (Music Emotion Recognition, MER) es un campo interdisciplinario que combina la ciencia de datos, el aprendizaje automático y la música para identificar y clasificar las emociones que las canciones evocan en los oyentes. Los sistemas MER hacen uso de la tecnología de procesamiento de señales moderna para lograr la extracción de características musicales y métodos de aprendizaje automático para lograr la regresión o clasificación de emociones musicales [1]. Aunque las investigaciones han avanzado significativamente en los últimos años, este campo enfrenta diversos desafíos y limitaciones tanto desde una perspectiva técnica como conceptual. Entre los más destacados se encuentran:

1. **Subjetividad de las emociones**: Uno de los mayores desafíos en MER es la naturaleza subjetiva de las emociones. Las mismas canciones pueden evocar respuestas emocionales diferentes en distintas personas, lo cual depende de factores como el contexto cultural, la experiencia previa o el estado de ánimo. Lo anterior dificulta la creación de etiquetas consistentes y fiables para entrenar modelos de aprendizaje automático.

2. **Representación y extracción de características musicales**: La música es un fenómeno complejo, compuesto por varios elementos como el timbre, la armonía, el ritmo y la melodía. Encontrar representaciones adecuadas que capten todos estos aspectos es un desafío. Si bien las características de bajo nivel como el espectrograma o el tempo pueden ser útiles, las emociones suelen estar asociadas con interacciones más abstractas entre diferentes elementos musicales.

3. **Interpretabilidad de los modelos**: Los modelos de aprendizaje profundo como las redes neuronales son eficaces para procesar señales musicales, pero a menudo son difíciles de interpretar. Esto limita la capacidad de los investigadores para comprender qué características de la música están siendo utilizadas para predecir una emoción en particular, dificultando la optimización y el rendimiento.

4. **Evaluación de rendimiento**: La evaluación del rendimiento de los modelos de MER es compleja, ya que las emociones no son entidades discretas que puedan clasificarse fácilmente. En lugar de emociones categóricas, las

emociones continuas, como las representadas en el espacio de valence-arousal, son más representativas, pero también más difíciles de modelar y evaluar.

El reto en este proyecto es desarrollar un sistema de reconocimiento de emociones a partir de un conjunto de 260 características de bajo nivel junto con sus valores de arousal y valence el cual arroje una precisión aceptable con base en los trabajos anteriores. Adicionalmente, se espera desarrollar un algoritmo que genere recomendaciones de canciones de artistas emergentes a partir de canciones de listas de reproducción populares, con el fin de dar a conocer estos artistas en surgimiento y mejorar la experiencia de los oyentes al momento de crear sus listas de reproducción.

Objetivos

Objetivo general

Diseñar un sistema recomendador de canciones para generar listas de reproducción de artistas emergentes a partir de playlist de artistas populares.

Objetivos específicos

1. Extraer las características existentes en el dataset MediaEval en nuevas canciones.
2. Desarrollar un modelo de predicción de los valores de arousal y valence
3. Desarrollar un modelo de clasificación que permita recomendar canciones por similaridad musical basado en emociones.

Marco Teórico

Para la presentación del marco teórico del proyecto se abordará la definición de los conceptos Affective Computing, Music Emotion Recognition y modelo arousal-valence. La definición de estos conceptos permitirá comprender las metodologías, tecnologías y aspectos fundamentales que serán el punto de partida para el desarrollo de la solución al problema planteado.

Affective Computing

La computación afectiva es un campo interdisciplinario que se enfoca en el desarrollo de sistemas y dispositivos capaces de reconocer, interpretar, procesar y simular emociones humanas. Este campo combina la informática, la psicología y la ciencia cognitiva (Vergara, 2024). La hipótesis principal de este concepto es que las emociones juegan un papel importante en la toma de decisiones humanas, por lo que resulta muy útil que los sistemas computacionales puedan entender y responder a los estados de ánimo de las personas.

Existen muchos métodos para capturar los datos emocionales, entre los más comunes encontramos el reconocimiento facial, el análisis de voz y los sensores fisiológicos. Los sensores pueden generar señales que posteriormente pueden dar una interpretación de la emoción o sensaciones que experimentan las personas, esos datos pueden ser usados por algoritmos de aprendizaje automático para dar solución a diversos problemas relacionados con el comportamiento y las emociones humanas.

Algunas aplicaciones de software que pueden beneficiarse de las aplicaciones de la computación afectiva son:

- **Mejora de la experiencia del usuario (UX)**: En el desarrollo de interfaces de usuario, la computación afectiva permite crear experiencias más personalizadas y atractivas.
- **Educación y entretenimiento**: Los sistemas podría monitorear el estado de ánimo y las emociones de los estudiantes para identificar donde pueden necesitar ayuda adicional.
- **Salud y bienestar:** Los sistemas basados en computación afectiva se pueden emplear para monitorizar el estado emocional de los pacientes, especialmente aquellos que padecen de trastornos mentales, ya que podría detectar cuando un paciente está pasando por un episodio de ansiedad o un cuadro de depresión para generar alarmas en los profesionales de la salud.

- **Entretenimiento:** La computación afectiva también tiene aplicación en el ámbito de entretenimiento. Por ejemplo, al ser usada en videojuegos, podría adaptar la música, los colores y la interfaz de acuerdo con el estado emocional del jugador.

A pesar de las aplicaciones y beneficios que ofrece la computación afectiva, esta también enfrenta una serie de desafíos en cuestiones de ética y privacidad. En términos de privacidad, la recolección de datos emocionales, dependiendo del contexto, podría ser intrusiva, además existe el riesgo de manipulación emocional donde los sistemas podría influir de maneras no deseadas o éticamente cuestionables. Es importante destacar que al manipular este tipo de aplicaciones que utilicen los datos emocionales de las personas se siga un conjunto de regulaciones y medidas, como pueden ser el tratamiento y la protección de los datos; además contar con el consentimiento informado de los usuarios tales que garanticen la privacidad y el uso adecuado de esos datos.

Music Emotion Recognition

Music Emotion Recognition (MER) es un campo interdisciplinario dentro de la ciencia de datos, el aprendizaje automático y la musicología que busca identificar y clasificar las emociones que una pieza musical provoca en los oyentes. El objetivo es automatizar el reconocimiento de las emociones asociadas a la música utilizando algoritmos que analizan diversas características de las canciones.

Los primeros estudios que adoptaron este enfoque en la clasificación de emociones en música se inspiraron en teorías psicológicas de las emociones humanas, como las propuestas por Paul Ekman y Robert Plutchik, quienes identificaron categorías básicas de emociones humanas.

El reconocimiento de emociones musicales (REM) (Music Emotion Recognition MER) se deriva de dos campos mayores: la recuperación de información musical (RIM) (Music Information Retrieval MIR) y la computación afectiva (Affective Computing) (Calvo, D’Mello, Gratch y Kappas, 2014; Picard, 2000).

Modelo Arousal-Valence

El modelo arousal-valence es una forma de describir y medir las emociones que las personas experimentan al percibir algún estímulo musical, el cual se representa en un espacio bidimensional de dos ejes. Los modelos dimensionales sugieren que la emoción se entiende mejor como algo que ocurre dentro de un espacio dimensional, más comúnmente un espacio que abarca los ejes de valencia (valence) y excitación (arousal) [7].
El arousal es la dimensión que refleja la energía invertida durante la emoción. Este eje mide el nivel de activación o excitación que provoca la música. Va desde un estado de baja activación (tranquilo, relajado) hasta alta activación (excitado, enérgico). La valencia mide la cualidad emocional de la experiencia, que va desde emociones negativas (tristeza, enojo) hasta emociones positivas (alegría, felicidad).

Ambos ejes proporcionan una forma clara de categorizar y medir las emociones, lo cual es importante en varias disciplinas, desde la psicología y la neurociencia hasta la inteligencia artificial y la música. La medición de emociones en los modelos de AI basados en Arousal y Valencia suelen ofrecer respuestas más personalizadas, lo cual es útil en asistentes virtuales, videojuegos y robots de interacción social, donde el objetivo es entender y responder a las emociones humanas de manera adaptativa.

En el presente trabajo, los ejes de arousal y valencia generados a partir de las características de audio permitirán conocer las emociones evocadas en diferentes momentos del clip musical, lo que será de gran importancia para el modelo de recomendación de canciones de artistas emergentes a partir de los valores de arousal y valencia de canciones populares en las plataformas de streaming de audio.

Estado del Arte

Para presentar el estado del arte del proyecto, se identifican y describen soluciones existentes frente al desarrollo de sistemas recomendadores de música basados en emociones.

Griffiths Darryl, Cunningham Stuart, Weinel Jonathan, Picking Richard, (2021), A multi-genre model for music emotion recognition using linear regressors

El artículo describe un enfoque innovador para el reconocimiento de emociones en música a través de un modelo de múltiples géneros basado en regresores lineales. El objetivo de dicha investigación era desarrollar un sistema que pudiera organizar la gran cantidad de música disponible para el público en general de una manera significativa y personalizada, proporcionando un mecanismo por el cual la música pudiera ser etiquetada emocionalmente.

Para entrenar el modelo, se realizaron dos estudios; el primero utilizó 20 canciones de distintos géneros musicales, como rock, jazz, clásica y reggae. Se recopilaron autoevaluaciones emocionales de 44 participantes, quienes escucharon cada canción y evaluaron las emociones percibidas e inducidas. Las características de audio, como energía, espectro y ritmo, fueron correlacionadas con las dimensiones de arousal y valence mediante análisis de regresión.

El segundo estudio validó el modelo con un nuevo conjunto de 40 canciones, recolectando evaluaciones de 158 participantes. Los resultados indicaron una fuerte correlación entre las características del audio y las emociones percibidas, especialmente en el arousal, con un coeficiente de determinación (R^2) del 85% para la predicción de arousal y 77.6% para la valencia cuando se ajustaron los datos eliminando valores atípicos.

Los resultados arrojaron que ese enfoque puede ser eficaz para calificar emocionalmente la música, particularmente en la predicción de valence, además se destaca la eficiencia del modelo y su capacidad de generalización a través de géneros musicales diversos. Sin embargo, los autores del artículo señalan que mejorar la predicción de la valence, ampliar el conjunto de datos y utilizar características más complejas podrían optimizar el modelo en futuros trabajos.

Yu Xia, Fumei Xu, (2022), Study on Music Emotion Recognition Based on the Machine Learning Model Clustering Algorithm

El artículo de investigación propone un sistema de reconocimiento de emociones musicales utilizando técnicas de machine learning y un algoritmo de clasificación por agrupamiento. El estudio se basó en el modelo bidimensional de emociones de Thayer, el cual mapea las emociones en un plano de arousal y valence. La metodología incluye la extracción de características musicales clave, como la energía, la melodía, el dominio del tiempo, el dominio de la frecuencia, el ritmo y la armonía, y la aplicación de modelos de regresión y clasificación para predecir las emociones.

En el artículo se presenta un clasificador híbrido que combina varios algoritmos de aprendizaje supervisado, tales como Support Vector Machine (SVM), k-nearest neighbors (KNN) y Redes Neuronales Difusas. El clasificador logró una tasa de reconocimiento de emociones del 84.9%, mejorando los resultados de otros enfoques anteriores.

Por otro lado, en el artículo convierten el problema de clasificación de emociones en un problema de regresión, donde se predicen las posiciones de las emociones en el plano de valence-arousal. Se utilizaron técnicas como la regresión por vectores de soporte y la regresión por función de base radial, obteniendo precisiones del 81% para arousal y 72% para valence.

El estudio analizó la proyección espacial basada en el análisis de componentes principales y el algoritmo basado en relieve. El análisis de componentes principales (PCA) se usó como un método eficaz para comprimir y extraer información en muestras basadas en una matriz de covarianza variable que podía reducir eficazmente el número de características que contenían ruido o redundancia. El algoritmo de relieve se usó ya que es uno de los algoritmos de ponderación de selección de características más comúnmente utilizados,

En cuanto a las mejoras en la precisión, comparado con métodos tradicionales como la clasificación por SVM, el enfoque de regresión híbrido mejoró la precisión en un 14%. Los experimentos demostraron que combinar múltiples técnicas de clasificación y regresión es más efectivo para el reconocimiento de emociones.

Van Loi Nguyen, Donglim Kim, Van Phi Ho, Younghwan Lim, (2019), A New Recognition Method for Visualizing Music Emotion

El artículo propone un enfoque híbrido que combina los modelos categórico y dimensional para el reconocimiento de emociones en música. Los investigadores hacen uso del modelo de valence-arousal de Thayer, el cual se divide en 36 categorías emocionales. En el estudio se recopilaron 300 clips de música con emociones anotados por expertos, extrayendo características acústicas para entrenar modelos supervisados.

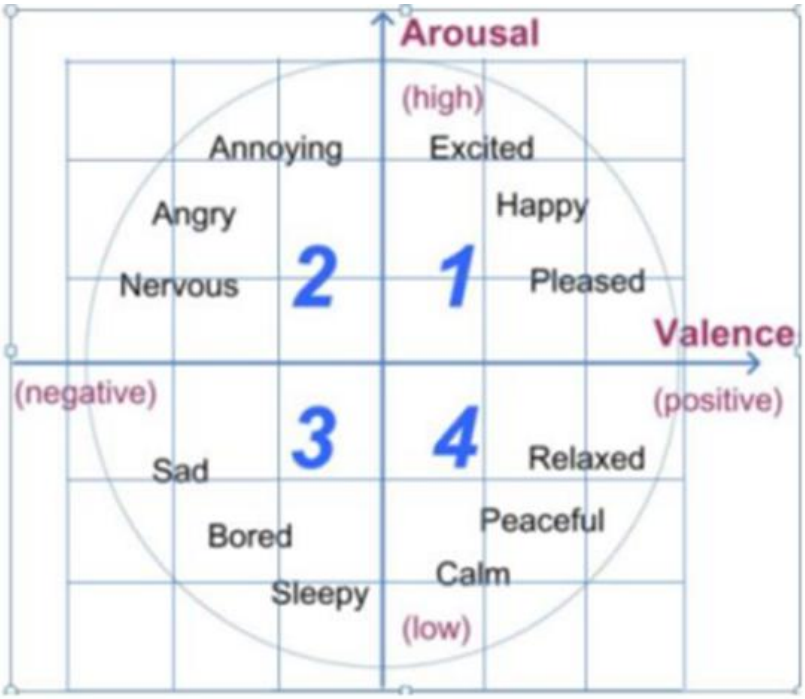

Figura 1. Modelo circumplejo de Russell

El sistema se desarrolló en dos etapas principales: la construcción del modelo de entrenamiento y la clasificación de emociones. Se seleccionaron características clave de la música, como energía, ritmo, espectro y armonía, y se utilizaron técnicas de aprendizaje automático para evaluar diferentes algoritmos de clasificación. La selección de atributos mejoró significativamente el rendimiento, reduciendo el tiempo de cómputo y aumentando la precisión.

El artículo ofrece las siguientes contribuciones originales relacionadas con el campo de la recuperación de información musical (MER/MIR):

A. El primer estudio que combina enfoques dimensionales y categóricos en la MER basados en el nivel de valence y arousal.

B. La precisión del método propuesto mejora notablemente en comparación con trabajos anteriores, especialmente para el reconocimiento de valence.

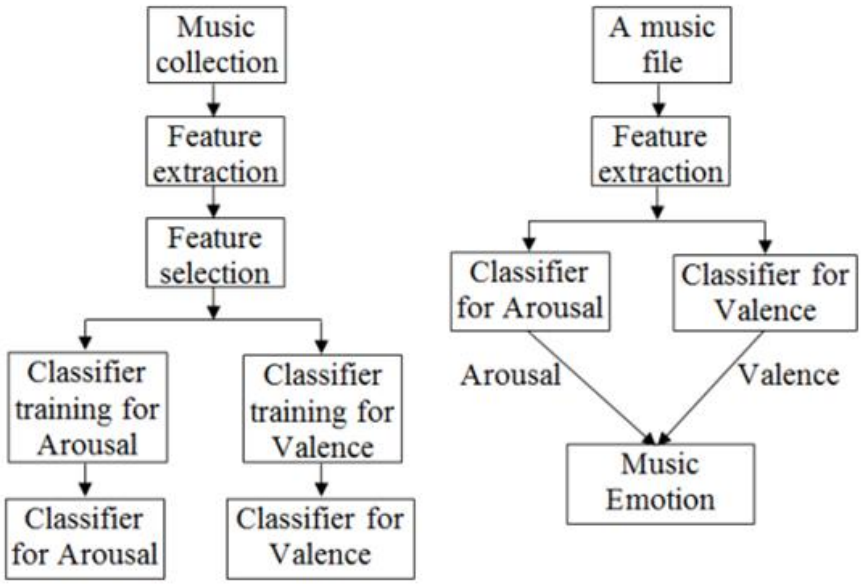

Figura 2. Diagrama del algoritmo de reconocimiento de emociones propuesto

Por otro lado, en el artículo se destaca la subjetividad en la percepción emocional como uno de los principales retos del reconocimiento de emociones musicales, también la dificultad de representar emociones de manera universal.

Finalmente, el modelo desarrollado mostró que el clasificador RandomForest obtuvo una precisión del 70% para el valor de arousal y 57.3% para valence, superando trabajos anteriores. Los resultados pueden tener aplicación en la visualización de emociones en la música. Aunque los resultados del método propuesto representan una mejora, los autores sugieren que la combinación de características de nivel bajo y medio, como letras y progresiones de acordes, junto con un dataset más grande, podría aumentar aún más la precisión en el futuro.

Metodología

1. Análisis exploratorio de los datos

La base de datos MediaEval consta de varios directorios con diferentes documentos CSV que contienen datos relacionados con las características extraídas de cada archivo de sonido, anotaciones de tipo arousal-valence sobre cada una de las canciones, 1802 archivos de sonidos, y una carpeta con metadata de cada uno de los recursos de audio. La primera actividad que se realizó fue el análisis exploratorio de los datos para:

1. Entender cómo estaban estructurados los datos.

2. Revisar cuáles eran los tipos de datos a manejar.

3. Validar si alguno de los datasets tenía valores faltantes o se debía realizar alguna actividad para organizar los datos.

Del análisis se puede destacar que el dataset con la metadata contaba con varios valores faltantes, especialmente en la categoría de géneros y el segmento de inicio y finalización de la canción; Sin embargo, para el análisis y la generación de modelos de predicción de valores valence-arousal esto no representaba ningún impedimento.

Del análisis también se puede mencionar que la base de datos está enriquecida con múltiples géneros musicales distribuidos de una forma relativamente equitativa con un promedio de 18 canciones por género músical. Lo anterior indica que no hay un sesgo significativo en los datos que pueda afectar el entrenamiento del modelo de predicción.

2. Elección de la ventana de tiempo por canción

En el análisis de los datos se pudo evidenciar que el dataset con los valores de arousal tomaba marcos de tiempo desde los 15000 milisegundos hasta los 626500 milisegundos, mientras que el de valence iba desde los 15000 milisegundos hasta los 626000 milisegundos. En ambos datasets, posterior a los 45000 milisegundos las canciones comenzaban a presentar muchos valores faltantes, dado que el tiempo de duración de cada canción es diferente. Se decide trabajar con un marco de tiempo desde los 15000 milisegundos hasta los 43500 milisegundos, ya que la canción anotada con menor duración llegaba a los 43500 milisegundos, por lo que trabajar con este rango de tiempo evitaría realizar una actividad de imputación de datos.

3. Generación de características para las nuevas canciones

Las características generadas en la base de datos MediaEval fueron extraídas con OpenSmile. OpenSmile es una herramienta ampliamente utilizada en la investigación de audio y análisis de señales acústicas, particularmente en música y habla la cual

fue desarrollada por el audEERING GmbH y es una biblioteca de código abierto que permite extraer características o features de señales de audio. La base de datos MediaEval selecciona un total de 260 características por audio, entre la cuales se pueden encontrar valores relacionados con el tono musical, la energía, el espectro de la señal, los Coeficientes MFCC, El Suavizado temporal de los contornos de pitch y otra cantidad enorme de características importantes para la predicción de los valores arousal y valencia.

Para la extracción de nuevas características era necesario contar con el script de configuración necesario que permitiera la extracción exacta de las 260 características. Dado que OpenSmile funciona desde el terminal local del equipo en el cual se encuentre instalado, fue necesario desarrollar una función en Python que permitiera la ejecución automatizada del script; de no haberse hecho de esta forma, hubiera sido necesario la extracción manual de las características desde el terminal local.

La función desarrollada en Python quedó automatizada para realizar la extracción de las características en varios audios simultáneamente, para esto, se debe configurar en python la ruta donde se encuentran alojadas las canciones a las que se le va a hacer la extracción de las características y la ruta de salida para los archivos CSV.

```
import os
import subprocess

# Rutas de archivos y carpetas playlist
audio_folder = "audios/playlist"
output_folder = "features_nuevos/playlist"
```

Dado que el objetivo del proyecto es poder generar listas de reproducción de artistas emergentes a partir de canciones populares en las plataformas de streaming de audio, se requiere realizar la extracción de características de audio de canciones de artistas emergentes. Para cumplir con el objetivo anterior, se decide utilizar la base de datos ENSA que contiene 59 archivos de audio de artistas emergentes. Adicionalmente, se escoge un conjunto de 14 canciones de una playlist popular entre las cuales encontramos los géneros pop, rock, bachata, electro pop y reggaeton. La extracción de las características de audio y predicción de valores de arousal y valence de la playlist de canciones populares se usará posteriormente para el modelo desarrollado de recomendación de listas de reproducción por similaridad emocional para artistas emergentes.

4. Limitaciones en trabajos relacionados con MER

En cuanto a las limitaciones de los modelos de aprendizaje automático generados para reconocimiento de emociones musicales, la subjetividad en el etiquetado musical es considerada como la limitación de mayor relevancia en este tipo de trabajos.

La percepción de las emociones es subjetiva y las personas pueden percibir diferentes emociones para una misma canción, lo que ocasiona que la predicción de emociones evocadas a partir de la música se enfrente a múltiples problemas. En primer lugar, la subjetividad a la hora de categorizar la música con un valor emocional hace que la evaluación del rendimiento de un sistema MER sea particularmente difícil, ya que resulta muy complejo llegar a un acuerdo común sobre el resultado de la clasificación. En segundo lugar, no es fácil describir las emociones de forma universal, porque los adjetivos que se utilizan para describir las emociones pueden ser ambiguos y el uso de adjetivos para la misma emoción puede variar de una persona a otra.

La limitación de la subjetividad en el etiquetado músical se observa claramente en los datos anotados por los evaluadores en la BD MediaEval, donde se puede apreciar cómo difiere en gran medida el etiquetado de un fragmento de canción de una persona a otra. Es importante mencionar este problema, ya que en varios trabajos relacionados ha supuesto un reto a la hora de entrenar modelos de machine learning para la predicción de valores de arousal y valence.

	sample_15000ms	sample_15500ms	sample_16000ms	sample_16500ms	sample_17000ms
0	-0.039130	-0.039130	-0.039130	-0.039130	-0.039130
1	-0.521740	-0.521740	-0.521740	-0.521740	-0.521740
2	-0.384860	-0.399610	-0.402170	-0.402170	-0.402170
3	0.088233	0.088233	0.088233	0.088233	0.088233
4	0.834780	0.834780	0.834780	0.834780	0.834780
5	-0.191140	-0.189590	-0.181190	-0.221960	-0.256300
6	-0.319560	-0.319560	-0.319560	-0.319560	-0.319560
7	0.143480	0.143480	0.143480	0.143480	0.143480
8	-0.173910	-0.173910	-0.173910	-0.173910	-0.173910
9	-0.169560	-0.169560	-0.169560	-0.169560	-0.169560

Figura 3. Diferencias en el etiquetado emocional de 10 evaluadores en el dataset de arousal de una canción (*annotations per each rater)*

5. Entrenamiento de modelos para la predicción de valores arousal-valence

Dado que se quiere predecir valores continuos de arousal y valence, se decide entrenar inicialmente 3 modelos: Random Forest Regressor, Red Neuronal Profunda

en Keras y Support Vector Regression con Kernel RBF. Para todos los modelos de entrenamiento se realiza una división de datos del 80% de los datos para entrenamiento y 20% para evaluación. Es importante mencionar que se utilizó la totalidad de los datos de la base de datos para realizar el entrenamiento de los modelos, es decir, los 1802 datos.

El modelo de Random Forest Regressor fue el más sencillo de entrenar, ya que fue el único en el que no se estandarizaron los datos y arrojó el segundo mejor valor de MSE con un 0.0223 para arousal y un 0.0216 en valence.

Para el segundo modelo, la Red Neuronal Profunda en Keras se diseñó un modelo con una estructura de red densa (fully conected) con 3 capas ocultas y una capa de salida. Cada capa oculta tiene unidades ReLU para ayudar en la convergencia, además se experimentó con funciones de activación y regularización con el fin de mejorar el rendimiento y minimizar el MSE. Para la preparación de los datos de la red neuronal, se normalizaron las características para facilitar el entrenamiento de la red neuronal. Para el entrenamiento del modelo se utiliza la función de pérdida de MSE y se agrega Dropout para reducir el sobreajuste, dado que el modelo puede aprender patrones específicos del conjunto de entrenamiento si no se regulariza. Finalmente el modelo utilizó el optimizador Adam ya que suele converger rápido en redes neuronales profundas y es un optimizador robusto. Este modelo arrojó un valor MSE de 0.0202 en arousal y 0.0192 en valence.

El tercer modelo se desarrolló con un modelo de regresión no lineal, es el caso de Support Vector Regression (SVR) con un kernel RBF. Dado que los modelos de SVR son sensibles a las escalas, se normalizaron las características previo al entrenamiento del modelo. Se utilizó el kernel RBF ya que en SVR permite al modelo aprender relaciones no lineales entre las características y las variables emocionales. Los hiperparametros que recibe este modelo son: ***C*** el cual controla el margen de error del SVR (valores más altos de C permiten un ajuste más estricto); ***epsilon*** el cual permite definir una zona de tolerancia en la que las predicciones no son penalizadas si están cerca de los valores reales. Los hiperparametros se configuran con los valores 1,0 y 0,1 respectivamente. La evaluación de este último modelo arroja valores de MSE de 0.0237 para arousal y 0.0238 para valence.

Adicionalmente al último modelo se decide implementar una estrategía de Feature Selection y GridSearchCV para validar si mejoraba la métrica MSE; no obstante, aparte de que el modelo demoró muchísimo más en su entrenamiento, arrojó las métricas más bajas de MSE con un 0.0811 para arousal y un 0.0548 para valence.

La siguiente tabla resume los resultados de la evaluación de los cuatro modelos mencionados anteriormente:

Modelo Usado	MSE Arousal	MSE Valence
RandomForestRegressor	0.0223	0.0216
Red neuronal profunda en Keras	0.0202	0.0192
Support Vector Regression (SVR) con un kernel RBF	0.0237	0.0238
Support Vector Regression (SVR) con un kernel RBF + Feature Selection + GridSearchCV	0.0811	0.0548

6. Generación de descriptores de error

Por la naturaleza del modelo bidimensional para caracterizar las emociones según los ejes de arousal y valence, la métrica de error MSE no es suficiente para conocer la precisión de los modelos entrenados. Si bien, evaluar la calidad del modelo predictivo a través de la diferencia entre el valor real y el valor predicho funciona en este modelo, no es suficiente para calificar su calidad de predicción. Es importante tener en cuenta en la métrica qué valores se predicen cómo positivos cuando su valor real era negativo y viceversa, ya que, aunque estos valores podrían no distar en una cantidad significativa de sus valores reales, el cambio de signo representa una emoción por lo general opuesta a la que debería evocar en la realidad. Dada la necesidad de conocer el cambio de signo en los valores predichos, se genera una métrica de error para realizar un conteo de esos valores; Adicionalmente se tuvieron en cuenta también los valores predichos que distan entre 0,2 y 0,3 decimales y los que distan por encima de los 0,3 decimales, cada uno con su contador independiente.

Los resultados obtenidos a partir de los 3 contadores para evaluar la calidad del modelo SVR se describen en las siguientes dos tablas

Métrica de error valores predichos de Arousal - Modelo SVR	
Valores predichos que distan entre 0,2 y 0,3 decimales	19520
Valores predichos que distan más de 0,3 decimales	40612
Valores predichos cuyo signo cambia	44649

Métrica de error valores predichos de Valence - Modelo SVR	
Valores predichos que distan entre 0,2 y 0,3 decimales	20415
Valores predichos que distan más de 0,3 decimales	31101
Valores predichos cuyo signo cambia	46547

Teniendo en cuenta que se tiene un total de 104,516 valores de arousal y valencia predichos, la métrica de los resultados del descriptor, indica una mala predicción, ya que el 44,5% de los valores predichos de Arousal se están prediciendo con un signo contrario al real y el 42,7% para el caso de los valores predichos en valence. En cuanto a los valores que están distando más de 0,2 decimales, para arousal se tiene el 49,3% de los datos y en valence llega hasta el 57,5%.

Por otro lado, al aplicar el descriptor de error en los valores predichos de arousal y valence del modelo de regresión de árboles aleatorios los resultados mejoran sustancialmente:

Métrica de error valores predichos de Arousal - Modelo Random Forest Regressor	
Valores predichos que distan entre 0,2 y 0,3 decimales	2569
Valores predichos que distan más de 0,3 decimales	1173
Valores predichos cuyo signo cambia	7059

Métrica de error valores predichos de Valence - Modelo Random Forest Regressor	
Valores predichos que distan entre 0,2 y 0,3 decimales	2574
Valores predichos que distan más de 0,3 decimales	1016
Valores predichos cuyo signo cambia	8304

En el caso del modelo entrenado con regresión de árboles aleatorios, la métrica de los resultados del descriptor, indica una mejor predicción, ya que solo el 6,7% de los valores predichos de Arousal se están prediciendo con un signo contrario al real y el 7,9% para el caso de los valores predichos en valence. En cuanto a los valores que están distando entre 0,2 y 0,3 decimales, tanto para arousal como para valence se tiene solo el 2,5%. Finalmente, los valores que distan más de 0,3 decimales, en el caso de arousal se tiene el 1,1% de los datos totales y en valence solo el 1,0% de todos sus datos.

Finalmente, al aplicar el descriptor de error en los valores predichos de arousal y valence del modelo entrenado con la red neuronal en Keras, los resultados no son los más óptimos, siendo valores cercanos a las métricas de error arrojados por el modelo SVR:

Métrica de error valores predichos de Arousal - Red Neuronal en Keras	
Valores predichos que distan entre 0,2 y 0,3 decimales	18505
Valores predichos que distan más de 0,3 decimales	43564
Valores predichos cuyo signo cambia	48607

Métrica de error valores predichos de Valence - Red Neuronal en Keras	
Valores predichos que distan entre 0,2 y 0,3 decimales	20831
Valores predichos que distan más de 0,3 decimales	31311
Valores predichos cuyo signo cambia	48913

7. Generación de nuevos datos de entrenamiento de Arousal y Valence a partir de filtrado de fragmentos con baja variabilidad

Anteriormente, en el punto 4 se enfatizó sobre las limitaciones en trabajos relacionados con MER y cómo el etiquetado de los valores de arousal y valence difieren significativamente entre los evaluadores. Al revisar la desviación estándar de los datos etiquetados, los valores de desviación estándar en cada fragmento de las canciones son bastante altos, lo cual puede influir en la predicción de los modelos entrenados.

Para mejorar los datos de entrenamiento, se generaron nuevos valores de arousal y valence a partir de los datos etiquetados por los evaluadores. Es menester resaltar que los valores con los que se estaba realizando el entrenamiento corresponden al promedio de los datos etiquetados de los 10 evaluadores en cada fragmento de las canciones (esa fue la estrategia que los desarrolladores de la base de datos MediaEval seleccionaron para generar los valores de arousal y valence). La alta desviación estándar presente en los datos etiquetados por los evaluadores indica una falta de consenso entre ellos, lo que sugiere que el valor promedio podría no ser confiable. A raíz del problema anterior, se opta por generar nuevos valores filtrando los fragmentos con baja variabilidad dentro de los datos etiquetados por los evaluadores y definiendo un umbral basado en experimentación para generar nuevamente el promedio, pero solo con los valores cuya desviación estándar fuera menor al de ese umbral.

8. Reentrenamiento del modelo con los nuevos valores de arousal y valence generados a partir del filtrado de fragmentos con baja variabilidad

Una vez obtenidos los nuevos valores de arousal y valence, se realiza el reentrenamiento del modelo de regresión de bosques aleatorios, ya que fue el modelo con mejores métricas de error. Es menester resaltar que en este nuevo entrenamiento se pasa de usar los 1802 datos originales del entrenamiento de canciones de MediaEval a un nuevo conjunto con 1753 datos, puesto que 49 de los datos

presentaron tal variabilidad que el algoritmo desarrollado para filtrar el nuevo dataset los descartó.

Los valores de arousal y valence predichos por el modelo reentrenado se pasan por el descriptor de error generado en este proyecto, el cual arrojó los siguientes resultados:

Métrica de error valores predichos de Arousal - Modelo Random Forest Regressor	
Valores predichos que distan entre 0,2 y 0,3 decimales	3691
Valores predichos que distan más de 0,3 decimales	2142
Valores predichos cuyo signo cambia	6254

Métrica de error valores predichos de Valence - Modelo Random Forest Regressor	
Valores predichos que distan entre 0,2 y 0,3 decimales	3351
Valores predichos que distan más de 0,3 decimales	1628
Valores predichos cuyo signo cambia	7514

Teniendo en cuenta que con los registros descartados del dataset se tiene un total de 101,674 valores de arousal y valencia predichos, la métrica de los resultados del descriptor, indica una mejora en la predicción de valores con signo contrario, ya que el 6,1% de los valores predichos de Arousal se están prediciendo con un signo contrario al real y el 7,4% para el caso de los valores predichos en valence. En cuanto a los valores que están distando entre 0,2 y 0,3 decimales, para arousal se tiene el 3,6% de los datos totales y en valence el 3,3%. Finalmente, los valores que distan más de 0,3 decimales, en el caso de arousal se tiene el 2,1% de los datos totales y en valence el 1,6% de todos sus datos. De lo anterior, se puede apreciar una mejora en la predicción de valores de signos contrarios tanto para arousal como para valence, pero no de los datos que distan entre 0,2 y 0,3 decimales, y los que distan más de 0,3 decimales, cuyos resultados demuestran una predicción un poco inferior.

9. Algoritmo de selección de canciones por similaridad emocional

Teniendo en cuenta que el modelo Random Forest Regressor entrenado inicialmente dio mejores métricas de error que el modelo reentrenado con los nuevos valores de arousal y valence, se toma la decisión de usar el modelo inicial para la generación de los valores de arousal y valence de las canciones de artistas emergentes y las del playlist popular. Con el fin de dar cumplimiento al objetivo 3 de este proyecto, se usa el algoritmo K-Nearest Neighors (KNN) para encontrar las 5 canciones más similares de la base de datos de canciones de artistas emergentes por cada canción de la playlist de canciones populares. El algoritmo KNN es adecuado para este problema puesto que puede calcular la distancia euclidiana entre los registros y selecciona los K registros más cercanos.

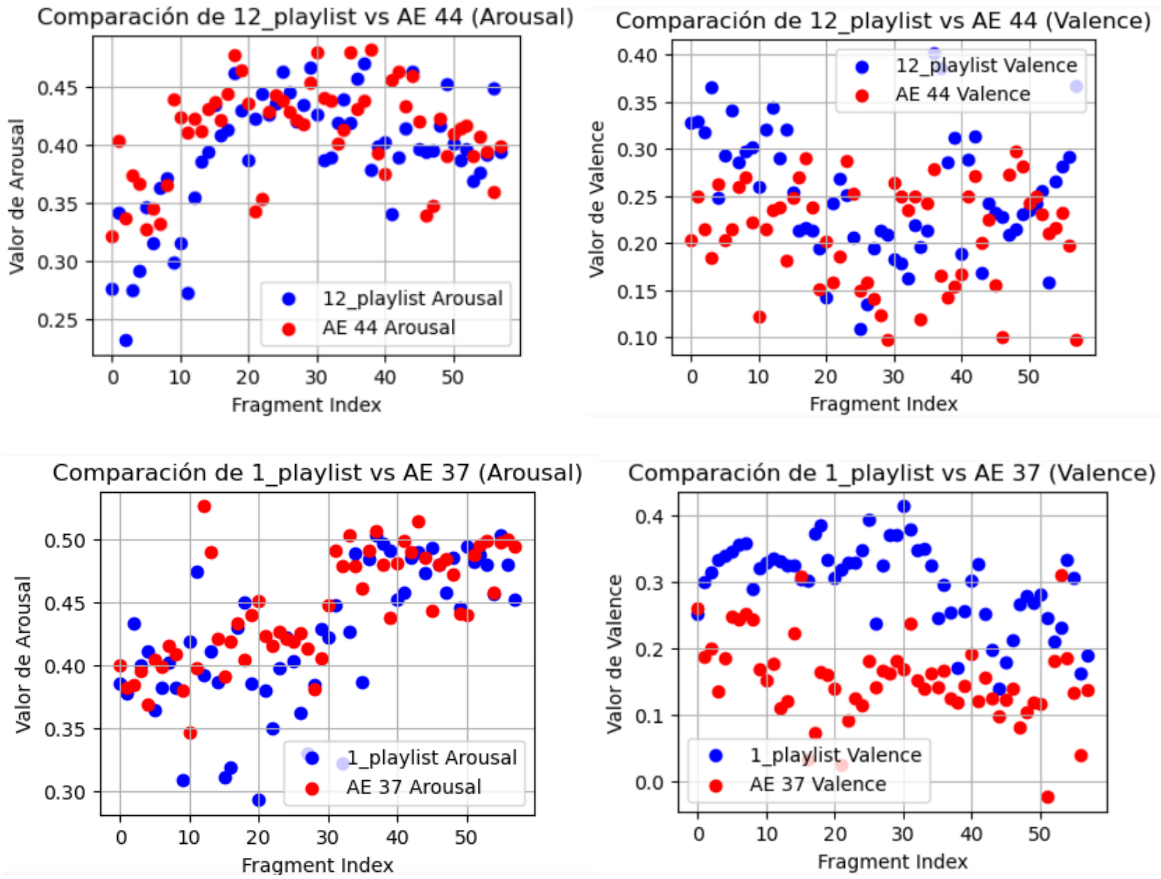

Figura 4. Se puede apreciar el diagrama de dispersión de valores de arousal y valence de dos de los resultados que arrojó el algoritmo KNN. La gráfica superior corresponde a la distribución de los valores de arousal y valence de la canción #12 de la playlist de canciones populares (*Untouched - The Veronicas*) vs la canción 44 de la base de datos de audios de artistas emergentes (*Exterminio - Unknown artist*). La gráfica inferior corresponde a la distribución de los valores de arousal y valence de la canción #1 de la playlist de canciones populares (*4Ever - The Veronicas*) vs la canción 37 de la base de datos de audios de artistas emergentes (*Depredador - Unknown artist*).

Conclusiones y trabajo futuro

En este proyecto, se utilizaron diferentes modelos de predicción para clasificar las emociones musicales mediante los valores de arousal y valence, siendo el modelo de regresión de árboles aleatorios el que mejores métricas arrojó con un 6,7% de error en cuanto a la predicción de signo contrario al real para arousal y de un 7,9% en valence. En cuanto a los valores que distan más de 0,2 decimales, en arousal solo se tiene un porcentaje de error del 3,6% y en valence un porcentaje de error de 3,5%.

En el estudio realizado, se debe tener en cuenta algunos factores que pudieron haber incidido en las métricas de error y que están relacionados con la base de datos MediaEval, la cual se empleó con fines de entrenamiento para los diferentes modelos usados en este proyecto. El primer aspecto a destacar es el número de canciones en la base de datos, ya que 1802 canciones podría ser un número relativamente pequeño en comparación con otros trabajos dentro del dominio MER. Por otro lado, la pequeña cantidad de evaluadores podría no dar una idea general y global de las emociones que evocan las diferentes canciones usadas en la base de datos, tampoco se conoce información demográfica de los evaluadores, ni información sobre sus gustos o preferencias musicales, lo cual pudo repercutir sustancialmente en los valores de arousal y valence generados, por lo que a futuro se recomienda enriquecer mejor la base de datos de entrenamiento para obtener una mejor predicción de los modelos entrenados.

Es importante mencionar que, aunque se realizó una combinación de selección de características y combinación de hiperparametros con GridSearchCV, no fue posible agregar demasiados valores para la búsqueda de hiperparametros óptimos con GridSearchCV dada la premura del proyecto. Desafortunadamente, el tiempo de entrenamiento de los modelos entrenados estaba tomando más de un día, por lo que a futuro se recomienda usar optimización bayesiana ya que es más eficiente en problemas con espacios de búsqueda grandes.

Para los proyectos relacionados con reconocimiento de emociones musicales, es importante reconocer que la percepción de las emociones es intrínsecamente subjetiva y las personas pueden percibir diferentes emociones para una misma canción, lo que va a depender de los gustos de las personas, su estado de ánimo actual o su cultura, por lo que es muy complejo llegar a un acuerdo general de clasificación emocional relacionada a la música.

Referencias bibliográficas

[1] Y. Feng, Y. Zhuang, and Y. Pan, “Music information retrieval by detecting mood via computational media aesthetics,” in *Proceedings IEEE/WIC international conference on web intelligence (WI 2003)*. IEEE, 2003, pp. 235–24

[2] Yu Xia, Fumei Xu. (2022). *Study on Music Emotion Recognition Based on the Machine Learning Model Clustering Algorithm*. Wiley Online Library, 3-4. https://doi.org/10.1155/2022/9256586

[3] Yang YH, et al., *Music Emotion Classification: A Fuzzy Approach*, Proceedings of the 14th ACM international conference on Multimedia,pp. 81-84, Oct 2006

[4] L. Lu, D. Liu, and H.-J. Zhang, *Automatic mood detection and tracking of music audio signals*, IEEE Transactions on audio, speech, and language processing, vol. 14, no. 1, pp. 5–18, 2005.

[5] Lujan Villar, (2019). *Reconocimiento de emociones musicales a través de datos y tecnologías digitales*, chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://comunicacionyhombre.com/wp-content/uploads/2020/01/ESTUDIO-2.pdf, pp 60.

[6] Van Loi Nguyen, Donglim Kim, Van Phi Ho, Younghwan Lim. (2017). *A New Recognition Method for Visualizing Music Emotion*. https://ijece.iaescore.com/index.php/IJECE/article/view/7483/6479

[7] Citron Francesca, Gray Marcus, Critchley Hugo, Weekes Brendan. (2014) *Emotional valence and arousal affect reading in an interactive way: Neuroimaging evidence for an approach-withdrawal framework,* pp 79-80.

[8] Vergara Sergio. *¿Qué es la computación afectiva y qué aplicaciones tiene en el desarrollo de software?* ITDO. Recuperado el 8 de octubre de 2024, de https://www.itdo.com/blog/que-es-la-computacion-afectiva-y-que-aplicaciones-tiene-en-el-desarrollo-de-software/

[9] Russell James A. *A circumplex Model of Affect.* Journal of Personality and Social Psychology, pp 1161-1178, Dic 1980