Universidad Nacional de la Plata

Facultad de Informática

Desarrollo de un modelo de elicitación de emociones a partir de las características de la música. Generación de un sistema recomendador.

T E S I S

para obtener el grado de

Doctor en Ciencias Informáticas

presenta

Yesid Ospitia Medina

Directora:
Sandra Baldassarri

Co-Director:
José Ramón Beltrán

Asesora:
Cecilia Sánz

La Plata, Argentina
Diciembre, 2023

"La música puede dar nombre a lo innombrable
y comunicar lo desconocido"
Leonard Bernstein

Índice general

Resumen

La computación afectiva, como área de investigación, ha logrado un importante desarrollo en los últimos años. Las actuales investigaciones han demostrado su utilidad, permitiendo medir la intensidad de las emociones a través de la tecnología, y con estos resultados implementar acciones que permitan crear beneficios para la humanidad en diferentes contextos. Uno de estos casos particulares, apuntan al estudio de la música y su relación con las emociones. Esta Tesis tiene como objetivo general el diseño de un sistema recomendador musical basado en emociones. Para abordar este objetivo se analizan las diferentes disciplinas científicas involucradas en el reconocimiento de emociones en la música. Primero, se lleva a cabo un estudio de conceptos relevantes tanto vinculados al área de identificación de emociones, con técnicas propias de computación afectiva, como en relación a las técnicas para el diseño de sistemas de predicción y de clasificación basados en *machine learning*. Al mismo tiempo, se identifican características intrínsecas y extrínsecas de piezas musicales. Se lleva adelante un proceso de revisión de la literatura sobre librerías de alto nivel, *dataset* musicales y trabajos en relación con sistemas de predicción, clasificación, y de recomendación de música. Asimismo, se revisan antecedentes en relación a sesgos que pueden intervenir en el proceso de recomendación de piezas musicales. Así, se identifica la importancia de diseñar un nuevo *dataset* musical, con algunas características novedosas, como lo son la inclusión de obras completas y originales de artistas noveles, y el etiquetado emocional sobre la variación temporal de la canción en relación con la estructura musical, para ello se utilizan series temporales. El desarrollo y diseño de este *dataset*, denominado ENSA (*Emotional Non-Superstar Artist-Dataset*), constituye uno de los aportes fundamentales de la Tesis. Posteriormente, se diseña un sistema recomendador híbrido, basado en el filtrado emocional, filtrado basado en contenido, y filtrado basado en similaridad. La propuesta hace uso de un modelo de *machine learning* para el reconocimiento de emociones a través de un sistema de etiquetado dimensional de *valence* y *arousal*. El nuevo *dataset* y las estrategias de agrupamiento por similaridad implementadas por el sistema recomendador, como otras medidas adicionales que se detallan en el desarrollo de esta Tesis, también permite dar un tratamiento al efecto de la popularidad, que generalmente aparece a través de un sesgo preexistente, que no solo afecta a oyentes, sino que también tiene un alto impacto en los artistas y sus posibilidades de desarrollo en la industria de la música. Estos últimos aportes se alinean con los objetivos propuestos en la Tesis y que se han alcanzado en su totalidad.

En cuanto al trabajo futuro, se resalta la importancia de continuar profundizando en las siguientes tres líneas: el *dataset ENSA*, el sistema recomendador, y el estudio de sesgos. Con respecto al *dataset*, es importante seguir avanzando en la inclusión de nuevas canciones, como también en la definición e inclusión de nuevas etiquetas que permitan llevar a cabo experimentos novedosos. En cuanto al sistema recomendador, se propone extender su funcionalidad incluyendo otras partes de la estructura musical (coro, puentes, solos) en las estrategias de recomendación. Y, finalmente, en relación a los sesgos, se propone identificar y analizar la aparición de nuevos sesgos (aquellos que clasifican como emergentes), y tanto desde el *dataset ENSA*, como desde las estrategias de recomendación, proponer nuevos tratamientos que permitan mitigar sus efectos.

Agradecimientos y dedicatoria

Agradecimientos

A Sandra, por darme la oportunidad de trabajar a su lado, y por estar siempre presente. A José Ramón, por compartirme su conocimiento, y por ese enorme aprecio que juntos tenemos por la música. A Cecilia, por todo su apoyo.

A mis amigos artistas, por participar en los experimentos de esta Tesis a través de sus obras musicales y entrevistas.

Dedicatoria
A la memoria de mis abuelos, Rosa y Jaime.
Y a mi madre, Amparito Medina.

Índice de tablas

Índice de figuras

Índice de abreviaturas

AAGPR *Adaptive aggregation of gaussian process regressors.*

API *Application Programming Interface.*

ARD *Averaged random distance.*

BLSTM- RNNs *Bi-directional long short-term memory recurrent neural networks.*

BPM *Beats per Minute.*

CF *Collaborative filter.*

CFNNS *Coin-Flipping Neural Networks.*

CHI *Conference on Human Factors in Computing Systems.*

CSV *Comma-separated values.*

dB *Decibeles.*

DF *Demographic filtering.*

DTW *Dynamic time warping.*

EBF *Emotion-based filtering.*

ENSA *Emotional Non-Superstar Artist-Dataset.*

FKNN *Fuzzy k-NNN.*

FNM *Fuzzy Nearest-Mean.*

GMER *General emotional classification of music.*

HCI *Human-Computer Interaction.*

HF *Hybrid filtering.*

HL *Hidden layers.*

IB *Based on interaction.*

IFPI *International Federation of the Phonographic Industry.*

ISMIR *International Society for Music Information Retrieval.*

LR *Learning rate.*

LSTM-RNN *Deep long-short term memory recurrent neural networks.*

MAE *Mean absolute error.*

MD *Metadata.*

MER *Music emotion recognition.*

MIR *Music information retrieval.*

MLP *Multi-layer perceptron.*

MRS *Sistemas recomendadores musicales.*

MTG *Music Technology Group.*

OpenSMILE *Open-source Speech and Music Interpretation by Large-space Extraction.*

PA *Personalized approach.*

PB *Based on popularity.*

PCA *Principal component analysis.*

PLB *Based on playlists.*

PMER *Personal emotional classification of music.*

R^2 *Determination coefficient.*

RFC *Random forest classifier.*

RFT *Random forest tree.*

RMSE *Mean square error.*

SAM *Self-Assessment Manikin.*

SB *Based on similarity.*

SVM *Support vector machine.*

SVR *Support vector regression.*

TUM *Technical University Munich.*

UC *User context.*

UX *User experience.*

V/A *valence and arousal.*

Capítulo 1

Introducción

En este capítulo se describe el tema de investigación que se desarrolla a lo largo de este documento de Tesis. Se presenta el contexto en el cual se encuentra inmersa la temática (Apartado 1.1), las motivaciones (Apartado 1.2), los objetivos (Apartado 1.3), las preguntas que orientan la investigación (Apartado 1.4), la metodología adoptada (Apartado 1.5), y la estructura de este documento (Apartado 1.6). Por último, se explicitan las decisiones adoptadas en torno al uso del masculino inclusivo en este trabajo (Apartado 1.7).

1.1 Contexto

La industria de la música tiene un importante aporte a la economía del mundo, y a lo largo del tiempo su modalidad de consumo ha ido presentando importantes cambios, siendo uno de ellos, la transición del consumo de medios físicos al consumo de servicios digitales alojados en la nube. En un reporte de consumo de música de 2022 generado por la IFPI (*International Federation of the Phonographic Industry*) [1] se resalta el hecho de que las ganancias por consumo de música a través de servicios de *streaming* han crecido en un 67 % con respecto al año anterior. Asimismo, en este informe se resalta la enorme dificultad para desarrollar nuevos talentos en un mercado musical global en el que cada día se suben a la red numerosas canciones para competir por la atención de los oyentes.

En la Figura 1.1 se puede observar la distribución de las ganancias en miles de millones de dólares (MMD) por diferentes criterios: medios físicos, servicios de *streaming*, descargas, derechos de autor, y servicios de *synchronisation* (el uso de música grabada en la publicidad, el cine, los vídeo juegos y la televisión). De esta gráfica es importante resaltar que a lo largo de la línea del tiempo los ingresos por medios físicos han disminuido desde 22.3 MMD en 1999 hasta 4.6 MMD en 2022. Mientras que para el caso de servicios de *streaming*, los ingresos han aumentado desde 0.1 MMD en 2005 hasta 17.5 MMD en 2022. De esta manera, se observa que el consumo de música consolida todo el desarrollo de una industria, en donde se evidencia una importante tendencia a utilizar herramientas computacionales soportadas en servicios de *streaming*.

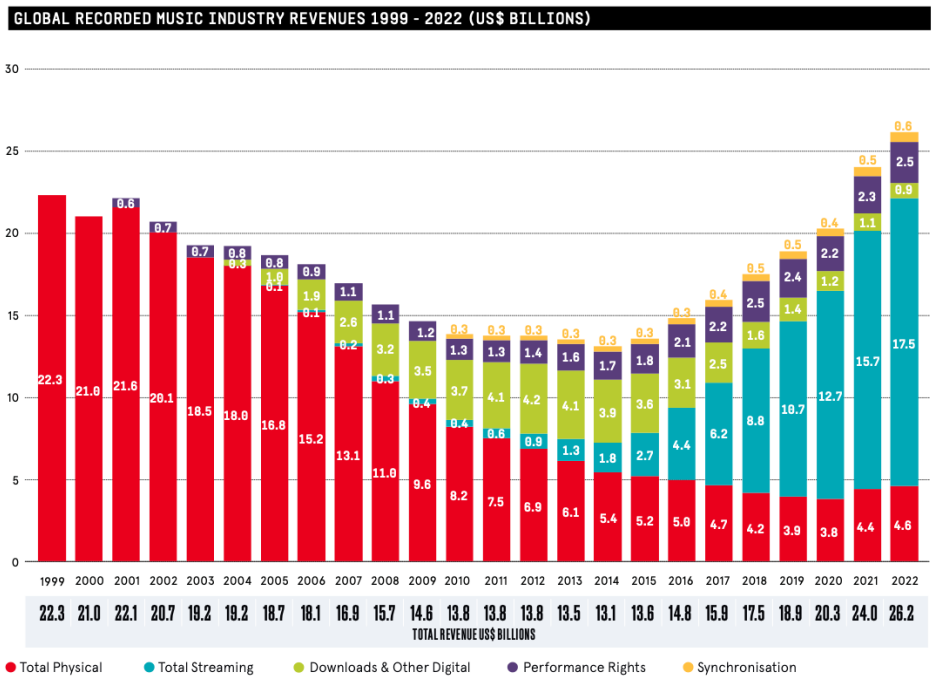


Figura 1.1: Ingresos de la industria mundial de la música grabada entre 1999 y 2022 (Miles de millones de dólares) [1].

Los sistemas recomendadores musicales son implementados en la actualidad por la gran mayoría de las plataformas de reproducción musical, y tienen como principal objetivo facilitar la búsqueda y el descubrimiento de nuevas canciones. En general, las estrategias más utilizadas consideran el perfil de usuario, el historial de reproducciones, el contexto, y la popularidad de las piezas musicales y de los artistas. Esta Tesis pretende ir un poco más allá, teniendo en cuenta la percepción emocional del oyente en las estrategias de recomendación. Para ello es importante entender la relación que tiene la música con la apreciación del oyente, y cómo finalmente el oyente considera sus preferencias sobre una canción. Existen diversos estudios que sugieren una relación importante entre la música y las emociones [8] [9], siendo las emociones consideradas como una reacción psicofisiológica, que tiene su origen a través de un proceso de evaluación mental, y que da lugar a experiencias asociadas con el bienestar o el malestar [2]. Desde el punto de vista computacional, esta Tesis busca comprender la complejidad de las herramientas existentes para caracterizar la música, como también la identificación de estrategias que permitan relacionar estados emocionales de los oyentes con características musicales, como el tempo, el tono y el modo [8]. Todo esto con el principal objetivo de plantear el diseño de un sistema recomendador musical novedoso que tenga en cuenta las emociones del usuario, y que además, considere el tratamiento de los sesgos, en particular aquellos relacionados con el efecto de la popularidad sobre artistas noveles.

1.2 Motivación

Los factores que motivan la investigación son los siguientes:

- Los modelos computacionales relacionados con el estudio de la música desde la computación afectiva son pocos, y se encuentran en una etapa temprana de desarrollo.
- Desde el punto de vista del consumo de la música, las herramientas de computación afectiva podrían en algún momento generar indicadores importantes que permitan comprender la percepción emocional de una persona (o grupos de personas) al escuchar ciertos tipos de piezas musicales.
- Existen notables desarrollos en el campo del reconocimiento de emociones en la música (*Music Emotion Recognition (MER)*). Sin embargo, la precisión del proceso de reconocimiento de emociones y sus aplicaciones reales tienen diversas limitaciones. Muchas de estas limitaciones se encuentran relacionadas con los procesos de etiquetado emocional, como también, con el diseño de algoritmos de recomendación, en particular, aquellos que se plantean bajo el enfoque de aprendizaje de máquina (*machine learning*). El aprendizaje de máquina se suele referir en idioma inglés, incluso en bibliografía en español, por lo que en adelante, en este documento se utilizará su denominación en inglés.
- A pesar de que en la actualidad los artistas pueden producir y distribuir con gran facilidad su música a través de diversas plataformas musicales, la posibilidad de crecer en la industria de la música es realmente difícil, en especial para los artistas noveles, debido a que muchas de las estrategias de recomendación de estas plataformas se basan en indicadores de popularidad.
- Aunque existen algunos *datasets* musicales para experimentar en el campo de sistemas recomendadores musicales, el acceso a dichos *datasets* es limitado, así como también es limitada la diversidad de características disponibles en estos *datasets* para la realización de experimentos novedosos.
- La efectividad de los sistemas recomendadores musicales depende mayormente de los desarrollos obtenidos en los sistemas de MER. En la medida que los recomendadores de piezas musicales evolucionen a mejores resultados, su aplicabilidad generará mayor credibilidad por parte de los usuarios finales (oyentes y artistas), como también una mejor experiencia de uso.

1.3 Objetivos

El objetivo general de la Tesis consiste en diseñar un sistema recomendador de piezas musicales, a partir de la relación entre las características intrínsecas de la música y las emociones percibidas por el oyente. Los objetivos específicos se indican a continuación:

- **Objetivo 1:** Determinar el estado actual de la computación afectiva en cuanto a la medición y reconocimiento de emociones a partir de la estimulación musical.
- **Objetivo 2:** Estudiar las características intrínsecas de la música.
- **Objetivo 3:** Determinar cuáles son las estrategias más efectivas para medir y reconocer emociones en la música, realizando experimentos que permitan establecer la adecuación necesaria de estas estrategias.
- **Objetivo 4:** Construir un modelo que permita establecer una relación entre las características intrínsecas de la música y las emociones percibidas por el oyente.
- **Objetivo 5:** Implementar un prototipo para la clasificación emocional de la música en base a sus características intrínsecas.
- **Objetivo 6:** Estudiar los diferentes tipos de sesgos existentes en las estrategias de recomendación, y proponer algunas medidas para su tratamiento.
- **Objetivo 7:** Implementar un experimento para el reconocimiento de emociones durante un proceso de apreciación musical
- **Objetivo 8:** Diseñar e implementar un nuevo *dataset* de piezas musicales de artistas noveles.
- **Objetivo 9:** Desarrollar un prototipo de sistema recomendador basado en la relación entre las propiedades intrínsecas de la música y las emociones percibidas por el oyente.

1.4 Preguntas que orientan la investigación

A continuación se presentan las preguntas que permiten orientar la investigación:

- **P1:** ¿Cómo se puede representar computacionalmente la relación entre la música y las emociones?
- **P2:**¿Puede la música mejorar la experiencia de un usuario o provocar emociones específicas?
- **P3:** ¿Cuáles son las técnicas de computación afectiva más apropiadas para determinar qué es lo que realmente siente un oyente?
- **P4:** ¿Qué *frameworks* existen actualmente para el reconocimiento de emociones en la música?
- **P5:** ¿Puede diseñarse computacionalmente un recomendador de piezas musicales basado en la percepción emocional de los usuarios?

Tabla 1.1: Áreas de investigación y palabras claves

Área	Palabras claves
Computación afectiva	Inteligencia emocional, emociones, estados de ánimo, detección de emociones, modelos afectivos.
Música	Música, composición musical, escritura de música, características de sonido.
Reconocimiento de emociones en la música	MER , reconocimiento de emociones en la música, reconocimiento de emociones en el sonido, análisis del sonido, reconocimiento de género en la música, características estándar de sonidos, características melódicas, estados de ánimo en la música, recuperación de información musical, librerías para la recuperación de información musical, sistemas de predicción en MER, sistemas de clasificación en MER, conjunto de datos musical *musical dataset*, sistemas difusos (*fuzzy*) MER, sistemas determinísticos MER, sistemas no determinísticos MER.
Sistemas recomendadores musicales	Sistemas recomendadores, sistemas recomendadores musicales, estrategias de recomendación, sesgos.

1.5 Metodología de la investigación

Este trabajo considera una metodología de investigación que incluye las siguientes fases:

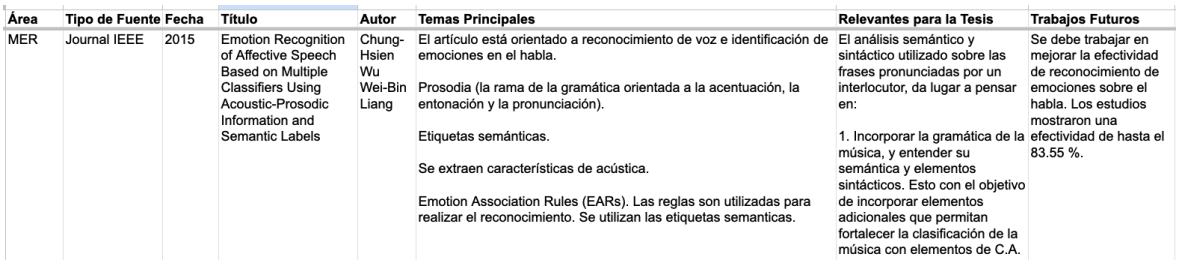
- La presentación de una base conceptual que soporta el marco teórico de la Tesis. En esta base conceptual se abordan las emociones y su relación con la música, la caracterización de emociones a través de modelos emocionales, y las aplicaciones de los sistemas recomendadores musicales basados en emociones (Capítulo 2). Para esta fase se consideran inicialmente referencias bibliográficas aportadas por los directores de la Tesis, expertos en el área consultados, material de los cursos tomados como parte del proceso de doctorado y bibliografía ampliamente citada en los temas vinculados. Luego, también se nutre de la revisión sistemática de literatura que se menciona a continuación.

- Una revisión sistemática de literatura [10] del estado del arte que se lleva a cabo de la siguiente manera:

  1. Se realiza una búsqueda a través de las áreas y palabras claves definidas en la Tabla 1.1. A lo largo del desarrollo de la Tesis para el abordaje de los principales temas de investigación, se considera trabajos que hayan sido publicados a partir del año 2013 en inglés y español. La selección de los trabajos considera atender a los diferentes objetivos propuestos y las preguntas de investigación planteadas. Para ello se tuvieron en cuenta principalmente las fuentes bibliográficas que se indican en la Tabla 1.2. El análisis de los antecedentes encontrados se presenta de manera integrada en el capítulo 3.

Tabla 1.2: Recursos bibliográficos

Tipo de recurso	Recursos específicos
Bases de datos	ACM Digital Library, IEEE Xplore, Elsevier Scopus, Springer Link, Google Scholar, ResearchGate.
Revistas	IEEE transactions on affective computing, Springer multimedia tools and applications, International Journal of Human-Computer Studies. Personal and Ubiquitous Computing (PUC).
Congresos relevantes	Jornadas Iberoamericanas de Interacción Humano-Computadora. Audio Mostly, International Society for Music Information Retrieval (ISMIR) Proceedings, ACM Conference on Human Factors in Computing Systems (CHI). International Conference on Affective Computing and Intelligent Interaction and Workshops, Affective Computing and Intelligent Interaction (ACII).



Área	Tipo de Fuente	Fecha	Título	Autor	Temas Principales	Relevantes para la Tesis	Trabajos Futuros
MER	Journal IEEE	2015	Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels	Chung-Hsien Wu Wei-Bin Liang	El artículo está orientado a reconocimiento de voz e identificación de emociones en el habla. Prosodia (la rama de la gramática orientada a la acentuación, la entonación y la pronunciación). Etiquetas semánticas. Se extraen características de acústica. Emotion Association Rules (EARs). Las reglas son utilizadas para realizar el reconocimiento. Se utilizan las etiquetas semanticas.	El análisis semántico y sintáctico utilizado sobre las frases pronunciadas por un interlocutor, da lugar a pensar en: 1. Incorporar la gramática de la música, y entender su semántica y elementos sintácticos. Esto con el objetivo de incorporar elementos adicionales que permitan fortalecer la clasificación de la música con elementos de C.A.	Se debe trabajar en mejorar la efectividad de reconocimiento de emociones sobre el habla. Los estudios mostraron una efectividad de hasta el 83.55 %.

Figura 1.2: Tabla de resumen para la consolidación de revisión sistemática. Elaboración propia a partir de Captura de pantalla.

2. Se realiza la lectura del *abstract* de cada artículo científico y se clasifica por área/tema de estudio.
3. Se selecciona o descarta cada artículo considerando su relación y aporte a los objetivos específicos de la Tesis.
4. Se realiza la lectura completa del artículo, generando apuntes a través de la herramienta *Mendeley*[1].
5. Se organizan e ingresan las notas anteriores en una tabla resumen en la que se resalta los siguientes atributos: el área, el tipo de fuente, la fecha de publicación, el título, los autores, los temas principales, los apuntes relevantes para la Tesis, y los trabajos futuros (y limitaciones) planteados por los autores. En la Figura 1.2 se muestra una captura de pantalla de esta tabla.

- El desarrollo de un sistema recomendador de piezas musicales a partir de las diferentes limitaciones encontradas en la revisión del estado del arte. Para el desarrollo del sistema recomendador se trabaja con prototipos evolutivos, además se llevan a cabo reuniones periódicas con los directores de la Tesis para discutir

[1]https://www.mendeley.com

sobre las barreras encontradas y los avances siguiendo una metodología ágil [11]. También, se realizan consultas a expertos del área de música y se trabaja con compositores noveles. Los prototipos evolucionan a partir de experimentos donde se miden las tasas de acierto y error. Cada experimento se presenta de manera detallada en los capítulos 4, 5 y 6 de este informe. Siguiendo esta metodología se diseñan y crean de manera progresiva los siguientes productos que fueron encontrándose como necesidades para abordar el sistema recomendador planteado en los objetivos y se conformaron en aportes de esta Tesis:

  - Un sistema para la predicción de emociones basado en un modelo dimensional afectivo a partir de características de sonido de piezas musicales.
  - Un sistema para la clasificación emocional no determinístico a partir de características de sonido de piezas musicales.
  - Un sistema para la clasificación emocional determinístico a partir de características de sonido de piezas musicales.
  - Un nuevo *dataset* musical con canciones originales de artistas noveles, que además incluye el etiquetado de varias características por parte de oyentes y artistas.
  - Un sistema recomendador de piezas musicales basado en el reconocimiento de emociones dinámico sobre la estructura de las canciones, en particular sobre el verso, y que implementa técnicas de agrupamiento por similaridad entre series temporales.

- La elaboración de conclusiones y planteo de líneas de trabajo futuro y limitaciones del estudio, se lleva a cabo a partir de contrastar los objetivos planteados y los resultados alcanzados en las diferentes etapas del proceso de Tesis y un proceso de análisis del recorrido realizado.

1.6 Estructura de la Tesis

Este documento se encuentra organizado en 7 capítulos. A su vez, cada capítulo internamente se organiza por apartados.

El **Capítulo 1** presenta de manera general el tema de investigación, detallando la motivación, los objetivos, las preguntas de investigación, la metodología, y la estructura que sigue todo el documento de la Tesis.

El **Capítulo 2** realiza una presentación de la fundamentación teórica que es importante para la lectura de esta Tesis; en esta base conceptual se introducen las emociones, la música, la medición de emociones a través de modelos afectivos computacionales, los analizadores de contenido, las librerías de alto nivel para el reconocimiento de emociones en la música, la predicción de valores con *machine learning*, la clasificación no determinística con sistemas *fuzzy*, la clasificación determinística con *machine learning*, los sistemas recomendadores, y los sesgos.

El **Capítulo 3** contiene la revisión del estado del arte, en donde se identifican y analizan diversos trabajos relacionados con las temáticas de interés de esta Tesis. Este capítulo cierra analizando los principales hallazgos encontrados en cuanto a problemas sin resolver, limitaciones, y líneas de investigación propuestas para desarrollar a futuro. Estos hallazgos y su discusión son los que orientan los aportes de esta Tesis en los siguientes capítulos.

Los **Capítulos 4, 5 y 6** presentan las contribuciones de esta Tesis. En el **Capítulo 4** se presenta el diseño de tres prototipos que permiten reconocer emociones en piezas musicales desde los siguientes enfoques: predicción, clasificación no determinística, y clasificación determinística. El **Capítulo 5** presenta el diseño del *dataset* musical *ENSA-Dataset* conformado por canciones originales de artistas noveles; adicionalmente, también se incluye un análisis del *dataset* considerando las etiquetas disponibles. El **Capítulo 6** presenta el diseño de un prototipo de sistema recomendador de piezas musicales basado en la estrategia de filtrado por similaridad; el sistema implementa técnicas de *machine learning* no supervisado para agrupar piezas musicales a partir de la variación temporal del *valence* y *arousal*.

El **Capítulo 7** presenta las conclusiones del trabajo, las implicaciones prácticas, los aportes con las producciones científicas de respaldo, las limitaciones del estudio, y las futuras líneas de trabajo.

Este documento además cuenta con 4 anexos. En el anexo A. Fundamentos de teoría musical se presentan algunos de los conceptos más relevantes dentro del campo de la música. En el anexo B. Experimento con *AcousticBrainz* se presenta el proceso de extracción de las características de tonalidad, modo, tempo, y emociones predominantes sobre una canción en particular. En el anexo C. Cuadro comparativo de librerías de alto nivel se presenta una comparación entre las librerías de alto nivel estudiadas en el apartado 3.1. Finalmente, el anexo D. Análisis del sistema de etiquetado en MediaEval presenta la revisión de algunos casos puntuales de etiquetado emocional sobre el *dataset* de MediaEval estudiado en el apartado 3.3.

1.7 Sobre el uso del masculino gramatical inclusivo

En esta Tesis se utiliza el masculino gramatical como término no marcado de la oposición de género, que puede referirse a grupos formados de varones y mujeres y, en contextos genéricos o inespecíficos, a personas de uno u otro sexo. En este sentido, se advierte que no funciona como una extensión del masculino con significado de sexo para denotar a todas las identidades sexogenéricas y autopercibidas por las personas. Sino como una consecuencia natural de la ausencia de sexo semántico en las entradas léxicas de los términos masculinos.

Capítulo 2

Fundamentación teórica

Este trabajo de Tesis aborda una línea de investigación multidisciplinar, por lo que a través de este capítulo se suministra una fundamentación teórica que permite identificar las disciplinas involucradas, las relaciones entre ellas, y los conceptos más relevantes. Inicialmente, en el Apartado 2.1, se introducen las emociones y su relación con la música desde el campo de la Psicología. En el Apartado 2.2 se presenta la computación afectiva como área de investigación, y se destaca su importancia para la caracterización de emociones y la posibilidad de generar mediciones sobre ellas. También se introducen algunas características particulares de la música, y su relación con la percepción emocional. El Apartado 2.3 presenta estudios relacionados con la identificación y medición de emociones percibidas por el oyente durante una apreciación musical. En el Apartado 2.4 se presentan los analizadores de contenido y sus principales características. El apartado 2.5 presenta las librerías de alto nivel para el reconocimiento de emociones en la música. En el Apartado 2.6 se introducen los principales conceptos teóricos relacionados con el diseño y análisis de sistemas de predicción de valores a través de *machine learning*. En el Apartado 2.7 se destacan las principales caracteristicas de un sistema de clasificación no determinístico (*fuzzy*). El Apartado 2.8 presenta los principales conceptos teóricos relacionados con el diseño y análisis de sistemas de clasificación determinísticos a través de *machine learning*. En el Apartado 2.9 se comenta la importancia de las emociones para definir y desarrollar sistemas recomendadores, en particular con referencia a la industria musical. Asimismo se plantean diferentes aplicaciones novedosas en este campo. El Apartado 2.10 describe los tipos de sesgos existentes, así como su impacto en el funcionamiento de los sistemas recomendadores musicales. Finalmente, en el Apartado 2.11 se presentan las conclusiones del capítulo. En relación a la metodología utilizada para conformar este capítulo se aborda como se indica en el capítulo 1, una revisión conceptual en base a bibliografía sugerida y a los resultados de la revisión sistemática de literatura, fundamentalmente a partir de las búsquedas con cadenas conformadas por las palabras claves de las filas 1 y 2 de la Tabla 1.1.

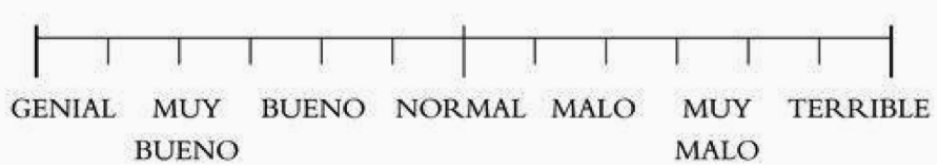


Figura 2.1: Escala de evaluación de los hechos [2]

.

2.1 Las emociones en la música

Según Rafael Santrandreu, lo que realmente nos afecta es lo que interpretamos acerca de nuestras experiencias, y no las experiencias en sí mismas [2]. Los efectos emocionales como la depresión, la ansiedad y la ira, son el resultado de la interpretación que le damos a los hechos externos que ocurren en nuestras vidas. Esta interpretación es única para cada persona, porque depende en esencia de su criterio de evaluación y calificación. La línea de evaluación de las cosas de la vida, es uno de los planteamientos de Santrandreu [2] para explicar la tendencia del ser humano a generar juicios sobre su entorno. Los seres humanos generalmente, tienden a calificar en todo momento lo que les ocurre. El resultado de esta calificación es ubicado sobre una escala, como la que se puede observar en la Figura 2.1. El proceso de evaluación depende de los pensamientos generados por cada individuo, y su resultado tiene una afectación sobre el estado de ánimo, como también sobre su bienestar y salud.

La ciencia empieza a mostrar que la inteligencia emocional es buena para la salud de las personas [2]. Por esta razón, es importante buscar mecanismos para controlar las tensiones que se pueden presentar; ejerciendo un control sobre las reacciones generadas a partir de hechos particulares. Del buen manejo emocional, depende en gran medida la misma expectativa de vida del ser humano. La gestión de las emociones, por lo tanto, permite mejorar la calidad de vida de una persona, y en general facilitar diversos procesos del mundo. Ante esta premisa inicial, se puede plantear la revisión de estrategias, que permitan impactar positivamente nuestra realidad, a través de una interacción con las emociones del ser humano.

Una de las formas que tienen los seres humanos para mejorar su estado de ánimo es a través de la música. La música es considerada un lenguaje, y se requiere entender sus principios fundamentales, para abordar su relación con las emociones. La música como lenguaje tiene propiedades que requieren de gramáticas de cierta complejidad [9]. Esta gramática es representada a través de escritura, y el músico en su rol de intérprete tiene la responsabilidad de leer e interpretar una pieza musical con exactitud; y quizás lo más complejo, transmitir la intención emocional del compositor. De esta manera, la música tiene una estructura que nos permite compararla con un lenguaje. Pero, tanto el compositor como el intérprete y el oyente, necesitan de recursos psicológicos para representar, decodificar y entender las gramáticas de la música.

La música tiene la capacidad de influir en nuestras emociones, y con ello, generar modificaciones en nuestro estado de ánimo [9] [12]. La gran mayoría de las personas han experimentado esa sensación de sentir cosas diferentes cada vez que escuchan una pieza musical [8]. Existen canciones que traen recuerdos de hermosos momentos, como

también de situaciones difíciles. Algunas canciones son escuchadas en momentos de debilidad, porque fortalecen emocionalmente al oyente. Algunos autores como Ian Cross y Elizabeth Tolbert plantean que los sonidos de la música tienen un significado otorgado a través de procesos cognitivos, que en muchos casos relaciona la intención expresiva del compositor con la respuesta emocional del oyente. [13]. Esta conexión entre la música y lo que se siente al escucharla, sugiere la capacidad que tienen las notas musicales para evocar emociones en las personas. Quizás por esto, Ludwig Van Beethoven alguna vez comparó la música con el lenguaje de Dios. Y Cyril Scott, comparó la melodía con el grito del hombre a Dios, y la armonía como la respuesta de Dios al hombre [9].

Si la música realmente tiene efecto sobre las emociones humanas, es posible considerarla como una herramienta de gestión emocional. Por medio de la música se podría facilitar un cambio emocional, como también generar unas condiciones que favorezcan la calidad de vida en un momento determinado.

2.2 Medida y caracterización de emociones en la música

La computación afectiva, como área de investigación, ha logrado un importante desarrollo en los últimos años. Las actuales investigaciones han demostrado su utilidad, permitiendo medir la intensidad de las emociones a través de la tecnología, y con estos resultados implementar acciones que permitan crear beneficios para la humanidad en diferentes contextos [14]. Uno de estos casos particulares, apuntan al estudio de la música y su relación con las emociones. Para ello, es necesario entender la música como fenómeno físico, y también como arte, para luego determinar su connotación emocional y su relación con la percepción del oyente.

En la música las notas musicales se constituyen como el elemento de menor granularidad. Cada nota desde una perspectiva física tiene una frecuencia de onda asignada (su frecuencia fundamental). La armonía es una sucesión de acordes que define en gran medida la intención emocional de una canción [9]. En la música occidental se establecen 2 modalidades fundamentales: modo mayor y modo menor. La diferencia desde el punto de vista físico está en la distancia en términos de la frecuencia que se encuentran entre las diferentes ondas de sonido generadas por cada nota. Esa característica física conecta con una característica emocional, y es que la música compuesta en modalidades mayores tiende a transmitir más felicidad que la compuesta en modalidades menores [8]. La modulación que consiste en cambiar de un tono a otro, y que suele utilizarse para generar transiciones dentro de la estructura de la canción como por ejemplo entre el verso y un coro, también es una estrategia de composición muy utilizada por los músicos, y su aplicación normalmente genera un cambio en la percepción emocional por parte del oyente [9]. Para mayor información sobre conceptos y definiciones de teoría musical se incluye el anexo A. Fundamentos de teoría musical.

Desde la Ciencias de la Computación se plantea el reto de la clasificación emocional de la música, iniciando con la reconstrucción de algunas características de alto nivel, que luego serán asociadas con estados emocionales [15]. La intensidad de la señal, el ritmo, el timbre, el modo y la tonalidad son algunas de las propiedades intrínsecas de la música, consideradas como relevantes durante el proceso de reconocimiento y clasificación.

Existen en la actualidad algunas herramientas o *frameworks* que se encuentran enmarcadas dentro de dos disciplinas: la recuperación de información musical (MIR), y el reconocimiento de emociones en la música (MER). En el caso de MIR se aborda el interés por estudiar y desarrollar herramientas que permitan reconstruir información musical a partir técnicas asociadas con el procesamiento de señales en archivos digitales de sonido, mientras que MER se encuentra orientado a la identificación de emociones sobre la música a partir de características musicales, generalmente reconstruidas a partir de MIR. El principio de funcionamiento de estos *frameworks* se centra en extraer un conjunto de propiedades intrínsecas analizando el sonido mediante diferentes técnicas; para posteriormente abordar la revisión de estas propiedades y determinar su clasificación emocional, teniendo en cuenta también la percepción emocional del oyente. La idea general de un sistema MER se concentra en establecer una relación entre determinadas emociones y ciertos valores de características de sonido particulares. Aunque se tienen avances con respecto a estos *frameworks*, el proceso de clasificación emocional aún tiene problemas de precisión, por la complejidad que se requiere para ejecutar este análisis [16]. Normalmente se realizan comparativas sobre estas herramientas para determinar la más apropiada según un criterio de evaluación previamente definido [17]. A continuación se listan y describen algunos de estos *frameworks*: Marsyas, MIR toolbox, y PsySound 3 [18].

- MARSYAS es una herramienta de análisis de sonido, que puede ser utilizada por recomendadores de música para la generación automática de listas de reproducción de canciones [19]. Estas listas de reproducción se generan a partir del estado emocional expresado por el oyente. La clasificación emocional de las canciones se realiza a partir del análisis de una serie de propiedades de sonido seleccionadas tales como: intensidad, tono, ritmo, timbre y tonalidad [17].

- MIR toolbox es un *framework* implementado a través de la integración de diferentes funciones codificadas en MATLAB, que permite reconstruir características musicales, como es el caso del timbre y la tonalidad, a partir de un sonido [20]. Una vez que se extrae esta información, se puede proceder a diseñar programas encargados de interpretar y clasificar el sonido [18].

- PsySound 3 analiza algunas características de sonidos como son el ritmo y la fluctuación de la intensidad de la señal [21]. Adicionalmente, se cuestiona la lentitud con la que responde durante el procesamiento de la señal, lo que afecta los

tiempos de respuesta de un sistema de clasificación de emociones en la música [16].

Aunque los sistemas MER definen algoritmos para la extracción de información del sonido, e implementan modelos para clasificar emocionalmente la música a través de sus propiedades intrínsecas; es importante obtener una retroalimentación (*feedback*) por parte del oyente, a través de otros medios, algo que no se considera en los *frameworks* mencionados anteriormente. Básicamente, con el objetivo de establecer un comparativo entre los reconocimientos obtenidos por un determinado *framework* / herramienta y la percepción emocional real del oyente. De aquí la importancia de analizar modelos emocionales y herramientas para medir esta percepción; lo que se aborda en mayor detalle en el siguiente apartado.

2.3 Modelos emocionales y percepción

La emoción es un sentimiento consciente y subjetivo que los seres humanos perciben cuando se enfrentan a ciertos estímulos. Las emociones también pueden describirse como esa respuesta que tienen las personas frente a diversos cambios físicos y psicológicos de sus entornos. Es importante resaltar que las emociones en las personas son sumamente complejas de estudiar, debido a que diversas variables determinan diferentes emociones en personas distintas [22]. En general, las emociones pueden clasificarse en negativas (como ansiedad, frustración, etc.), positivas (como felicidad, calma, etc.) o neutras [22]. La detección y clasificación de las emociones percibidas por parte de una persona se suele obtener utilizando diferentes métodos de evaluación de la experiencia de usuario (UX). Algunos de estos métodos pueden estar apoyados en diferentes tecnologías que permiten reconocer las emociones a partir de diferentes canales, visibles y no visibles, como es la captura de datos provenientes de sensores fisiológicos, reconocimiento facial con cámaras, reconocimiento de emociones en la voz (*voice recognition*) y biosensores [23] [24] [25]. En otros casos, los métodos pueden involucrar de manera más explícita al usuario mediante auto-evaluación, ya sea con cuestionarios y/o entrevistas en donde los usuarios expresan lo que sienten en un momento determinado [26] [27] [28]. Por ejemplo, un método muy utilizado es el uso de tarjetas emotivas (*emocards*) [3], las cuales le permite a un usuario expresar la emoción percibida a partir de una determinada experiencia. En la Figura 2.2 se puede observar una propuesta de *Emocards*. Otros ejemplos interesantes y ampliamente utilizados son las técnicas de evaluación pictórica no verbales conocidas como *Self-Assessment Manikin* (SAM) que permite indicar el nivel de placer, la excitación y la dominancia asociados a la reacción afectiva de una persona ante un estímulo[29], o *Pick-A-Mood* (PAM) que permite seleccionar entre 8 emociones posibles, más la neutral [30].

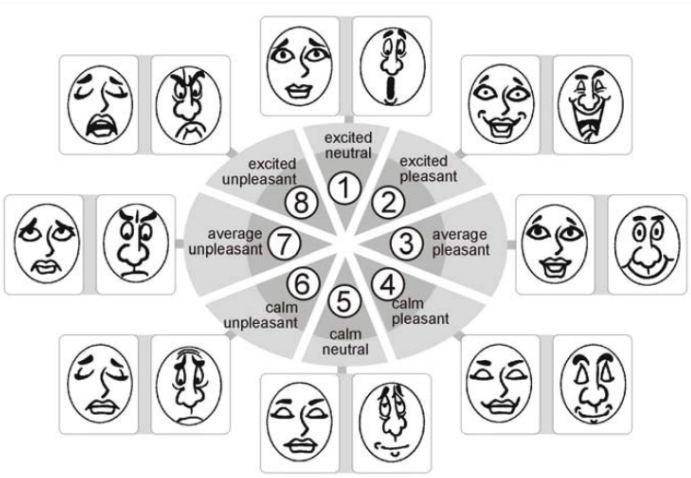


Figura 2.2: Tarjetas (*Emocards*) asociadas a categorías emocionales [3]

Para la representación de las emociones existen diferentes modelos, siendo los modelos categóricos y los dimensionales algunos de los más utilizados [31] [32]. Los modelos categóricos son más simples de utilizar, ya que el usuario debe elegir etiquetas discretas basadas en emoticones o adjetivos relativos a emociones. Las categorías emocionales básicas más conocidas son las seis emociones universales propuestas por Ekman [33], que incluyen la alegría, la tristeza, el miedo, la ira, el disgusto y la sorpresa. Aunque el modelo categórico resulta ser muy intuitivo y se ajusta con relativa facilidad a la experiencia de usuario, es importa tener en cuenta que las listas discretas de emociones no siempre describen la gran diversidad de emociones que se dan en la interacción social cotidiana. Para el caso de la música, el modelo categórico suele ser muy utilizado en sistemas informáticos relacionados con reproducción musical, en donde se aplican etiquetas globales a las canciones, es decir, etiquetas emocionales fijas para toda la duración de una canción. Este tipo de descriptores emocionales musicales son cuestionados, en especial, por los artistas, quienes insisten en que la estructura musical de una canción genera cambios emocionales a lo largo del tiempo [34].

Los modelos dimensionales, a diferencia de los categóricos, consideran dos o más ejes dentro de un plano cartesiano, lo que permite ampliar el nivel de detalle en la clasificación de emociones al ubicarlas dentro de un plano. Dentro de este campo de estudio destacan los modelos de los investigadores Whissell [35], Russell [4] y Plutchik [36], quienes consideran las emociones como un espacio 2D continuo, cuyas dimensiones son la evaluación y la activación. En la Figura 2.3 se muestra una de las variantes de los modelos dimensionales propuestos por James A. Russell [4], en donde se puede observar algunas emociones ubicadas a través de puntos sobre un espacio bidimensional, el eje y corresponde al nivel de *arousal*, el eje x corresponde al nivel de *valence*. Por una parte, el *arousal* se relaciona con el nivel de energía y activación en la persona (el cual puede ser nulo, bajo, medio, alto), y puede tener diversas traducciones al español, como por ejemplo: excitación, intensidad y energía. Por otra parte, el *valence* se relaciona con la evaluación de la emoción, considerando que las emociones positivas (sentirse contento, feliz, relajado) implica que la persona se sienta bien, mientras que

las emociones negativas (sentirse enojado, triste, deprimido) generan una sensación de malestar en la persona. Normalmente la traducción para *valence* al español corresponde solamente a valencia. Con el objetivo de estandarizar la presentación de ambos términos a lo largo de este documento, y evitar confusiones de traducción, ambos términos son presentados en inglés, respectivamente *valence* y *arousal* (V/A). La aplicación del modelo dimensional en la música presenta como ventaja la posibilidad de diseñar descriptores emocionales más detallados, pero en contraste, agrega complejidad en el proceso de etiquetado, al ser necesario un proceso de capacitación más elaborado para garantizar que las etiquetas emocionales de los oyentes sean consistentes, y en general, de calidad.

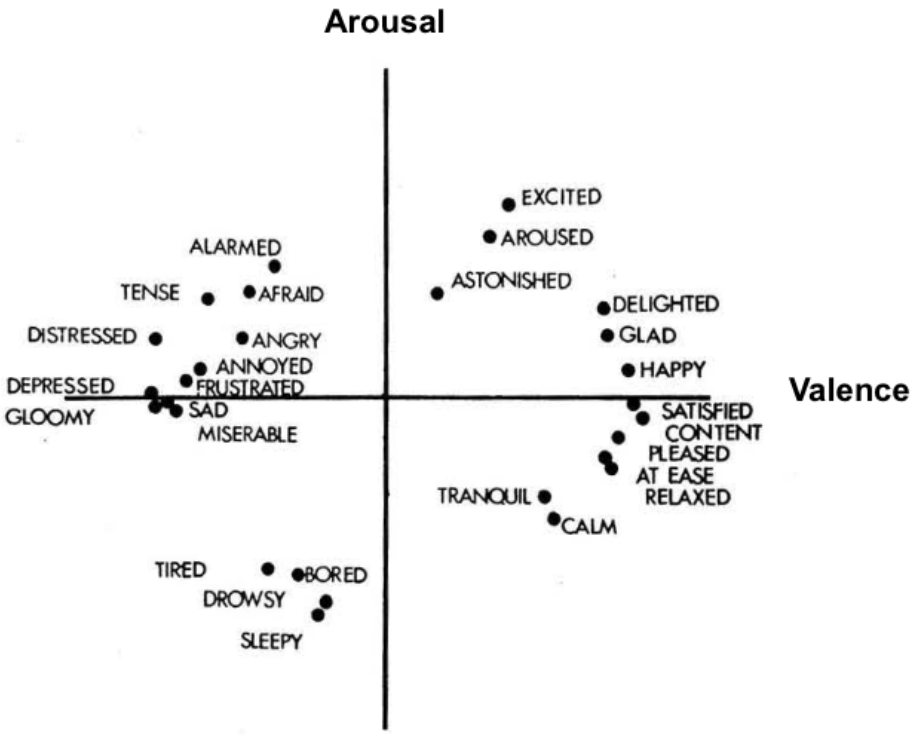


Figura 2.3: Modelo dimensional de 28 emociones. [4]

Resulta ser una buena estrategia utilizar las fortalezas de cada tipo de modelo, considerando la posibilidad de generar un modelo híbrido [37]. También es importante tener en cuenta que en una comunicación HCI (*Human-Computer Interaction*) los seres humanos transmiten sus comportamientos de manera continua; y como se mencionaba antes, este es el caso de la experiencia de una apreciación musical. Por lo que resulta necesario definir si se quiere realizar una medición en un punto de tiempo concreto (anotación estática), o si esta medición considera un intervalo del tiempo (anotación dinámica) [38]. Para el caso de la música, que se experimenta a lo largo del tiempo, resulta conveniente considerar la anotación dinámica, para estudiar los cambios emocionales del oyente frente a los cambios de las características musicales.

En el contexto de la música, el proceso de etiquetado implica que el oyente evalúe una canción especificando a través de etiquetas (*tags*, *labels*) su percepción emocional. Si

el modelo es categórico el oyente debe seleccionar una emoción en particular, mientras que si el modelo es dimensional, el usuario selecciona valores dentro de escalas definidas para *valence* y *arousal*. Bajo este esquema la clasificación emocional no se realiza a través del análisis de las características del sonido, sino de la percepción del oyente [39]. Sin embargo, definir una apropiada semántica a una etiqueta resulta una tarea compleja. La gran cantidad de términos que pueden utilizarse para dar un calificativo emocional a través de etiquetas, generalmente incluye cierto grado de subjetividad en el proceso de evaluación emocional por parte del usuario final [40].

Lo interesante del uso de modelos de clasificación emocional, es que permiten establecer una relación entre la clasificación obtenida por la percepción del oyente, y los análisis automáticos de las características de sonido obtenidos por diferentes *frameworks*. La comprensión y medición de estas relaciones puede ser considerado un factor relevante en el proceso de diseño de un sistema recomendador musical.

2.4 Analizadores de contenido en el sonido

Los analizadores de contenido de audio enfrentan diversas complejidades, en su mayoría relacionadas con las técnicas de procesamiento de señal que son necesarias para el análisis de piezas musicales grabadas en archivos de formato digital. Un analizador de contenido debe reconstruir las características intrínsecas de la música a partir de una etapa inicial de procesamiento de señal. Para lograrlo, por un parte, se implementan diversas técnicas en sus algoritmos de procesado de señal, extracción, selección y clasificación de características [41]. Por otra parte, se deben considerar las diferentes propiedades de sonido, que son determinadas por los diversos formatos disponibles para la grabación de la pieza musical. En [42], se resalta el efecto generado por algoritmos de compresión de sonido (como el formato mp3), los cuales determinan unas condiciones diferentes para el procesamiento de señal desde herramientas computacionales.

En cuanto a los analizadores de contenido, hay que precisar su nivel de estudio específico. Por una parte, se encuentra el estudio de la operación interna, en donde se busca entender, analizar, mejorar e incluso proponer nuevas técnicas, vinculadas directamente con el procesamiento de señal y algunas fases posteriores. Por otro lado, se encuentran las librerías alto nivel, en donde se ofrecen una serie de funcionalidades de extracción de características de la música, para las cuales se pueden aplicar modelos de clasificación, siendo uno de ellos, la clasificación emocional de la música. En general, los analizadores de contenido se concentran más en características de bajo nivel (relacionadas con señales digitales), mientras que las librerías de alto nivel se concentran más en características de alto nivel (relacionadas con la música, y clasificaciones entendibles por el oyente y el artista). Un analizador de contenido debe reconstruir las características intrínsecas de la música a partir de una fase inicial encargada del procesamiento de señales. Para ello, se implementan varias técnicas en el algoritmo de procesamiento de señales, extracción, selección y clasificación [41]. En la Figura 2.4 se muestra el proceso que sigue típicamente a un analizador de contenido, desde la etapa de procesamiento de la señal digital, seguida de la extracción de características de bajo nivel, y la reconstrucción de características de alto nivel, para finalmente aplicar

modelos de clasificación para obtener piezas musicales clasificadas.

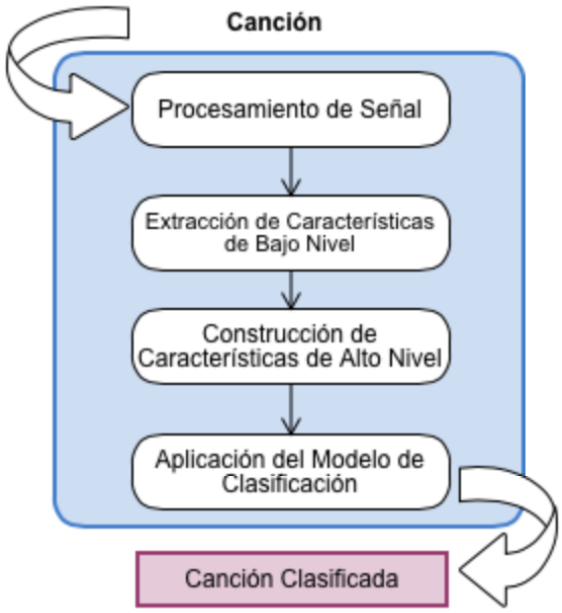


Figura 2.4: Proceso general de un analizador de contenidos. Elaboración propia.

Los analizadores de contenido tienen el objetivo principal de extraer las características de sonido. En algunos casos, los analizadores incluyen funcionalidades adicionales implementado modelos que permiten clasificar por criterios como género musical, estilo, artista, compositor, emociones y más [43]. Las librerías de alto nivel permiten hacer parte del proceso de los analizadores de contenido, pero son más orientadas a la fácil utilización, pues no requieren conocimientos avanzados en temas relacionados con el procesamiento de señales digitales, y generalmente ofrecen funcionalidades complementarias para procesos de clasificación.

2.5 Librerías de alto nivel para MER

Las librerías de alto nivel presentadas como APIs (*Application Programming Interface*) están orientadas a ofrecer un conjunto de funcionalidades para la extracción y clasificación de características intrínsecas de la música. Este tipo de herramientas se encuentran dirigidas a un perfil de usuario no experto en procesamiento de señales de sonido. El potencial de cada una de estas librerías depende en gran medida de su definición técnica interna para el procesamiento de señales digitales, que hace parte de las diversas técnicas desarrolladas en los sistemas MIR.

Algunas características generales de las diferentes librerías son:

- Tienen un límite en cuanto a la variedad de características intrínsecas de la música que pueden reconstruir.

- Tienen una determinada confiabilidad, en cuanto a la reconstrucción efectiva de una determinada característica intrínseca de la música y su respectiva clasificación.
- En consideración a su licenciamiento pueden ser de libre uso o comercial.
- Algunas son de código abierto, otras sencillamente funcionan como cajas negras.
- Algunas se encuentran disponibles como servicios en la nube, y requieren de acceso a internet para su utilización. Estos servicios también presentan algunas restricciones, como por ejemplo el número de veces que se puede consumir un servicio web por hora.

La utilización de una librería de extracción de características musicales requiere de un proceso riguroso de selección para identificar las fortalezas y debilidades de cada una de estas soluciones. Podría incluso considerarse la utilización combinada de diferentes librerías, con el propósito de lograr los mejores resultados posibles. Las características de sonido que una librería puede detectar se suelen clasificar generalmente en dos categorías: alto nivel y bajo nivel. En algunos casos las características de más alto nivel propias de la música como son el ritmo, la armonía y el modo, se clasifican dentro de las categorías de ritmo y tonales [44]. Es importante comprender el tipo de característica de sonido con el que trabaja cada librería y de igual manera su relación directa con conceptos musicales.

2.6 Predicción de valores con *machine learning*

Machine learning es la ciencia de conseguir que las computadoras actúen sin haber sido explícitamente programadas. En general, el *machine learning* puede ser utilizado para resolver problemas en donde se requiere hacer predicciones de valores, como también para resolver problemas en donde se necesita la clasificación de datos. A continuación se presenta una fundamentación teórica para los sistemas de predicción basados en *machine learning*. Se describen los siguientes elementos: métricas para la evaluación del desempeño, algoritmos de *machine learning*, y una técnica particular para la reducción de características conocida como el análisis de componentes principales (PCA).

Métricas para la evaluación del desempeño: Existen diversas métricas utilizadas para evaluar el desempeño de los modelos predictivos, cuyo objetivo consiste en analizar la diferencia entre la predicción de una observación y el valor real de la misma. Se describen algunas de las métricas más utilizadas de acuerdo los trabajos incluidos en la revisión del estado del arte:

- Error absoluto medio (MAE): se considera la diferencia absoluta entre el valor objetivo y el valor predicho por el modelo, la puntuación se da de manera lineal, lo que significa que las diferencias individuales se ponderan por igual. Su valor oscila entre 0 y 1, y se busca siempre minimizar y llevar a 0. La expresión matemática se define en 2.6.1.

$$MAE = \left(\frac{1}{n}\right) \sum_{i=1}^{n} |y_i - x_i| \tag{2.6.1}$$

- Error cuadrático medio (RMSE): representa la desviación estándar de la muestra de las diferencias entre los valores predichos y los valores observados. RMSE es útil cuando no se desean errores grandes. La expresión matemática se define en 2.6.2.

$$RMSE = \sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n}(y_i - x_i)^2} \tag{2.6.2}$$

- Coeficiente de determinación (R^2): refleja el nivel de bondad del ajuste de un modelo con respecto a una variable que se pretende explicar. Su valor oscila entre 0 y 1, y entre más cercano a 1 se encuentre, mayor se considera su nivel de ajuste. La expresión matemática se define en 2.6.3.

$$R^2 = \frac{\sigma^2 xy}{\sigma^2 x \sigma^2 y} \tag{2.6.3}$$

Algoritmos de *machine learning*: en cuanto a las alternativas para implementar modelos que permitan predecir valores bajo el concepto de regresión, a continuación se mencionan algunos de los algoritmos identificados dentro de la literatura revisada.

- *Support vector machine (SVM)*: algoritmo de aprendizaje supervisado que se fundamenta en el *Maximal Margin Classifier*, y a su vez en el concepto de hiperplanos, utilizado para clasificación múltiple y regresión.

- *Deep long-short term memory recurrent neural networks (LSTM-RNN)*: algoritmo de *machine learning* basado en redes neuronales, con la característica particular de que posee neuronas recurrentes en la capa oculta. Esto le permite al algoritmo procesar datos secuenciales.

- *Bi-directional long short-term memory recurrent neural networks (BLSTM- RNNs)*: algoritmo de *machine learning* basado en redes neuronales con una capa oculta de neuronas que tienen un comportamiento recurrente. En este caso, la información secuencial se reconoce en ambas direcciones, de futuro a pasado, o de pasado a futuro.

- *Adaptive aggregation of gaussian process regressors (AAGPR)*: estrategia que considera el diseño de múltiples regresores individuales, en donde cada uno es entrenado con una característica diferente, y posteriormente, los resultados individuales son agregados para obtener un resultado general. Esta estrategia es utilizada cuando se desea considerar los niveles de importancia por cada característica considerando su influencia en los valores de salida (predicciones).

- *ConvNet*: es una clase de red neuronal artificial, más conocida como red convolucional artificial, utilizada más específicamente para análisis de imágenes. Sin embargo, como *backbone* ya definido, tiene aplicaciones adicionales en campos de estudio como sistemas recomendadores y problemas relacionados con series temporales.

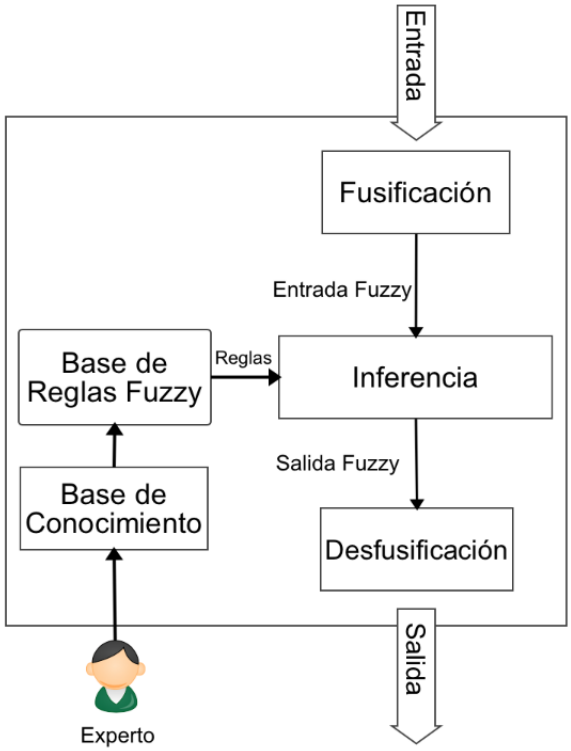


Figura 2.5: Sistema de inferencia difuso. Elaboración propia.

Análisis de componentes principales (PCA): es un algoritmo de reducción de dimensiones que normalmente se suele aplicar después del proceso de selección de características. Este método analiza las correlaciones existentes entre las características inicialmente disponibles para el modelo, y luego, genera una nueva representación de dichas características a través de un nuevo conjunto de características más pequeño, a las cuales se les identifica como componentes. Por una parte, esta reducción de características facilita el análisis, y la visualización del nuevo *dataset*, especialmente si se logra una reducción a 2 dimensiones, y por otra parte, disminuye la complejidad computacional en los procesos de entrenamiento.

2.7 Clasificación no determinística con *Fuzzy*

La lógica difusa (*Fuzzy Logic*) ha sido ampliamente utilizada para la toma de decisiones en entornos de información imperfecta. De hecho, existen varios trabajos aplicados al campo de los sistemas de control para la predicción, selección, monitorización, control y optimización [45]. En la teoría de conjuntos tradicional (no difusa), las operaciones se realizan con valores discretos para la lógica binaria (*crisp*), mientras que en los sistemas difusos estas operaciones se realizan con valores de pertenencia (lógica de valor continuo). Los principales componentes de un sistema difuso se muestran en la Figura 2.5.

A continuación se presenta la descripción para cada uno de los componentes:

- Entrada: las variables que definen el antecedente.
- Salida: las variables que definen el consecuente.
- Fusificación: proceso en el que se asigna el grado de pertenencia de los valores *crisp* de las entradas a los conjuntos difusos asociados.
- Inferencia: el conjunto de reglas difusas que se evalúan en los antecedentes (entrada).
- Desfusificación: proceso en el que a partir del conjunto difuso obtenido anteriormente (tras el proceso de inferencia) se toma como entrada para dar un valor *crisp* de salida.
- Base de reglas difusas: representación del conocimiento mediante conjuntos difusos.
- Base de conocimiento: el conocimiento del experto.
- Experto: la persona que es experta en el área de conocimiento.

En cuanto a las principales características de la lógica difusa, cabe destacar las siguientes:

- Permite la evaluación multivaluada, lo cual es una versión extendida de la lógica clásica.
- Permite el uso de cuantificadores sobre adjetivos, lo que se suele utilizar en las expresiones humanas.
- Define conjuntos difusos, permitiendo trabajar con elementos cuyo grado de pertenencia a cada conjunto viene determinado por su función de pertenencia [46].
- Es posible definir un módulo de inferencia difusa a partir de reglas lógicas.

2.8 Clasificación determinística con *machine learning*

A continuación se presenta una fundamentación teórica para los sistemas de clasificación basados en *machine learning*. Se describen los siguientes elementos: matriz de confusión, métricas para la evaluación del desempeño, y algoritmos de clasificación.

Matriz de confusión: permite la visualización del desempeño del modelo de clasificación. Cada columna en la matriz representa el número de predicciones por clase, mientras que cada fila representa la cantidad de elementos que pertenecen a la clase real. La matriz de confusión permite realizar un comparativo entre la predicción de la clase y la clase real desde el punto de vista de coincidencias. En la Figura 2.6 se observa una matriz de confusión de ejemplo, que involucra un sistema de clasificación binario,

con la clase benigno y maligno, en donde además se señalan los valores verdaderos negativos (TN), falsos positivos (FP), falsos negativos (FN), y verdaderos positivos (TP). Esto valores se encuentran incluidos en las expresiones matemáticas de las métricas de desempeño que se detallan a continuación.

Figura 2.6: Captura de visualización de una matriz de confusión de ejemplo con *Python*.

Métricas para la evaluación del desempeño: existen diversas métricas utilizadas para evaluar el desempeño de los modelos de clasificación, cuyo objetivo consiste en analizar la coincidencia entre la predicción de una observación y el valor real de la misma. Se describen a continuación algunas de las métricas más utilizadas de acuerdo los trabajos incluidos en la revisión del estado del arte; para todos los casos el umbral de análisis varía entre 0 y 1, siendo 1 el valor máximo de referencia, y además considerado como objetivo ideal de desempeño.

- Precisión (*Accuracy*): determina un nivel de coincidencia general entre todos los elementos clasificados correctamente, y el total de elementos. La expresión matemática se define en 2.8.1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.8.1}$$

- Exactitud (*Precision*): determina qué porcentaje de los elementos identificados como pertenecientes a una clase (*True Positive*), en realidad lo son. La expresión matemática se define en 2.8.2.

$$Precision = \frac{TP}{TP + FP} \tag{2.8.2}$$

- Recuerdo (*Recall*): determina qué porcentaje de los elementos identificados como pertenecientes a una clase (*True Positive*), se ha identificado correctamente. La expresión matemática se define en 2.8.3.

$$Recall = \frac{TP}{TP + FN} \tag{2.8.3}$$

- Medición F *(F-measure)*: F1 combina Precisión y Recall en una sola medida. La expresión matemática se define en 2.8.4.

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{2.8.4}$$

Algoritmos de clasificación: muchos de los algoritmos utilizados en predicción suelen ser también utilizados (con algunas variaciones) para clasificar, como es el caso de SVM y SVR. También es importante resaltar, que algunas técnicas complementarias en pre-procesado de datos como PCA, son independientes de los enfoques de predicción y clasificación, y pueden ser utilizadas en ambos casos, pues finalmente es un asunto más relacionado con los *datasets* para abordar fases iniciales de selección y reducción de características. A continuación se mencionan los algoritmos que se encontraron con mayor frecuencia en la literatura revisada.

- *Random Forest Classifier (RFC) y Random Forest Tree (RFT)*: algoritmo clasificador basado en la generación de árboles predictores, en donde cada árbol es generado a partir de valores aleatorios, y a través de aprendizaje supervisado se generan las reglas de decisión. Este algoritmo puede ser implementado tanto para clasificación como para regresión.

- *Multi-layer perceptron (MLP)*: red neuronal artificial que posee una arquitectura integrada por múltiples capas, con la capacidad de resolver problemas que no son linealmente separables.

2.9 Sistemas recomendadores musicales

El objetivo principal de los sistemas recomendadores musicales (MRS) consiste en facilitar a los usuarios oyentes la búsqueda y el descubrimiento de nuevas canciones que sean de su interés. La efectividad de los procesos de recomendación se mide en función del nivel de aceptación (*likes/dislikes*) expresado por el usuario oyente en relación a la canción recomendada [47] [48] [49]. Este tipo de aplicaciones permite a un usuario expandir fácilmente su conocimiento sobre nuevas piezas musicales, como también permite a los productores y músicos darse a conocer, y aprovechar la retroalimentación (*feedback*) recibida por parte de los oyentes.

Parte del éxito de un sistema recomendador musical depende de la identificación de las preferencias de usuario, y de cómo explotar adecuadamente esta información para generar las recomendaciones [49]. Para lograr realizar recomendaciones acertadas, se suele estudiar cuáles son las mejores herramientas para extraer y clasificar los atributos de la música, y luego establecer una relación entre estos atributos y las emociones que experimenta un oyente ante una determinada pieza musical. Para ello es necesario considerar dos de los problemas más importantes de este proceso: la construcción de preferencias para usuarios nuevos, y la recomendación de nuevas canciones; situación conocida como el problema del arranque en frío (*cold-start problem*) [49].

Para abordar estos problemas, el recomendador debe aprender de las preferencias de cada usuario y aplicar un modelo efectivo de clasificación [49]. Adicionalmente, el

recomendador podría implementar mecanismos para obtener retroalimentación (*feedback*) por parte del usuario, y evaluar la percepción que se tiene de sus recomendaciones. Resulta también importante considerar el contexto (actividad, momento del día, ubicación) del oyente y sus reproducciones históricas, para así enriquecer las recomendaciones futuras [50].

El algoritmo de procesamiento de sonido, el sistema de etiquetado de la música, y el esquema de evaluación de emociones; son elementos típicamente utilizados hasta el momento para estudiar la relación de la música con las emociones. Sin embargo, es importante considerar otros elementos que permitan que esta línea de investigación siga avanzando en resultados y modelos más precisos. La actividad del compositor, y toda la información relativa a este proceso, puede ser una entrada (*input*) que marque la diferencia en los resultados obtenidos hasta el momento, y que incluso permita generar nuevos casos de aplicación. Por lo que, a pesar de que esta Tesis no tiene como eje central el proceso de composición musical, se tiene en cuenta al compositor a través de entrevistas, encuestas, y experimentos que permitan comprender sus necesidades, dificultades y posibilidades de aportar a las estrategias de recomendación.

En general, los MRS están integrados por los siguientes elementos: artistas, oyentes, canciones y estrategias de recomendación. Los artistas publican sus canciones en una plataforma digital de música con el fin de impulsar sus carreras comerciales. Los oyentes utilizan la plataforma digital de música para encontrar las canciones específicas que quieren escuchar, además, también están interesados en descubrir nuevas canciones que les puedan gustar. Para lograr este proceso de descubrimiento, los MRS generan el emparejamiento (*match*) entre una canción y un oyente específico a través de estrategias de recomendación [51]. Las funcionalidades y las posibilidades de éxito en los MRS están definidas, en la mayoría de los casos, por las diferentes estrategias de recomendación, que han ido evolucionando en los últimos años [52] [53] [54] [55]. Las estrategias que se presentan a continuación no definen una taxonomía en el campo de MRS, sino que se explican teniendo en cuenta los diferentes enfoques encontrados a través de un proceso de revisión de la literatura.

- **Filtrado colaborativo (CF):** El CF genera predicciones automáticas según los intereses de un usuario mediante la recopilación de información sobre las preferencias de un amplio conjunto de usuarios, en la mayoría de los casos procedentes de las redes sociales [47].

- **Filtrado demográfico (DF):** El DF se basa en la clasificación de los perfiles de los usuarios por criterios como la edad, el estado civil, el sexo, etc [55].

- **Filtrado basado en contenido (CBF):** El CBF recomienda canciones en función de sus características internas, que pueden ser de bajo nivel (características de sonido a nivel de señal) o de alto nivel (características musicales). Es necesario que exista una relación entre los valores de estas características y el grado de aceptación por parte de la percepción del usuario [52].

- **Filtrado híbrido (HF):** HF funciona combinando los distintos tipos de filtrado. En general, este enfoque proporciona mejores resultados en comparación con la

implementación de un solo tipo de filtrado, sin embargo, requiere un proceso de ajuste y optimización detallado [51].

- **Contexto de usuario (UC):** UC incluye cualquier información que pueda utilizarse para caracterizar el contexto del usuario; esta información puede obtenerse de diferentes fuentes (información personal, información de los sensores, información de la actividad del usuario) [47].

- ***Metadata* (MD):** MD incluye un grupo de datos que describen la canción. Estos datos pueden clasificarse por categorías, por ejemplo en [51] se proponen 3 grupos de *metadatos*: editoriales, culturales y acústicos.

- **Filtrado basado en emociones (EBF):** El EBF considera la relación entre las emociones humanas y las características intrínsecas de la música; basándose en esta relación, el EBF tiene como objetivo identificar los valores de las características que generan un interés en el usuario y evocan una emoción particular [53] [56].

- **Enfoque personalizado (PA):** El PA consiste en el diseño e implementación de modelos centrados en el usuario que permiten crear experiencias de recomendación altamente efectivas. Algunos autores reconocen el PA como un paradigma, por lo que podría incluir estrategias basadas en la interacción con el usuario y el contexto [51].

- **Basado en listas de reproducción (PLB):** las listas de reproducción diseñadas por el usuario incluyen canciones que llegan a ser más importantes que las inicialmente recomendadas por el sistema. A partir de estas listas, es posible estudiar las características de las canciones allí contenidas, para luego realizar recomendaciones basadas en similaridad [57].

- **Basado en popularidad (PB):** esta estrategia genera una tendencia a recomendar las canciones que son comercialmente más famosas, o que tienen una presencia destacada en el mercado musical gracias a las grandes inversiones en estrategias de marketing [58]. Típicamente hace uso de contadores de reproducción (*music play counts*). Este tipo de estrategia suele crear un sesgo que afecta a los artistas noveles, porque su probabilidad de ser recomendados es muy baja [59].

- **Basado en similaridad (SB):** los sistemas basados en la similaridad calculan el grado de proximidad entre las características de una canción y otra; este grado de proximidad se utiliza para generar las recomendaciones [56].

- **Basado en interacción (IB):** los sistemas basados en la interacción se centran en analizar el comportamiento del usuario con respecto al uso del sistema, considerando aspectos como: cuándo el usuario genera una reproducción, qué canción se reproduce y cuántas veces, cuál es la relación entre la canción y determinados días de la semana, entre otros [60]. Generalmente, se utilizan diferentes tipos

de registros para almacenar los diversos eventos generados por el usuario en el sistema.

Después de la presentación de las diferentes estrategias, es necesario aclarar que la definición de una taxonomía para las estrategias de recomendación en el campo de MRS podría considerarse un reto, debido a que en muchos de los trabajos revisados algunas estrategias con los mismos objetivos se presentan con nombres diferentes. Además, algunos autores presentan algunas estrategias en un nivel superior como un paradigma, por ejemplo, PA, pero otros únicamente se refieren a ellas como estrategias.

2.10 Sesgos en sistemas recomendadores

Un sistema informático con problemas de sesgo discrimina injustamente algunos elementos específicos al negar o disminuir la posibilidad de que aparezcan en un proceso de interacción entre el usuario final y el sistema [61]. En el caso de los MRS, un sistema particular de recomendación de música que siempre recomienda las canciones más populares, y nunca o muy raramente recomienda canciones producidas por artistas noveles (*non-superstar*) es un claro ejemplo de sesgo. También es importante destacar el impacto económico que los sesgos pueden generar en un determinado modelo de negocio. En el caso de la industria musical, el problema no solamente radica en la fama del artista, sino también en las ganancias que puede recibir este artista, teniendo en cuenta que la mayoría de los servicios de *streaming* de música pagan al artista en función de sus contadores de reproducciones de canciones [54] [62].

Según Friedman y Nissenbaum, los sesgos en informática pueden clasificarse en 3 categorías [61]: preexistentes, técnicos y emergentes, los cuales se describen brevemente a continuación:

- **Sesgos preexistentes:** Sesgos generados por instituciones, prácticas y actitudes sociales. Este tipo de sesgo es promovido por la sociedad, tiene una relación directa con la cultura y puede ser incluido de forma explícita o implícita por los clientes, los diseñadores de sistemas y otras partes interesadas.

- **Sesgos técnicos:** Los sesgos tienen su origen en limitaciones técnicas o consideraciones técnicas. Este tipo de sesgo surge de las limitaciones técnicas, que pueden estar presentes en el hardware, el software y los periféricos. Para el caso del software, en los sesgos técnicos es muy importante analizar y tratar los algoritmos descontextualizados, que promueven el procesamiento injusto de los datos.

- **Sesgos emergentes:** Este sesgo únicamente puede detectarse en un contexto real de uso, y suele aparecer una vez ha finalizado la fase de diseño de un MRS. Normalmente es un resultado de los cambios en los conocimientos de la sociedad, la población o los valores culturales.

En la mayoría de los casos, los sistemas informáticos, así como la ciencia en general, intentan ayudar y mejorar diferentes aspectos de la vida, como los modelos de negocio,

los servicios de entretenimiento, los servicios de salud, las políticas sociales, y más. A pesar de este hecho, los sesgos preexistentes podrían promover una percepción negativa desde la perspectiva de los usuarios finales en relación con los sistemas informáticos, ya que para el usuario final no es claro dónde y por qué se genera un trato injusto sobre él. En cuanto a los sesgos técnicos existen dos escenarios relevantes para analizar, primero, el caso en el que los sistemas informáticos promueven sesgos debido a diversas debilidades en el proceso de diseño de sus algoritmos; y segundo, el caso en el que aunque el personal técnico ha identificado los sesgos preexistentes, no implementa ninguna mejora técnica para mitigarlos. Ambos escenarios han sido considerados para marcar los sesgos técnicos en los trabajos que se analizan en la siguiente sección.

2.11 Conclusiones

En este capítulo se presentaron los conceptos más generales e importantes para el desarrollo y la comprensión de esta Tesis. Como premisa fundamental se reconoce la estrecha relación entre las características intrínsecas de la música y las emociones percibidas por el oyente, resaltando la importancia de representar esta percepción emocional a través de modelos computacionales, estudiados y propuestos por el campo de la computación afectiva; estos modelos afectivos son considerados uno de los factores más importantes a tener en cuenta en el diseño de las estrategias de recomendación de las plataformas de reproducción musical. Luego, se describen los analizadores de contenido y su aplicación más particular, destacando su importancia para la extracción de características de sonido a partir del formato digital de las piezas musicales. Seguidamente, se presenta la fundamentación teórica de los sistemas de predicción basados en *machine learning*, sistemas de clasificación no determinísticos basados en *fuzzy*, y los sistemas de clasificación determinísticos basados en *machine learning*. El capítulo cierra abordando el concepto de más alto nivel de esta Tesis, los sistemas recomendadores musicales, y adicionalmente el tema se complementa con una fundamentación sobre sesgos.

Capítulo 3

Estado del arte

En este capítulo se presentan los trabajos y desarrollos previos más importantes relacionados con los objetivos de esta Tesis. Esta revisión de literatura se plantea con base en las preguntas de investigación y los objetivos propuestos en el capítulo 1. Los diversos trabajos relacionados con sistemas recomendadores musicales (MRS) también fueron clasificados de acuerdo a su temática principal. Esta clasificación se representa a través del diagrama de bloques de la Figura 3.1, en donde la raíz del diagrama corresponde a los MRS, luego los bloques que aparecen en los siguiente niveles, corresponden a las diferentes temáticas que sirven de soporte. Cada temática estudiada y presentada en los apartados de este capítulo sigue el proceso de revisión sistemática de literatura explicado previamente en el apartado 1.5. Se consideran trabajos desde el año 2013 en adelante para todos los casos, las áreas y palabras claves definidas en la Tabla 1.1, y las fuentes bibliográficas definidas en la Tabla 1.2.

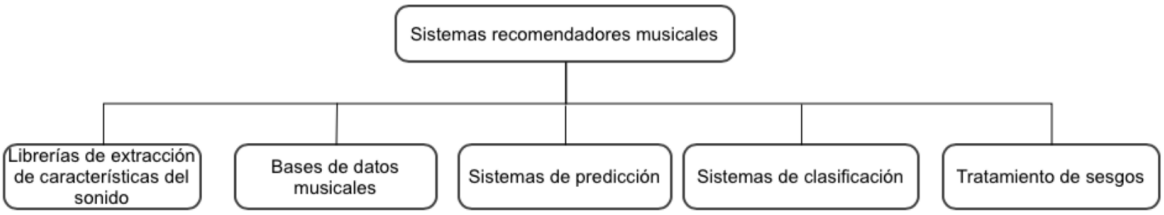


Figura 3.1: Diagrama de bloque para la organización de la literatura. Elaboración propia.

Estas temáticas también pueden ser entendidas como los principales elementos y/o fases para el diseño de un MRS, y la forma en que se relacionan, se expone a continuación:

1. Las librerías de extracción de características de sonido permiten caracterizar una canción, pero deben disponer de un repositorio de canciones en donde puedan aplicarse, y es allí donde radica la importancia de estudiar las bases de datos musicales existentes; entre las cuales destaca MediaEval por su completo etiquetado emocional, el cual resulta de gran interés para los objetivos de esta Tesis.

2. Al disponer de una o más bases de datos musicales, y respectivamente de las características de sonido de cada una de las canciones, el siguiente paso en el diseño de un MRS corresponde a la estrategia de recomendación. Esta estrategia tiene como objetivo establecer una relación de preferencia entre la canción y un oyente, y en consideración a los objetivos de esta Tesis, la recomendación se concentra en la percepción emocional, lo que da lugar a generar recomendaciones de canciones según la emoción que pueda evocar. Partiendo de algunos de los modelos de computación afectiva más utilizados (dimensionales y categóricos), anteriormente presentados en el apartado 2.3, las canciones se pueden etiquetar emocionalmente con una coordenada de *valence* y *arousal*, como también a través de una categoría en particular (feliz, triste, aburrido, deprimido). Teniendo en cuenta estas dos alternativas de etiquetado, desde el punto de vista de las estrategias de recomendación basadas en algoritmos de *machine learning*, implicaría considerar respectivamente el diseño de un sistema de predicción, o de un sistema de clasificación. Adicionalmente, es importante resaltar que dentro de un escenario de clasificación es posible hacerlo de manera no determinística, como también determinística, por lo que resulta interesante comprender las ventajas y desventajas.

3. Finalmente, el análisis de sesgos permite comprender el impacto que éstos tienen sobre el desempeño de los MRS. Para luego, proponer e implementar algunas acciones que permitan mitigar dicho impacto, y con ello, mejorar la experiencia de los usuarios: oyentes, y artistas.

De esta manera, en el **Apartado 3.1** se aborda una revisión enfocada en librerías que permiten la extracción de características de alto nivel de la música. Algunas de estas librerías también implementan procesos de clasificación. En el **Apartado 3.2** se revisan *datasets* de piezas musicales disponibles para llevar a cabo experimentos de MER. De la anterior revisión se selecciona MediaEval y se profundiza en su análisis a través del **Apartado 3.3**. En el **Apartado 3.4** se analizan trabajos relacionados con el reconocimiento de emociones a través de algoritmos de predicción. En el **Apartado 3.5** se estudian trabajos de MER para la clasificación emocional de la música bajo el enfoque de sistemas no determinísticos (*fuzzy*). Luego, en el **Apartado 3.6** se analizan trabajos de MER para la clasificación emocional de la música bajo el enfoque de un sistema determinístico basado en *machine learning*. En el **Apartado 3.7** se abordan trabajos relacionados con sistemas recomendadores de piezas musicales (MRS), y además, son clasificados por las estrategias de recomendación disponibles para MRS. Posteriormente, en el **Apartado 3.8** se realiza un análisis de sesgos con un enfoque especial en el análisis del efecto de la popularidad sobre procesos de recomendación. Adicionalmente, se presentan algunas recomendaciones para el tratamiento de sesgos en MRS. Finalmente, en el **Apartado 3.9** se presentan las conclusiones generales de la revisión del estado del arte, considerando un análisis integral que aborda todos los tipos de trabajos estudiados en los diferentes apartados indicados anteriormente.

3.1 Librerías de alto nivel para la extracción de características musicales

En este apartado se presenta el análisis de cuatro librerías de alto nivel para el reconocimiento de emociones en la música: *Spotify API*, *jMIR*, *AcousticBrainz* y *OpenSmile*. Luego del proceso de búsqueda que considera las palabras claves de la fila 3 de la Tabla 1.1, en particular *recuperación de información musical* y *librerías para la recuperación de información musical*, se dio paso al proceso de selección. A partir de allí se puso como foco quedarse con trabajos que presenten librerías que sean las más citadas en relación al reconocimiento de emociones en la música. Luego de la revisión en detalle de los trabajos finalmente seleccionados, se analizaron un total de 4 librerías que se describen en este apartado. Además, se decidió indagar de cada una, cuáles eran las características que permitía considerar, atendiendo a sus posibilidades para el reconocimiento de emociones. Las librerías fueron analizadas de acuerdo a su descripción funcional y también a pruebas realizadas por el tesista. A continuación, se describen cada una de estas librerías.

3.1.1 Spotify API

Spotify [63] es una de las principales plataformas de reproducción de música. Cuenta con un amplio repositorio de canciones de diversos géneros, al que se puede acceder desde un entorno web, como también desde aplicaciones móviles. El sistema recomendador de *Spotify* y la manera de clasificar la música para facilitar el acceso y las búsquedas, es una de las características más relevantes que ha permitido su exitosa acogida por parte de los usuarios. Dentro de las diversas funcionalidades ofrecidas por *Spotify*, se encuentra la posibilidad de utilizar el *API* de servicios [63]. Este *API* de servicios tuvo sus inicios con el proyecto *Echonest* [64], y una vez fue absorbido por *Spotify*, algunos aspectos técnicos de implementación y funcionalidad cambiaron. Entre los cambios más notables se tiene que actualmente el *API* de *Spotify* únicamente permite aplicar sus servicios sobre canciones que se encuentran en su repositorio, siendo esto una gran limitación para ejercicios de experimentación.

A continuación, se describen algunas de las funcionalidades más relevantes del API:

- Facilitar integraciones de aplicaciones propias en entorno web y móvil a los servicios de *Spotify*.
- Acceder al catálogo de música de *Spotify*.
- Obtener información (*metadata*) en formato *JSON* sobre artistas, álbumes, y canciones almacenadas en el catálogo de *Spotify*.
- Acceder de forma directa a las listas de reproducción creadas para un perfil de usuario en particular.

El *API* de *Spotify* cuenta con una entidad conocida como *audio object features*, esta entidad dispone de 18 características, entre ellas las más relacionadas con las emociones

y la música se presentan en la Tabla 3.1. Respecto al reconocimiento de emociones en la música, se tiene la posibilidad de extraer las características *valence* y *energy*. Desde el punto de vista de los modelos emocionales, estas dos características pueden analizarse como una coordenada de tipo *energy-valence* [4], de tal manera que la emoción puede localizarse en un plano bidimensional.

Tabla 3.1: Características extraídas por Spotify API

Característica	**Descripción**
Key	Identificación de la tonalidad de la canción.
Mode	Identificación del modo (mayor o menor) de la canción.
Tempo	Asociado al ritmo y a la velocidad de la canción. Permite generar un estimado de los *beats* por minuto.
Liveness	Permite identificar si la canción es en vivo. Los valores cercanos a 1 indican una alta probabilidad de que la canción sea en vivo.
Energy	Representa el nivel de intensidad y actividad en la canción. Valores cercanos a 1 indican mayor intensidad presente en la canción.
Instrumentalness	Permite identificar si una canción tiene contenido vocal. Cuanto más cercano a 1 sea su valor se determina mayor probabilidad de que la canción no tenga contenido vocal.
Valence	Esta característica se asocia a lo positiva que puede ser una canción. Los valores más cercanos a 1 determinan emociones positivas, mientras que los valores cercanos a 0 indican emociones negativas.

En cuanto a los **trabajos de aplicación**, destaca en [65] y [66] la utilización de la biblioteca *Echonest*, en sus primeras etapas de desarrollo, para el reconocimiento de emociones en el contenido audiovisual, logrando la clasificación sonora en un modelo dimensional (*valence-arousal*), con algunos ajustes particulares que permiten reconocer las siguientes emociones: excitación, feliz, relajado, triste y enfadado. Este modelo es utilizado por un sistema de recomendación para facilitar el consumo de contenido audiovisual a través de una estrategia de *streaming* adaptativo. En [67] la librería se utiliza para extraer características de audio con las que la música se clasifica según la similitud musical entre los artistas. Esta información es utilizada por un sistema de recomendación que también analiza la información de contexto y dibuja la ubicación de cada pieza musical en un modelo dimensional de emociones (*valence-arousal*). En [68] la librería se utiliza para analizar la característica de bajo nivel *beat synchronous* y genera una clasificación emocional de la música, junto con el procesamiento de otras características extraídas con Matlab; el proceso de clasificación implica el etiquetado

manual de 1000 canciones, y luego cada una se clasifica a través de un modelo dimensional (*valence-arousal*).

3.1.2 jMIR

jMIR [5] es un software de código abierto implementado en lenguaje *Java*, que se caracteriza por su flexibilidad y avanzada capacidad funcional para la recuperación de información musical. Se encuentra integrado por un grupo de componentes, que pueden ser utilizados en conjunto, dependiendo de las necesidades del experimento. De manera general en la Figura 3.2 se describe el proceso de extracción de características, y para cada una de las actividades del proceso, se relaciona el componente de *jMIR* involucrado [5].

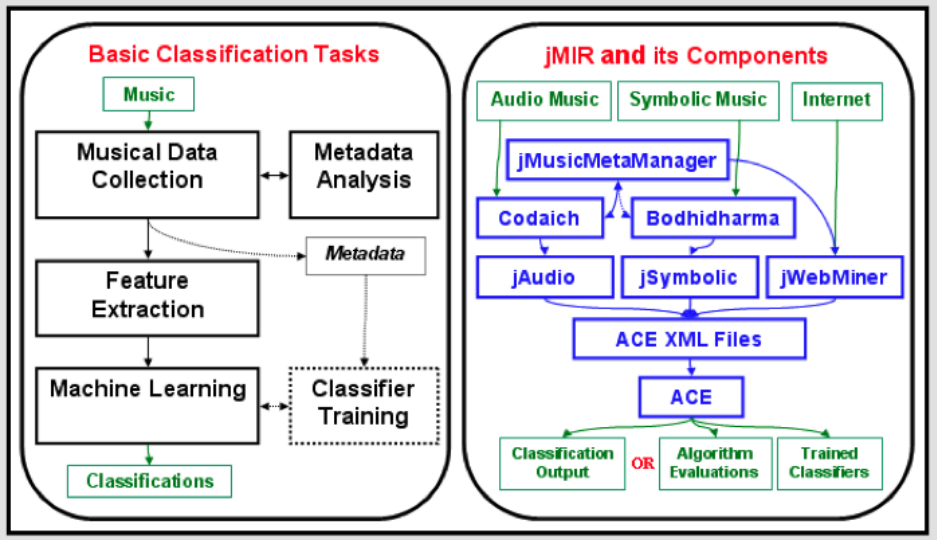


Figura 3.2: Actividades presentes en la clasificación de la música y los componentes jMIR asociados [5]

Entre los componentes de *JMIR* destaca *JAudio*, el cual permite la extracción de características de sonido desde un archivo de sonido digital. *jAudio* fue diseñado para trabajar directamente con sonido de manera general y, en principio, no implementa una herramienta concreta que permita analizar características de alto nivel de la música. Entre las principales ventajas del *jAudio*, y sus principales funcionalidades, se pueden destacar las siguientes:

- Funciona como una aplicación local, por lo que no hay dependencia de un canal de comunicación.
- Permite parametrización y extensión de uso por parte del usuario. Por ejemplo, es posible crear características de sonido adicionales a las ya definidas por defecto en la librería, lo que amplía las posibilidades de experimentación.
- Se encuentra desarrollado en lenguaje *java* y puede ser integrado con otros aplicativos.

- Permite exportar las características reconocidas a un archivo *XML* con todas los metadatos asociados a características de sonido. Este formato es interpretado por los diferentes módulos disponibles en *jMIR*.

- Se puede integrar con *jSymbolic*, el cual permite analizar características más específicas de la teoría musical, pero para este componente se requiere información simbólica adicional que generalmente se obtiene de un archivo de formato *MIDI*.

Para obtener clasificaciones de tipo emocional, género musical y entre otros, *jMIR* cuenta con el módulo *ACE2* que permite definir taxonomías y características. Con ello, es posible cargar en *ACE2* los resultados obtenidos con *jAudio* durante el proceso de extracción de características de bajo nivel. Cada uno de estos resultados cargados en *ACE2*, se considera una instancia particular, y posteriormente *ACE2* puede aplicar diversos modelos de clasificación, obteniendo así características de alto nivel que han sido definidas previamente por el usuario.

jAudio permite extraer 26 características principales de sonido de bajo nivel, a partir de las cuales, se puede definir *metafeatures*, lo que es básicamente una característica parametrizada por el usuario, y que se calcula a partir de las características principales que *jAudio* puede extraer. En la Tabla 3.2 se presentan algunas de las características de bajo nivel más relacionadas con el contexto de la música y las emociones.

Tabla 3.2: Características de bajo nivel extraídas por jAudio

Característica	**Descripción**
Power spectrum	Permite medir la intensidad y actividad en la canción. Se presenta la magnitud de la señal en decibeles (dB). Está muy relacionado con la característica *energy* definida en otras librerías.
Beat Histogram	A partir de un histograma que representa la regularidad del ritmo, se logra determinar el *tempo* de una canción. Otras características que permiten calcular el *Beat Histogram* son: *Strongest Beat, Beat Sum, Strength of Strongest Beat* [69]. Esta característica permite mostrar la frecuencia con la que se presenta una determinada velocidad en diferentes momentos de una canción.
Pitch	El rango de frecuencias obtenido entre el *low pitch* y el *high pitch* de la canción, es una medida que sirve como punto de partida para estimar la nota predominante de la canción, y con ello deducir la tonalidad y el modo.

jMIR también permite integrar modelos de clasificación, los cuales deben ser diseñados inicialmente, para luego proceder a cargarlos y entrenarlos a través del componente *ACE2*. Para el caso concreto de *jAudio*, es posible implementar un modelo de clasificación emocional en función del *Power spectrum* y del *Beat Histogram*; en términos musicales, esto permitiría personalizar la clasificación emocional de una canción de

acuerdo a su velocidad y ritmo. También es posible proponer diferentes *metafeatures* que relacionen valores de *pitch*, creando asociaciones entre características tonales y rítmicas, con emociones. El modelo de clasificación emocional podría ser categórico o dimensional [4], dependiendo en gran medida del sistema de anotaciones utilizado. El modelo necesita de algún mecanismo para establecer el proceso de anotación emocional y, muy probablemente, ese mecanismo debería aplicarse como un desarrollo adicional con el que debe integrarse el *ACE2*, o con cualquier otra librería externa que permita realizar clasificaciones con algoritmos predictivos, como los utilizados en las técnicas de *machine learning*.

Con respecto a los trabajos de aplicación, en [70] se describe con gran detalle la capacidad funcional de *jMIR*, y en relación con *jAudio*, se destaca de manera general su amplio alcance en los problemas de procesamiento de señales de sonido. En [71] *jSymbolic* se utiliza para identificar e interpretar el timbre, el ritmo, la dinámica y la melodía. En este trabajo, *jAudio* se utiliza para reconocer las características de *MFCC*, *Spectral Centroid*, *Spectral Flux* y *Zero Crossings*. Este trabajo se centra en la identificación de las emociones en la música a partir del análisis de la progresión de los acordes, de manera que se pueda identificar la tonalidad y el modo de la canción y, con ello, se pueda inferir la emoción y situarla dentro de un modelo dimensional (*valence-arousal*).

3.1.3 AcousticBrainz

AcousticBrainz [72] es una librería cuya funcionalidad principal es facilitar la recuperación de información musical sobre canciones. *AcousticBrainz* se soporta sobre las funcionalidades ofrecidas por *Essentia toolkit* [73], que en general comprende toda la capacidad de procesamiento y análisis de sonido. *AcousticBrainz* ha sido desarrollado por el esfuerzo colaborativo entre el *Music Technology Group* de la *Universitat Pompeu Fabra* y el proyecto *Music Brainz* [74].

AcousticBrainz clasifica las características que se pueden extraer de una canción en dos categorías fundamentales: bajo nivel y alto nivel. Dentro de las características de bajo nivel se encuentran aquellas que comprenden **31 descriptores acústicos, 9 rítmicos y 8 tonales**. En la Tabla 3.3 se presentan las características más representativas frente al contexto de la música y las emociones. *AcousticBrainz* permite identificar 4 emociones básicas [75]: feliz, agresivo, triste y relajado. También es posible identificar algunas características adicionales como el caso de acústico, electrónico y de fiesta; los cuales podrían estar relacionados con las emociones dentro de un proceso de clasificación y/o anotación.

Tabla 3.3: Características extraídas por AcousticBrainz

Característica	Descripción
bpm	Pulsaciones por minuto.
chords_key	Corresponde a la identificación de la tonalidad de la canción.
chords_scale	Corresponde a la identificación del modo.
mood_acoustic	Clasifica la canción en acústica o no acústica. Cuanto más cercano a 1 sea su valor se determina mayor probabilidad de que la canción corresponda a una versión acústica.
mood_aggressive	Clasifica la canción por emoción de agresividad. Cuanto más cercano a 1 sea su valor, mayor probabilidad de que la canción clasifique con emoción agresiva.
mood_electronic	Clasifica la canción de acuerdo al tipo de sonido. Cuanto más cercano a 1 sea su valor, mayor probabilidad de que la canción tenga un tipo de sonido electrónico.
mood_happy	Clasifica la canción por emoción de felicidad. Cuanto más cercano a 1 sea su valor mayor probabilidad de que la canción clasifique con emoción feliz.
mood_party	Clasifica la canción por estado emocional de: fiesta / celebración. Cuanto más cercano a 1 sea su valor, mayor probabilidad de que la canción clasifique con emoción fiesta.
mood_relaxed	Clasifica la canción por emoción de relajación. Cuanto más cercano a 1 sea su valor, mayor probabilidad de que la canción clasifique con emoción de relajación.
mood_sad	Clasifica la canción por emoción de tristeza. Cuanto más cercano a 1 sea su valor, mayor probabilidad de que la canción clasifique con emoción triste.
moods_mirex	Clasifica la canción dentro de alguno de los 5 *clusters* predefinidos para categorías de estado emocional (*mood*).

Con respecto a la característica *moods_mirex* se detallan a continuación las emociones incluidas dentro de cada de uno de los clusters [76]:

- *Cluster 1*: apasionado, entusiasta, seguro, bullicioso, ruidoso.
- *Cluster 2*: alegre, animado, divertido, dulce, amable / de buen carácter.
- *Cluster 3*: Emocionalmente inteligente, conmovedor, melancólico.

- *Cluster 4*: humorístico, tonto, cursi, peculiar, caprichoso, ingenioso, irónico.
- *Cluster 5*: agresivo, ardiente, tenso / ansioso, intenso, volátil, visceral.

Algunos trabajos en los que se aplica esta librería son: en [77] para la clasificación de la música por género musical para una base de datos de 120 canciones, en [78] se utiliza para la detección de 4 emociones básicas a través de un modelo afectivo categórico que reconoce las emociones: feliz, enojado, triste, relajado; adicionalmente también se valida la precisión del reconocimiento de la emoción, variando la selección de características que se extraen y utilizan en el modelo de clasificación; y finalmente en [44] se utiliza la librería para detectar los valores de *valence* y *arousal* en grabaciones musicales, mostrando la importancia de combinar características de bajo nivel con características de alto nivel para lograr mejores resultados en las clasificaciones emocionales de la música.

AcousticBrainz trabaja con modelos definidos y entrenados computacionalmente, para que a partir de características de bajo nivel, estos modelos puedan construir características de alto nivel que generalmente funcionan como clasificadores. Estas características de alto nivel contemplan:

- Identificar si la voz de la canción es masculina o femenina.
- Identificar el género musical de la canción.
- Clasificar la canción por emociones.

En el anexo B. Experimento con *AcousticBrainz* se presenta un ejemplo práctico en donde se utiliza la librería para extraer algunas características musicales y emocionales.

3.1.4 OpenSMILE

OpenSMILE[1] (*open-source Speech and Music Interpretation by Large-space Extraction*) ofrece un conjunto de herramientas de código abierto para la extracción y clasificación de características de sonido relacionadas con señales de voz y música [79]. *OpenSMILE* tuvo sus inicios en el año 2008 a través del esfuerzo colaborativo de los investigadores Florian Eyben, Martin Wöllmer y Björn Schulle en la *Technical University Munich* (TUM). Posteriormente, a partir de 2013, la compañía audEERING toma los derechos de *OpenSMILE* y continúa con su desarrollo. *OpenSMILE* es una de las herramientas más ampliamente utilizadas en el reconocimiento automático de emociones en computación afectiva [80]. Entre las más recientes características de *OpenSMILE 3.0* se encuentra el API *opensmile-python* que permite una fácil integración de *OpenSMILE* con lenguaje *Python*.

Las funcionalidades de *OpenSMILE* se encuentran agrupadas en 8 categorías:

1. **Fundamentales:** extracción de hasta 27000 *features*, multiplataforma (Windows, Linux, Mac, Android, iOS), procesamiento en tiempo real, alta modularidad y reusabilidad de componentes.

[1] https://www.audeering.com/research/opensmile/

2. **Entradas/salidas de audio:** lectura y escritura de archivos PCM WAV, lectura de archivos FFmpeg, grabación de sonido en tiempo real.

3. **Formato de archivos de *features*:** lectura y escritura de archivos CSV, WEKA ARFF, HTK. Escritura de archivos LibSVM.

4. **Procesamiento de señales:** funciones *Windowing*, transformada rápida de Fourier, filtro *pre-emphasis*, filtros FIR, autocorrelación, Cepstrum.

5. **Características relacionadas con voz:** *signal energy, loudness, mel-/bark-/octave-spectra, MFCC, PLP-CC, pitch, voice quality (Jitter, Shimmer), formants, LPC, line Spectral Pairs (LSP), Spectral Shape descriptors.*

6. **Características relacionadas con música:** *pitch classes (semitone spectrum), CHROMA and CENS features, weighted differential.*

7. **Procesamiento de datos:** *mean-variance normalisation, range normalisation, delta-regression coefficients, vector operations, moving average filters*

8. **Resúmenes de características:** *means, extremes, moments, segments, samples, peaks, linear and quadratic, regression, percentiles, durations, onsets, DCT coefficients, zero-crossings, modulation spectrum.*

Es importante aclarar que todas las características extraíbles a partir de estas funcionalidades se desarrollan a un bajo nivel, por lo que es necesario un procesamiento adicional para relacionarlas con características comprensibles desde el contexto de las emociones y de la música. La empresa audEERING también ofrece otras herramientas basadas en *OpenSMILE*, las cuales a partir de las características de bajo nivel son capaces de generar características de alto nivel; este es el caso de *DevAIce*, *AI SoundLab*, *EnterAIn play* y *EnterAIn Observe*.

Uno de los proyectos en donde se aplica *OpenSMILE* y que resulta de gran interés para esta Tesis es el de *MediaEval* [81]. En este caso, *OpenSMILE* fue utilizado para realizar la extracción de 260 características de bajo nivel (*low-level features*) cada 500 ms sobre un conjunto de 1802 canciones, para luego, relacionar los valores obtenidos en cada ventana de tiempo con las etiquetas emocionales suministradas por diferentes oyentes. Los detalles de *MediaEval* se profundizan y presentan en el Apartado 3.3

3.1.5 Consideraciones generales

A partir de la revisión de cada una de las librerías anteriormente mencionadas y del anexo C. Cuadro comparativo de librerías de alto nivel, se puede destacar que:

- De manera general, las librerías tienen como funcionalidad principal la extracción de características de audio, sin embargo, la cantidad, la diversidad y el nivel de clasificación (de bajo nivel o de alto nivel) difieren entre sí.

- *jAudio* y *OpenSMILE* trabajan exclusivamente con características de bajo nivel, por lo que requieren de conocimientos previos sobre el análisis de señales digitales, para luego diseñar y calcular las características de alto nivel por fuera de la librería.
- No todas las librerías de alto nivel revisadas implementan modelos de clasificación. En algunos casos, el sistema de clasificación debe ser diseñado como una extensión de las librerías, como es el caso de *jAudio* y *OpenSMILE*.
- Las librerías utilizan modelos de clasificación emocional general (GMER) [82]. GMER consiste en aplicar una misma clasificación emocional de la música para todos los usuarios, y no una clasificación personalizada para cada usuario (PMER). El nivel de diferencia entre la clasificación general y la clasificación personal suele describirse a través de la característica *residual modeling*.
- El enfoque difuso (*fuzzy*) es implementado únicamente por AcousticBrainz. En este caso se realiza a través de *clusters*, indicando el nivel de pertenencia que una determinada canción tiene a cada *cluster* [82]. Por lo general, se da un vector como [0.1, 0.3, 0.9, 0.1] como salida del clasificador, donde cada posición en el vector se asocia con un *cluster* determinado, y cada *cluster* contiene una serie de emociones.

3.2 Datasets musicales

En este apartado se presentan y analizan algunos de los *datasets* existentes para realizar experimentos en los campos del reconocimiento de emociones de la música (MER), recuperación de información musical (MIR), y sistemas recomendadores musicales (MRS). Es importante resaltar, que a pesar de que en la actualidad existen algunos *datasets* disponibles, muchos de ellos presentan diferentes limitaciones que requieren ser comprendidas y consideradas en el diseño de un sistema recomendador musical [83] [84].

3.2.1 Revisión de Datasets

Para el proceso de revisión sistemática sobre *datasets*, se consideraron los artículos ya encontrados en las búsquedas que consideraban las palabras clave de las filas 3 y 4 de la Tabla 1.1 del apartado 1.5. Con estas palabras claves se armaron las cadenas de búsqueda. En la revisión del *abstract* de los trabajos se verificó la mención del uso de algún *dataset*, también se revisaron las palabras clave y los títulos de las secciones. Para el proceso de selección de análisis de *datasets* encontrados, se ha considerado el nivel de relevancia y utilización que éstos tienen en algunos de los congresos más reconocidos en los campos de investigación que son de interés para esta Tesis, tal es el caso del *Audio Mostly* y el *ISMIR Conference* (*International Society for Music Information Retrieval Conference*), como también por parte de comunidades científicas expertas, como sería el caso particular del MTG (*Music Technology Group*) de la Universitat Pompeu Fabra de Barcelona.

Se seleccionó un total de 9 *datasets* para su análisis. Este análisis se valora novedoso porque considera conjuntos de datos musicales en donde: se evalúan los archivos (clips) para identificar si son canciones completas con estructura musical etiquetada, se identifica la duración promedio de los archivos, se identifica el modelo afectivo utilizado, se identifica si los artistas no son superestrellas (noveles), se identifica si existe etiquetado emocional disponible por parte de artistas y oyentes. Cabe destacar que el tesista como parte de la metodología de análisis ha realizado algunos experimentos sobre los *datasets*. En particular sobre *MediaEval*, y es por esto, que este *dataset* se describe más en detalle, ya que en su estudio se detectaron ciertas características de interés, lo que dio lugar a realizar experimentos para analizar de manera práctica sus posibilidades.

A continuación se explica detalladamente cada uno de los criterios que aparecen en la Tabla 3.4:

- ***Dataset:*** El nombre con el que se conoce la base de datos musical (*dataset*).
- **Año:** El año en que fue generado el artículo científico que relaciona el *dataset*.
- **Archivos:** El número de archivos de audio cuyos *metadatos* se han incluido en el *dataset*.
- **Duración:** La duración media de los archivos de audio.
- **Audio disponible:** Si el conjunto de datos incluye los archivos de audio o no.
- **Estructura musical:** Si el conjunto de datos incluye o no *metadatos* relacionados con la identificación completa de la estructura musical de una canción, siendo la estructura más típica la introducción, el verso, el coro y el solo. La estructura musical permite realizar experimentos basados en la similaridad de las partes de la estructura, lo cual es muy útil teniendo en cuenta que una canción es una experiencia emocional que se produce a lo largo del tiempo.
- **Modelo afectivo:** El tipo de modelo afectivo utilizado, que puede ser categórico o dimensional.
- **Artistas no-famosos:** Si el conjunto de datos incluye canciones producidas por artistas noveles (*non-superstar*).
- **Etiquetado emocional por artista:** Si el artista ha etiquetado emocionalmente sus propias canciones o no.
- **Etiquetado emocional por oyente:** Si el oyente ha etiquetado emocionalmente las canciones o no.

Un hallazgo muy importante en los datasets GTZAN [85], Ballroom [86], Magna-TagATune [87], AudioSet [92], TUT Acoustic Scene [91], UrbanSound8k [89], ESC-50 [90] es la duración de los archivos de sonido, la cual es muy corta, y varía entre 1 y 30 segundos. En la mayoría de los casos no se dispone de canciones completas, sino de fragmentos de sonido que pueden corresponder a sonidos de ambiente (claxon de

Tabla 3.4: Revisión de *datasets* musicales

Dataset	Año	Archivos	Duración	Audio disponible	Estructura musical	Modelo afectivo	Artistas no-famosos	Etiquetado emocional por artista	Etiquetado emocional por oyente
GTZAN [85]	2002	1000	30s	✓	✗	None	✗	✗	✗
Ballroom [86]	2006	698	≈30s	✓	✗	None	✗	✗	✗
MagnaTagATune [87]	2009	25,850	≈30s	✓	✗	None	✗	✗	✗
Million Song Dataset [88]	2011	1M	-	✗	✗	Categórico	✗	✗	✓
UrbanSound8k [89]	2014	8732	⩽4s	✓	✗	None	✗	✗	✗
ESC-50 [90]	2015	2000	5s	✓	✗	None	✗	✗	✗
TUT Acoustic Scene [91]	2016	1560	30s	✓	✗	None	✗	✗	✗
Mediaeval [81]	2016	1744	≡45s	✓	✗	Dimensional	✗	✗	✓
		58	[46s, 627s]						
AudioSet [92]	2017	≈2.1M	10s	✗	✗	Categórico	✗	✗	✗

un carro, tráfico, animales, naturaleza, etc), o en el mejor de los casos, pequeños fragmentos de sonido que corresponden a una interpretación musical muy limitada. En el caso del *dataset* Million Song Dataset [88], la documentación no detalla con claridad la duración promedio de los archivos, y aunque es un conjunto de datos de canciones, estas corresponden a versiones de canciones famosas y no a canciones originales de artistas que no son superestrellas, además, los archivos de audio no están disponibles. El *dataset* MediaEval [81] resulta muy interesante por la cantidad de anotaciones emocionales que se incluyen, de momento es quizás una de las más completa para realizar experimentos de clasificación emocional. Sin embargo, no hay anotaciones para cada parte de la estructura de la canción, ni hay datos que hagan referencia a un análisis más profundo de la perspectiva del artista. El etiquetado emocional sólo se realiza en los primeros 45 segundos de cada clip (1744 clips tienen una duración de exactamente 45 segundos, mientras que los 58 clips restantes tienen una duración que varía entre 46 y 627 segundos). Una cosa más a destacar es que ninguno de los conjuntos de datos incluye canciones con una intención artística real y con interés en formar parte de la industria musical, por lo que no hay ningún tipo de anotación por parte de los artistas que permita un análisis más profundo desde su perspectiva, como por ejemplo la estructura de la canción, y su propia intención emocional relacionada con esa estructura musical.

3.2.2 Limitaciones relevantes

Teniendo en cuenta las consideraciones anteriores de los *datasets* musicales revisados, y desde el punto de vista de sistemas de recomendación musical, se identifican las siguientes limitaciones:

- En la mayoría de los *datasets* se incluyen fragmentos de sonido (clips) que generalmente no corresponden a canciones, y por lo tanto, no cuentan con una estructura musical.
- No hay un análisis en profundidad desde el punto de vista de la música que implique al artista para entender su intención emocional y sus técnicas de composición.
- La mayoría de los análisis que incluyen canciones de estructura musical completa, se centran en canciones comerciales y no en canciones originales de artistas

noveles, lo que genera un sesgo preexistente relacionado con la popularidad, anteriormente explicado en el apartado 2.10. Esta situación genera una alta inconformidad por parte de los artistas que no son súper estrellas, porque una gran mayoría de servicios de *streaming* de música recomienda con muy baja probabilidad sus obras musicales [54].

- No se tienen muchos *datasets* que incluyan canciones reales, y *metadata* asociada, que facilite a la comunidad científica avanzar hacía experimentos que permitan analizar algunos aspectos menos explorados en MER y MRS en la actualidad.

3.3 MediaEval Dataset

En este apartado se presenta las características del *dataset* de MediaEval [81]. Este *dataset* ha sido seleccionado para ser analizado en mayor profundidad, teniendo en cuenta que con respecto a todos las demás *datasets* revisadas anteriormente, MediaEval es el *dataset* con mayor información disponible y detallada en cuanto a etiquetado emocional, lo cual es un tema relevante para los objetivos definidos de esta Tesis. La base de datos de MediaEval contiene 1802 archivos en formato MPEG layer 3 (MP3), con una frecuencia de muestreo de 44100Hz, y para cada una de ellas se dispone de 260 características de bajo nivel (*low-level features*). Cada canción es analizada durante 45 segundos, y el valor de cada característica de sonido se extrae cada 0.5 segundos (500 ms). El proceso de extracción de las características de sonido fue realizado a través del software *openSMILE*[2]. El proyecto MediaEval[3] también incluye los archivos de sonido, y adicionalmente, también incluye archivos de extensión CSV (*comma-separated values*) en donde se puede encontrar evaluaciones de percepción emocional para cada una de las canciones. El proceso de evaluación emocional fue llevado a cabo aplicando un modelo emocional dimensional, lo que implica establecer un valor de *valence* y *arousal* (V/A). Adicionalmente, el proceso de etiquetado emocional fue realizado con un enfoque dinámico (*over-time*), lo que significa que cada 500 ms (para este caso) el usuario indicó su percepción emocional estableciendo una coordenada V/A. Para mayor información sobre el etiquetado emocional dinámico, en el anexo D. Análisis del sistema de etiquetado en MediaEval se puede encontrar la revisión de algunos casos puntuales del *dataset*.

La cantidad y distribución de archivos de sonido sobre el plano bidimensional emocional V/A se muestra en la Figura 3.3; dicha distribución fue generada a partir de la ubicación de las coordenadas medias V/A de cada canción dentro de un plano de dos dimensiones, generando así una clasificación por cuadrantes. Es importante resaltar que se evidencia un desbalanceo de datos para los diferentes cuadrantes, por ejemplo, Q1 tiene 886 canciones, mientras Q2 tan solo tiene 218 canciones. Esta condición del *dataset* se debe analizar en profundidad durante los diferentes experimentos que se desarrollarán en esta Tesis.

[2]https://www.audeering.com/research/opensmile/
[3]http://www.multimediaeval.org/datasets/

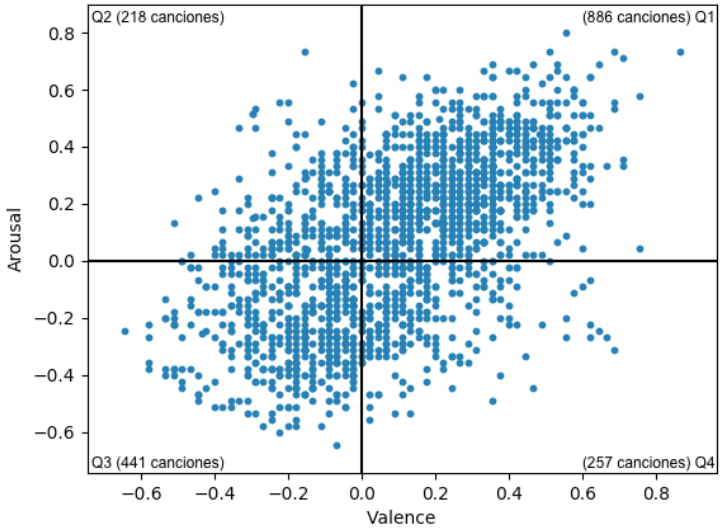


Figura 3.3: Distribución de las canciones de MediaEval en un espacio dimensional V/A. Elaboración propia [6].

3.4 Sistemas de predicción

En este apartado se presentan y analizan algunos trabajos enfocados en el reconocimiento de emociones en la música a través de modelos predictivos de valor aproximado. En general, el problema del reconocimiento de emociones en la música puede ser analizado bajo el enfoque de modelos predictivos, teniendo en cuenta, que una canción se describe por un conjunto de características de sonido que pueden ser de alto y bajo nivel, y dichas características se pueden relacionar con la percepción emocional (*valence* y *arousal*) de los oyentes. De este modo, es posible entrenar un modelo predictivo a partir de un conjunto de datos previamente anotado, y luego evaluar los resultados de las predicciones para determinar el desempeño del modelo; una forma de hacerlo es a través de la comparación entre los valores indicados por el usuario, y los valores predichos por el algoritmo de predicción. Esa comparación permite calcular los niveles de error, para luego evaluar estrategias adicionales que permitan minimizar estos mismos. Es muy importante diferenciar los sistemas de predicción de clases (sistemas clasificadores) de los sistemas de predicción de valores aproximados; en el primer caso se dispone de una matriz de confusión, en la que se especifican las métricas de exactitud, precisión y recuerdo (lo que se explica en el apartado 2.8). En el segundo caso, la tasa de éxito suele analizarse mediante el error medio absoluto (MAE), el error cuadrático medio (RMSE) y el coeficiente de determinación (R^2) (lo que se explica en el apartado 2.6).

Para la búsqueda de trabajos se siguió el protocolo de revisión sistemática de literatura con las palabras clave de la fila 3 de la Tabla 1.1. Para el proceso de selección se consideraron dos criterios: por una parte, la utilización de un modelo emocional dimensional para el etiquetado de emociones, y por otra parte, la implementación de modelos de *machine learning* para realizar las predicciones [6]. Así en la Tabla 3.5 se

presenta un total de 4 artículos que cumplían con estos criterios de inclusión, y que permiten atender al objetivo 3 de la Tesis. La Tabla 3.5 resume el tamaño del conjunto de datos, el número de características, las técnicas de predicción y las tasas de éxito. Las técnicas utilizadas en estos trabajos son: *support vector machine* (SVM), *deep long-short term memory recurrent neural networks* (LSTM-RNN), *bi-directional long short-term memory recurrent neural networks* (BLSTM- RNNs) y *adaptive aggregation of gaussian process regressors* (AAGPR). Estos trabajos utilizan un enfoque de modelo dimensional, que ubica la emoción como una coordenada en dos ejes: *valence* (qué tan positiva o negativa es la emoción) y *arousal* (para determinar la intensidad como nivel de energía, y/o excitación).

Tabla 3.5: Comparación de algunos trabajos relacionados

Artículo	Canciones	Características	Técnica de Predicción	RMSE
Fernandes [17]	194	554	SVM	V:0.24 , A:0.22
Coutinho et al [93]	744	260	LSTM-RNN	V:0.13 , A:0.15
Xu et al [94]	489	65	BLSTM-RNNs	V:0.46 , A:0.34
Fukayama et al [95]	744	65	AAGPR	V:0.24 , A:0.22

Es importante destacar que los conjuntos de datos tienen diferentes tamaños (con respecto a las canciones y las características), las técnicas de clasificación son diversas, y los valores de RMSE varían entre 0.13 y 0.46 para *valence*, y 0.15 y 0.34 para *arousal*. Esto sugiere que las diferentes configuraciones y condiciones experimentales tienen una fuerte relación con las tasas de éxito.

3.5 Sistemas no deterministicos (*Fuzzy)*

El problema de la clasificación en los sistemas MER puede resolverse mediante dos grandes paradigmas: el determinista y el no determinista. En un modelo determinista (según la teoría de conjuntos tradicional), un elemento puede ser miembro de diferentes conjuntos con un grado de pertenencia absoluto, es decir, pertenece o no pertenece. En los modelos no deterministas, un elemento puede ser miembro de varios conjuntos, con diferentes grados de pertenencia entre 0 y 1 [46]. En general, la mayoría de los trabajos científicos de clasificación en el campo del MER se basan en el paradigma determinista, y aplican técnicas centradas en el enfoque de *machine learning* [96]. Este enfoque de solución requiere un conjunto de datos pre-etiquetados. El modelo de clasificación predecirá una clase (emoción) [4], y su éxito dependerá en gran medida del proceso de entrenamiento, en el que se deben analizar las métricas correspondientes a la técnica elegida para mejorar la eficacia del proceso de clasificación. En la mayoría de los casos, este tipo de modelo se implementa con redes neuronales [97].

La clasificación musical por emociones también puede analizarse como un problema de reglas lógicas, a través de un paradigma no determinista, el cual considera la subjetividad de la percepción humana. Los psicólogos han descubierto que los valores de algunas características de alto nivel están relacionados con la percepción emocional de los usuarios; por ejemplo, la mayoría de los oyentes suelen percibir emociones positivas para un tempo superior a 70 pulsaciones por minuto (bpm) [8]. A pesar de que las

reglas no son claras ni universales, existe un consenso sobre cómo algunos valores de las características musicales evocan algunas emociones particulares. Si se dispone de los conocimientos que explican la relación entre los distintos valores de las características de la música y la emoción transmitida, sería posible crear todas las reglas de inferencia para clasificar una pieza musical dentro de un modelo emocional; aunque este escenario implica un esfuerzo mucho mayor en comparación con el enfoque de *machine learning*, en el que las reglas se descubren en un proceso de aprendizaje iterativo y automático. Esta forma particular de resolver un problema de clasificación se implementa típicamente mediante sistemas difusos, cuya base conceptual y arquitectura fue explicada en el apartado 2.7.

Con el principal objetivo de analizar las aportaciones de la lógica difusa en MER, se llevó a cabo una búsqueda con las palabras clave de la fila 3 de la Tabla 1.5. Luego se seleccionaron aquellos trabajos que aplicaran lógica difusa para las estrategias de etiquetado y de clasificación de emociones. Es importante aclarar que en el caso de la estrategia de clasificación, sólo se marcan los trabajos que definen reglas de inferencia difusa. La mayoría de estos trabajos, aunque clasifican, lo hacen con el enfoque de *machine learning* y no con el de lógica difusa. la Tabla 3.6 muestra los trabajos que fueron seleccionados bajo el criterio mencionado.

Tabla 3.6: Trabajos con lógica difusa aplicada en MER

	Estrategias MER	
Artículo	Etiquetado emocional	Clasificación
Yang et al [98]	✓	-
Jun et al [99]	✓	✓
Zhu et al [100]	✓	-
Yang et al [82]	✓	-
Naji et al [101]	✓	-
Huang et al [102]	✓	-

La importancia de la subjetividad en la percepción humana cuando un oyente intenta clasificar la música emocionalmente se destaca en [98]. Este trabajo utiliza un conjunto de datos de 243 canciones e implementa el clasificador Fuzzy k-NNN (FKNN) y el Fuzzy Nearest-Mean (FNM). Una canción no siempre tiene una única clasificación para todos los oyentes, por lo que, en este caso, es importante manejar varias clases con diferentes grados de pertenencia. El trabajo de Yun et al. [99] presenta un prototipo para el reconocimiento de emociones implementado con MATLAB. Haciendo hincapié en la importancia de la semántica utilizada y en la información imprecisa que indican las emociones, propone el motor de inferencia difusa de *Mamdani* con 24 reglas que han sido diseñadas mediante el conocimiento de expertos. La retroalimentación de los oyentes respecto a las emociones percibidas y el ajuste de las reglas del sistema de inferencia en base a la retroalimentación es una contribución relevante de este trabajo.

En [100], los autores definen un vector de emociones de la música (cada posición del vector representa el grado de pertenencia a una emoción) para diseñar un sistema de

exploración para realizar búsquedas de canciones de forma difusa. La complejidad de la subjetividad en la percepción emocional también se menciona en [82], y se propone la lógica difusa como estrategia para obtener resultados menos deterministas. En [101] se aplica la lógica difusa como estrategia para reducir el número de características. En cuanto al proceso de clasificación, se trabaja con una red neuronal en cascada (CFNNS), que es una técnica similar al perceptrón multicapa (MLP). Mientras que la percepción emocional del oyente suele ser el foco principal en los diferentes trabajos de MER, [102] presenta una interesante contribución al permitir a los compositores etiquetar sus piezas musicales emocionalmente mediante el uso de marcadores de expresión. El enfoque de la lógica difusa se aplica a través de diferentes adjetivos utilizados para las marcas emocionales. En este caso, el conjunto de emociones se representa como un conjunto difuso y, por tanto, se aplican estrategias de fusificación y defusificación. La información de los compositores y de los oyentes se capta a través de cuestionarios y luego se compara.

Es importante señalar que aunque todos los trabajos revisados anteriormente [98] [99] [99] [100] [82] [101] [102] aplican la lógica difusa a la estrategia de etiquetado emocional de las canciones, sólo uno de ellos [99] considera la estrategia de clasificación con procesos de inferencia completamente basados en reglas. Esto motiva un análisis sobre las posibles ventajas y desventajas de la lógica difusa, para ambas estrategias. Según la revisión bibliográfica del MER centrada en la aplicación de la lógica difusa, es importante destacar los siguientes hechos:

- En la mayoría de los trabajos revisados, la lógica difusa se utiliza ampliamente para el proceso de etiquetado de las emociones en las canciones. Hay dos razones principales: la ambigüedad en la determinación de las emociones a través del lenguaje humano, y la diferente percepción emocional de diferentes oyentes para la misma canción.

- Existen muy pocos trabajos que implementen sistemas de clasificación basados únicamente en motores de inferencia difusa. Aunque es posible formular reglas lógicas para realizar inferencias difusas, hay que destacar la alta complejidad operativa del proceso de inferencia que vendrá determinada por la cantidad de características de sonido incluidas en las reglas lógicas, así como por todos los condicionales necesarios según los criterios de clasificación.

- La capacidad de aprendizaje de las redes neuronales no está disponible en el enfoque de la lógica difusa. Aunque es posible diseñar un sistema de reglas difusas, se requiere un enorme esfuerzo para extraer los conocimientos de los usuarios finales y hacerlos parte del sistema difuso; además, sería necesario analizar el diseño de un verdadero proceso dinámico sobre la retroalimentación de las reglas, esto con el fin de mejorar constantemente el sistema de clasificación difuso.

- El diseño de reglas de inferencia, con umbrales claramente definidos para el manejo de datos difusos, contribuye a obtener buenos resultados en términos de clasificación. En el *machine learning*, los umbrales de clasificación se aprenden, pero

en este proceso, diversos aspectos como el ruido de los datos, los datos desequilibrados, el sobreajuste y otras cuestiones, afectan a las métricas de rendimiento de los procesos de clasificación.

- En el enfoque de los sistemas difusos, el conocimiento que se codifica en las reglas de inferencia pertenece a un experto concreto. Para desarrollar clasificadores personalizados, en los que las reglas de inferencia pueden variar (como es el caso de la música), es necesario diseñar un sistema de inferencia completamente independiente para cada usuario.

3.6 Sistemas de clasificación determinísticos

Este apartado se concentra en la revisión de trabajos de MER basados en modelos de clasificación determinísticos, en los que normalmente, las canciones son clasificadas con grados de pertenencia absolutos (pertenece o no pertenece) a las diferentes clases existentes dentro del modelo, para este caso, emociones particulares. Para la búsqueda de trabajos se siguió el protocolo de revisión sistemática de literatura con las palabras clave de la fila 3 de la Tabla 1.1. Se vuelve a señalar los trabajos de predicción estudiados en el apartado 3.4 teniendo en cuenta que algunos de ellos son extendidos a trabajos de clasificación determinística. En particular, para los intereses de este apartado, se seleccionaron aquellos trabajos que cumplieran con enfocarse en MER y que se basaran en modelos de clasificación determinísticos, filtrando un total de 6 trabajos que son presentados en la Tabla 3.8.

En la Tabla 3.7 se listan trabajos que corresponden a predicción, y en la Tabla 3.8 se listan trabajos que corresponde a clasificación. Estas tablas muestran las características más importantes de los conjuntos de datos (*datasets*) utilizados: número de canciones, duración de cada canción, número de características de sonido, estado de balanceo de los datos y tipo de anotación (estática o dinámica), también se específica las técnicas de clasificación o predicción y sus respectivas métricas de evaluación. En el caso de los trabajos relacionados con sistemas de clasificación (Tabla 3.8), también se especifican las clases/categorías utilizadas en el trabajo: cuadrante, *cluster* y, en algunos casos, media de todos los cuadrantes.

Algunos trabajos se centran sólo en realizar predicciones como [17], [103], [44] o [104], otros trabajos se enfocan sólo en resolver problemas de clasificación como [16], [78], [105]; también hay varios trabajos que realizan predicciones y posteriormente extienden sus sistemas para lograr clasificaciones, como [106] o [107]. Generalmente, estos trabajos que incluyen predicción y clasificación utilizan conjuntos de datos anotados en modelos dimensionales, en los que se establece un sistema de coordenadas de *valence* y *arousal* (V/A) [4]. Además, algunos de estos trabajos implementan una fase de pre-procesamiento de los datos, pero únicamente con el objetivo de identificar clases (cuadrantes, *clusters*, otros), aunque esto puede ser innecesario si el *dataset* también se encuentra anotado categóricamente con un conjunto discreto de emociones [37].

Tabla 3.7: Sistemas de predicción emocional. Métricas de referencia: root-mean-square error (RMSE), averaged random distance (ARD), determination coefficient (R^2).

Artículo	Canciones	Duración	*Features*	Balanceado	Anotación	C. técnica	RMSE	ARD	R^2
Fernandes [17]	194	25 s	454	No	Estático	SVM	(V:0.24, A:0.22)	-	-
Schmidt [106]	240	15 s	-	-	Dinámico	SVR	-	0.238	-
Panda [103]	189	25 s	556	No	Estático	SLR, KNN, SVR	-	-	(V:40.6 %, A:67.4 %)
Bai [107]	744	45 s	548	-	Estático	SVR, RFT, PCA	-	-	(V:29.3 %, A:62.5 %)
Grekow [44]	324	6 s	654	Sí	Estático	SMOreg	-	-	(V:58 %, A:79 %)
Hennequin [104]	18,644	30 s	-	-	Estático	ConvNet	-	-	(V:17.9 %, A:23.5 %)

Tabla 3.8: Sistemas de clasificación emocional. Métricas de referencia: *Accuracy*, *F-measure*.

Artículo	Canciones	Duración	*Features*	Balanceado	Anotación	C. técnica	Clases	*Accuracy*	*F-measure*
Schmidt [106]	240	15 s	-	-	Dinámico	SVR	4 cuadrantes	50.18 % (promedio)	-
Panda [16]	903	30 s	253	Sí	Estático	SVM	5 grupos	-	[0.37, 0.37, 0.61, 0.40, 0.53]
Grekow [78]	324	6 s	471	Sí	Estático	SMO	4 cuadrantes	[0.81, 0.90, 0.87, 0.77]	[-,0.72, 0.65, 0.54]
Zhang [105]	400	35 s	-	Sí	Estático	RFC	4 cuadrantes	83.29 % (promedio)	-
Bai [107]	744	45 s	548	-	Estático	SVR RFT PCA	4 cuadrantes	59.2 % (promedio)	-
Panda [108]	900	30 s	898	No	Estático	SVM	4 cuadrantes	-	[0.77, 0.85, 0.71, 0.68]

El tipo de etiquetado de las emociones es otro criterio importante de comparación, así como también un factor clave de éxito en el campo de MER. Es importante resaltar que ninguno de los trabajos mencionados en la Tabla 3.7 y Tabla 3.8 tienen en cuenta la intención emocional del compositor, en lugar de ello, el etiquetado emocional disponible siempre es el del oyente. Siqi Huang en [102] resalta la escasez de modelos de clasificación musical, tanto discretos como dimensionales, que consideren en su diseño el punto de vista del compositor, en particular, la intención emocional. Entre estos trabajos, existen principalmente dos enfoques para anotar las canciones: estático y dinámico. En el proceso de anotación estático el usuario establece un valor de *valence* y otro de *arousal* para indicar su percepción emocional, más relacionada con la respuesta anímica a la música, este etiquetado normalmente se hace para toda la canción. En un proceso de anotación dinámico, el usuario genera una evaluación temporal dinámica sobre su percepción emocional, de forma continua, lo que quiere decir que etiqueta emocionalmente la canción durante toda la experiencia de apreciación musical. Por ejemplo, Schmidt *et al* utilizó la anotación dinámica, con una ventana de tiempo que varía de 2 a 15 segundos [106]. Aunque la mayoría de los trabajos de predicción y clasificación emocional se basan en anotación estática [17] [107] [44] [104], existen algunos trabajos que aplican la anotación dinámica temporal, pero esta anotación sobre el tiempo (*over-time*) se promedia posteriormente y, por tanto, al final se tiene un único valor de *valence* y *arousal*. Este es el caso del trabajo de Panda *et al* [103], en el que se implementa la anotación dinámica, pero estas anotaciones se promedian y se utilizan como un único valor global dentro del modelo de predicción, por lo que el trabajo finalmente se clasifica como "anotación estática".

Además del enfoque dinámico o estático que normalmente se define en los procesos de etiquetado, para obtener buenos resultados en un sistema de clasificación y, en particular, en la clasificación de piezas musicales por emociones, es fundamental contar con un gran número de datos de piezas musicales distribuidas de manera balanceada para cada emoción. Este balanceo de los datos es una cuestión fundamental en los sistemas

de clasificación, porque impacta los valores obtenidos en las tasas de éxito [84]. En general, en la mayoría de los sistemas de predicción, la cantidad de datos es un criterio muy importante porque determina su capacidad de generalización. Sin embargo, si la disponibilidad de datos por anotación categórica no está distribuida de manera balanceada, las clases minoritarias pueden verse afectadas en el desempeño de un sistema de clasificación, porque las técnicas de *machine learning* tienden a especializarse en la predicción de las clases mayoritarias.

La información sobre el balanceo de los datos en los trabajos estudiados se muestra también en las Tablas 3.7 y 3.8. El valor de la columna *Balanceado* indica el estado original del balanceo de datos antes de cualquier fase de pre-procesamiento. Se puede observar que en algunos trabajos no se incorporan detalles sobre el balanceo de datos, especialmente en los sistemas de predicción (ver Tabla 3.7) en los que no tiene mayor relevancia. En cuanto a los sistemas de clasificación (ver Tabla 3.8), los datos de los trabajos [16], [78] y [105] están balanceados y no requieren ningún tratamiento especial. En cambio, [108] presenta datos desbalanceados, y aunque no se detalla la estrategia utilizada para balancear los datos, aparentemente se eliminan algunas canciones del conjunto de datos original. Y, por último, hay otros trabajos sin información sobre el equilibrio de los datos [106] [107].

En cuanto a los resultados de los trabajos revisados de MER, se puede observar que tanto para la predicción (Tabla 3.7) como para la clasificación (Tabla 3.8) las tasas de éxito varían de valores bajos a medios. Sólo hay unos pocos trabajos que tienen valores más altos, pero es difícil sacar alguna conclusión, debido a la falta de uniformidad en los conjuntos de datos y a la naturaleza desbalanceada de la mayoría de los conjuntos de datos disponibles.

3.7 Sistemas recomendadores musicales

El proceso de búsqueda y selección para conformar este apartado tuvo en cuenta las palabras clave de la fila 4 de la Tabla 1.1. Para la selección se consideraron trabajos que presentaran sistemas recomendadores musicales. Un total de 18 trabajos fueron incluidos, ya que se focalizaban en esta temática, estos son presentados en la Tabla 3.9, en donde se describen las estrategias de recomendación implementadas o mencionadas, y la relación entre cada trabajo y la estrategia de recomendación. Antes de analizar en detalle cada trabajo y sus estrategias, es muy importante destacar cómo la selección de las estrategias de recomendación varía de un trabajo a otro; lo que se considera un primer hallazgo que motiva a discutir y entender la conveniencia de cada estrategia en el campo de los MRS. También es importante resaltar que la mayor parte de la literatura de MRS se centra en la experiencia del usuario oyente, hay algunos trabajos que destacan la importancia de analizar la experiencia del usuario artista, que en general viene determinada por las posibilidades reales de impulsar una carrera comercial [54].

Tabla 3.9: Revisión de literatura de MRS.
CF: Filtrado colaborativo , DF: Filtrado demográfico, CBF: Filtrado basado en contenido, HF: Filtrado híbrido, UC: Contexto de usuario, MD: Metadata, EBF: Filtrado basado en emociones, PA: Enfoque personalizado, PLB: Basado en listas de reproducción, PB: Basado en popularidad, SB: Basado en similaridad, IB: Basado en interacción

		Estrategias MRS											
Año	**Artículo**	**CF**	**DF**	**CBF**	**HF**	**UC**	**MD**	**EBF**	**PA**	**PLB**	**PB**	**SB**	**IB**
2020	Shah et al [47]	✓	-	-	-	✓	-	-	-	-	-	-	-
	Paul et al [52]	-	-	✓	-	✓	✓	✓	✓	-	-	-	-
2019	Zheng et al [109]	-	-	-	-	-	-	-	✓	✓	-	-	-
	Fessahaye et al [110]	✓	-	✓	-	-	✓	-	-	✓	-	-	-
	Yucheng et al [53]	-	-	-	-	✓	-	✓	-	-	-	-	-
	Bauer et al [59]	-	✓	-	-	-	-	-	-	-	✓	-	-
	Andjelkovic et al [56]	-	-	-	-	-	-	✓	-	-	-	✓	✓
	Chen et al [111]	✓	-	-	-	-	-	-	-	-	-	-	-
	Ferraro et al [58]	-	-	-	-	-	-	-	-	-	-	-	✓
	Katarya et al [60]	✓	-	-	✓	✓	-	-	-	-	-	-	✓
2018	Garcia-Gathright et al [57]	-	-	-	-	-	-	-	-	✓	-	-	✓
	Deshmukh et al [51]	✓	-	✓	✓	✓	✓	✓	✓	-	-	-	-
	Schedl et al [112]	-	✓	-	-	✓	-	✓	✓	✓	✓	-	✓
2017	Bauer et al [54]	-	-	✓	-	-	✓	-	-	-	✓	-	-
2016	Cheng et al [113]	-	-	-	✓	-	-	-	✓	-	-	-	-
	Vigliensoni et al [55]	-	✓	-	-	✓	-	-	-	-	-	-	✓
	Katarya et al [114]	-	-	-	-	✓	-	✓	-	-	-	-	-
2013	Bobadilla et al [115]	✓	✓	✓	✓	-	-	-	-	-	-	-	-

Entre las diferentes estrategias de recomendación, el filtrado colaborativo es probablemente una de las más utilizadas, quizás por su sencillez técnica frente a otros modelos más sofisticados. Sin embargo, es imprescindible disponer de una comunidad digital y de un flujo representativo de información fiable para garantizar un rendimiento mínimamente óptimo. Por el contrario, los *metadatos* son una de las estrategias menos utilizadas en los trabajos revisados, y esto ocurre fundamentalmente cuando dichos *metadatos* no se construyen automáticamente. Hay algunos casos en los que los *metadatos* se generan a través de un proceso de filtrado basado en el contenido, en el que, por ejemplo, se extraen automáticamente las características del sonido de una canción y luego, con modelos previamente entrenados, se determina el género musical, las emociones evocadas, entre otros. En estos casos, se trata de una estrategia de filtrado híbrida y no puramente de *metadatos*. Aunque todavía existen implementaciones de CF en trabajos recientes [47], [110], [111], [60], todos ellos implican estrategias adicionales, en este sentido, estos trabajos implementan una estrategia de filtrado híbrido.

Resulta muy interesante notar que los trabajos más recientes [47], [52], [109], [110] exploran estrategias de enfoque personalizado, filtrado basado en emociones, filtrado basado en el contenido, y contexto del usuario. Se evidencia una tendencia de implementar un enfoque personalizado para mejorar los MRS, que suele basarse en técnicas de *machine learning*, ya que permite diseñar estrategias más dinámicas para generar recomendaciones a través de un proceso de aprendizaje. Este proceso de aprendizaje es probablemente la ventaja más importante del enfoque de *machine learning* en contraste

con los sistemas tradicionales que implementan reglas fijas, y podría ser utilizado para reconocer emociones, y hacer predicciones basadas en el contexto del usuario. A pesar de las contribuciones del enfoque de *machine learning* a las estrategias más recientes, hay evidencia de insatisfacción desde el punto de vista de los oyentes y artistas [116], y la razón de esta insatisfacción está relacionada con los sesgos, por lo tanto, este hallazgo motiva a explorar en profundidad el impacto de los sesgos en MRS en el siguiente apartado 3.8.

La importancia de lograr una comparación objetiva entre los MRS también ha despertado un gran interés por definir un adecuado proceso de evaluación. Sin embargo, la evaluación de los sistemas de recomendación musical es algo realmente difícil de definir, debido a que en la mayoría de los casos dependerá de los intereses particulares de los actores que intervienen en la industria musical, las estrategias de recomendación implementadas, y entre otras cosas más [112]. Teniendo en cuenta la tendencia a utilizar estrategias de recomendación basadas en el campo de *machine learning*, muchas de las métricas de los MRS relacionadas con la novedad, o serendipia de un elemento, se definen en términos de métricas de evaluación comúnmente utilizadas en ese campo, como la exactitud, la precisión, el recuerdo y el error cuadrático medio. En los últimos años han surgido algunas medidas novedosas para el campo de los sistemas recomendadores musicales, y estas medidas, denominadas *más allá de la exactitud* (*beyond-accuracy*), manejan particularidades de los MRS como la utilidad, la novedad o la serendipia de un elemento [117].

3.8 Tratamiento de sesgos en MRS

En su definición más general, el término sesgo implica en la mayoría de los casos una discusión relacionada con el trato injusto, considerando que en algunas situaciones, ciertos actores y/o elementos tienen mayores oportunidades que otros. Este trato desigual promueve condiciones perjudiciales que incluso actualmente motivan discusiones morales [61]. De acuerdo a la fundamentación teórica expuesta en el apartado 2.10 existen tres tipos de sesgos: preexistentes, técnicos y emergentes. Los sesgos preexistentes están relacionados con predisposiciones sociales y de contexto. Los sesgos técnicos tienen relación con los algoritmos y los datos. Y los sesgos emergentes, son aquellos que pueden surgir durante la utilización del sistema recomendador. Para el abordaje de este apartado se trabajó con los artículos resultantes de la búsqueda específica con las palabras clave de la fila 4 de la Tabla 1.1, y se seleccionó un total de 9 trabajos de acuerdo a su aporte novedoso que permite llevar a cabo una discusión sobre sesgos en el campo de MRS. Estos trabajos se incluyen en la Tabla 3.10, en donde para cada uno se identifica el tipo de sesgo y su relación con cada estrategia de recomendación involucrada. A partir de esta información, se revelan las siguientes conclusiones:

- En general, las estrategias de recomendación más afectadas por los sesgos son el filtrado colaborativo (CF) y el basado en la popularidad (PB).
- Siete de los nueve trabajos hablan de sesgos preexistentes, lo que sugiere la importancia de este tipo de sesgo.
- Únicamente un trabajo analiza el sesgo emergente, siendo el sesgo menos estudiado.

Tabla 3.10: Revisión bibliográfica de sesgos para trabajos de MRS.
CF: Filtrado colaborativo , DF: Filtrado demográfico, CBF: Filtrado basado en contenido, MD: Metadata, PA: Enfoque personalizado, PB: Basado en popularidad, SB: Basado en similaridad

			Estrategia MRS						
Año	**Artículo**	**Sesgo**	**CF**	**DF**	**CBF**	**MD**	**PA**	**PB**	**SB**
2020	Perera et al [118]	Preexistente	-	-	-	-	-	✓	-
	Melchiorre et al [119]	Preexistente	✓	✓	-	✓	-	-	-
		Emergente	✓	-	-	-	-	-	-
	Abdollahpouri et al [116]	Preexistente	-	-	-	-	-	✓	-
		Técnico	-	-	-	-	-	✓	-
	Sánchez-Moreno et al [120]	Preexistente	✓	-	-	-	-	✓	-
		Técnico	✓	-	-	-	-	✓	-
	Patil et al [121]	Técnico	-	-	✓	-	-	-	-
	Abdollahpouri et al [122]	Preexistente	✓	-	-	-	-	✓	-
		Técnico	✓	-	-	-	-	✓	-
	Shakespeare et al [123]	Preexistente	✓	-	-	-	-	✓	-
		Técnico	✓	-	-	-	-	✓	-
2019	Ferraro et al [124]	Preexistente	-	-	-	-	✓	✓	-
		Técnico	-	-	-	-	✓	✓	-
2018	Flexer et al [125]	Técnico	-	-	✓	-	-	-	✓

Es importante destacar algunos hallazgos individuales para cada trabajo incluido en la Tabla 3.10, porque ayudan a comprender mejor cómo operan los sesgos sobre las estrategias de recomendación en algunos casos de estudio específicos. Existe un sesgo preexistente en [118] relacionado con la estrategia basada en la popularidad (PB). A pesar de que la calificación de las canciones está influenciada por las estrategias de marketing, los datos de calificación de las canciones se utilizan para tratar el problema del arranque en frío. En [119], se implementan ambas estrategias, el filtrado demográfico (DF) y el filtrado colaborativo (CF) con sesgo preexistente. En este caso, el sesgo preexistente se genera debido a un proceso de minería de datos, en el que la red social Twitter es la fuente principal. En la mayoría de los casos, estos datos son incompletos y poco fiables debido a la incoherencia que se suele presentar entre la personalidad del usuario deducida de Twitter y el comportamiento real del usuario. Existe una relación muy estrecha entre el comportamiento de la sociedad y Twitter, cualquier cambio en la sociedad también cambiará los datos de Twitter, y cualquier sistema que dependa de estos datos se verá afectado en tiempo real, lo que constituye la principal característica

de un sesgo emergente. En [116] se destaca un sesgo preexistente en la estrategia de filtrado basado en la popularidad (PB), ya que la calificación de las canciones está influenciada por las estrategias de marketing, como resultado, el número de veces que se reproducen las canciones (contadores de reproducciones) por los oyentes tiende a aumentar. Los algoritmos de recomendación implementados por MRS utilizan los contadores de reproducciones como entrada principal, estos algoritmos no implementan ninguna acción para mitigar el efecto de popularidad, por lo que también promueven un sesgo técnico. La explotación de la información social de las redes sociales es una cuestión clave en [120] para implementar una estrategia de filtrado colaborativo (CF) basada en el algoritmo de similitud de vecinos, que promueve un sesgo preexistente. Los vecinos se encuentran en función de la similitud de las valoraciones de los usuarios considerando únicamente las mismas canciones valoradas por ambos vecinos, que suelen ser las más famosas de la industria musical, y esto revela que también se aplica la estrategia basada en la popularidad (PB). Cualquier otra canción que no sea común entre los usuarios es descartada por el algoritmo de similitud de vecindad, aunque exista la posibilidad de que les gusten estas canciones. Este proceso de descarte genera un sesgo técnico.

El análisis presentado en [121] se centra en los sesgos técnicos, especialmente en los algoritmos basados en modelos matemáticos como la descomposición del valor singular, la clasificación personalizada bayesiana, los autocodificadores y el *machine learning*. Estos algoritmos, típicamente implementados con filtrado basado en el contenido (CBF), incluyen un cierto nivel de ruido en sus capas internas, lo que impacta en sus índices de precisión (*accuracy rates*), y en la mayoría de los casos, impacta negativamente en las expectativas del usuario. Este trabajo no analiza ningún sesgo preexistente ni las posibles relaciones entre el sesgo preexistente y el sesgo técnico. Este hecho podría ser una debilidad desde el punto de vista de muchas partes interesadas, como los artistas, los oyentes, los desarrolladores de software y otros, ya que no existe una visión más detallada del problema que permita comprender los impactos reales en el modelo de negocio de la industria musical. Según [122] en los sistemas de recomendación un pequeño número de ítems aparecen frecuentemente en los perfiles de los usuarios, en cambio, un número mucho mayor de ítems menos populares aparecen muy raramente. Este sesgo tiene su origen en dos fuentes diferentes: los datos y los algoritmos. En el caso de los datos, el proceso de calificación se basa en el grado de fama de cada artista (sesgo preexistente), lo que genera una tendencia al desequilibrio de los datos de calificación. En cuanto a los algoritmos, no están diseñados para tratar la naturaleza desequilibrada de los datos de calificación, por lo que recomiendan en exceso los ítems populares (sesgo técnico), perjudicando al mismo tiempo las posibilidades de aumentar la popularidad de los ítems menos populares. La discriminación por género, con raíces en factores socioculturales, es el principal foco de atención del estudio de sesgos presentado en [123]. Existe una distribución muy desequilibrada por género según el proceso de análisis realizado sobre los conjuntos de datos LFM-1b y LFM-360k [126] [127], de manera que los artistas de género masculino constituyen la mayoría (82 %) de los artistas para los que se puede identificar el género.

Ferraro [124] explica cómo el problema del arranque en frío en muchos casos se trata mediante estrategias basadas en los índices de popularidad. Esta información de

calificación depende fundamentalmente de datos extraídos de las redes sociales, que amplifican un sesgo preexistente. Asimismo, Ferraro propone realizar un análisis más profundo sobre la perspectiva del usuario, implementando una evaluación centrada en el usuario que permita optimizar un problema multiobjetivo con una perspectiva multistakeholder, que ayude a mitigar posibles sesgos técnicos. En [125] se propone la responsabilidad ética de producir sistemas justos e imparciales como un nuevo reto para la comunidad de la minería de datos, destacando la importancia de revisar y mejorar las condiciones de los conjuntos de datos, así como el diseño de los algoritmos en el campo de *machine learning*. En este trabajo, las canciones más recomendadas se denominan canciones *hub*, y las menos recomendadas o nunca recomendadas se denominan *anti-hubs*. Este proceso de clasificación de las canciones es consecuencia de una debilidad promovida por las estrategias de *clustering* a través de un enfoque de *machine learning* no supervisado, en el cual, las canciones más recomendadas son las más cercanas a un centro de *cluster* específico, mientras que las canciones menos recomendadas son las más alejadas con respecto al mismo centro del *cluster*.

A continuación se presenta un conjunto de recomendaciones para el tratamiento de los sesgos en sistemas recomendadores musicales. Estas recomendaciones se han propuesto teniendo en cuenta los hallazgos discutidos anteriormente. Las recomendaciones que se detallan a continuación se han debatido a lo largo de este apartado y formulado para cada tipo de sesgo (preexistente, técnico y emergente).

En cuanto a los sesgos preexistentes, es importante destacar que en un proceso tradicional de desarrollo de software, los requerimientos funcionales se basan en un modelo de negocio específico, en este caso, el de la industria musical. El modelo de negocio define las reglas de negocio, en donde normalmente se consideran los intereses y objetivos de todos los actores interesados (*stakeholders*) [128]. En general, cualquier producto final debe responder a las expectativas de los actores interesados, por lo que si alguno de ellos no es invitado a participar en el proceso de desarrollo del producto, es muy probable que sus intereses no se tenga en cuenta, generando consecuentemente un sesgo preexistente. De acuerdo con la discusión de las secciones anteriores, los artistas que no son superestrellas deberían considerarse partes interesadas para cualquier proyecto relacionado con sistemas recomendadores musicales. Lamentablemente, en muchos casos pareciera no existir un interés real por comprender sus necesidades. En consecuencia, se ven afectados negativamente por el efecto de popularidad que promueven los actuales modelos de negocio de la industria musical.

En vista de lo anterior, las siguientes recomendaciones son importantes para el tratamiento de los sesgos preexistentes:

- Identificar a todas las partes implicadas en el caso de negocio y evaluar sus intereses.
- Analizar cómo afectan a cada una de las partes interesadas los requerimientos y limitaciones definidos por el modelo de negocio.
- Tener en cuenta las necesidades de todas las partes interesadas en el proceso de desarrollo del producto.

- Mantener una estrecha comunicación con todas las partes interesadas que permita tomar decisiones enfocadas a mejorar un trato justo para todos y todas.

En cuanto a los los sesgos técnicos, hay dos formas de analizar el asunto. Por un lado, está el caso en que el sesgo técnico es inevitable porque es consecuencia de un sesgo preexistente. En este caso, el personal técnico del proyecto sigue órdenes y aplica las reglas de negocio definidas por las partes interesadas. Por otro lado, está el caso en que las condiciones de los conjuntos de datos, las estrategias de recomendación y sus algoritmos presentan debilidades técnicas, como podría ser un *dataset* desbalanceado, o un algoritmo basado únicamente en popularidad (típicamente variables contadoras). Considerando el caso en donde se presentan debilidades técnicas, las siguientes recomendaciones son importantes para el manejo del sesgo técnico:

- Comprender en detalle el modelo de negocio y los datos implicados, así como el punto de vista de cada parte interesada.
- Aplicar un proceso de evaluación riguroso para identificar las estrategias de recomendación más adecuadas teniendo en cuenta las necesidades específicas del modelo de negocio.
- Evitar la estrategia de recomendación basada en la popularidad, especialmente si el caso de negocio incluye artistas que no son superestrellas.
- Tener precaución con la información extraída de las redes sociales para aplicar una estrategia de recomendación basada en el filtrado colaborativo. Hay que evaluar la calidad de esta información y la forma en que se integrará al sistema recomendador.
- Diseñar métricas, o seleccionar algunas disponibles, para evaluar la calidad de los datos según las definiciones del modelo de negocio, y las estrategias de recomendación seleccionadas para implementar en el sistema recomendador.
- Diseñar descriptores musicales más cercanos a la realidad indicada por los artistas, sugiriendo la importancia de marcar las diferentes partes de la estructura de la canción, e implementando estrategias de recomendación para cada una de estas partes.
- Considerar la estrategia de filtrado basada en emociones. Para lograr mejores resultados, el reconocimiento de las emociones en música debe hacerse a lo largo del tiempo, teniendo en cuenta que la acción de escuchar música es una experiencia emocional dinámica dada por la estructura de la canción [34].
- Se recomienda ser cuidadoso con los conjuntos de datos desbalanceados, especialmente cuando están implicados en estrategias basadas en el enfoque de *machine learning*, ya que el proceso de aprendizaje favorece a la clase mayoritaria y, en consecuencia, las predicciones mostrarán un gran sesgo por clases. Existen diferentes formas de tratar los datos desbalanceados. Por una parte, es posible implementar clasificadores binarios (uno por clase). Por otra parte, se puede pensar en implementar estrategias de balanceo [83].

- Analizar cómo tratar la información subjetiva proporcionada por la percepción humana a través de los procesos de etiquetado, especialmente en el caso del etiquetado emocional, lo cual impacta directamente en cualquier proceso futuro de análisis de datos sobre conjuntos de datos musicales [129].

- Considerar siempre que la música es un arte y debe tratarse como tal. Aunque la música puede estudiarse como una señal digital, hay algunas características de alto nivel, como las emociones y los conceptos musicales, que deben comprenderse en profundidad e incluirse en cualquier proceso de diseño de estrategias de recomendación.

Para terminar, en cuanto a los sesgos emergentes, es importante resaltar que éstos sólo pueden detectarse en un contexto real de uso, por lo que un proceso de seguimiento continuo será la clave para gestionarlos. La aplicación de estrategias de retroalimentación podría ser una buena forma de descubrir nuevas percepciones de trato injusto desde el punto de vista de cada parte interesada identificada en el pasado, así como de nuevas partes interesadas que pudieran aparecer en un futuro cercano. Esta informaaión de retroalimentación serviría para proponer nuevos cambios en el sistema recomendador, con el fin de mitigar los sesgos emergentes actuales.

3.9 Conclusiones

A continuación se presentan las conclusiones generadas para cada uno de los apartados anteriores.

En cuanto a las librerías de extracción de características: La recuperación de las características de la música desde archivos digitales, comprende dar solución al problema de la extracción de propiedades de sonido a través de analizadores de contenido, que usualmente son implementados a través de librerías. El éxito de estos procesos de extracción depende en gran parte de la efectividad de las técnicas internas y el análisis de las diversas combinaciones de características de bajo nivel, para la reconstrucción de características de más alto nivel. Las características de alto nivel generalmente son propuestas como modelos de clasificación. Entre los modelos más utilizados para clasificar la música se suele encontrar la clasificación por emociones. Dicha clasificación es compleja, y en muchos de los casos es validada por la evaluación de expertos, quienes, mediante anotación o etiquetado, asocian emociones con una canción determinada. Para los modelos automáticos de clasificación emocional muchas de las librerías tienen sus propias propuestas. Sin embargo, dichos modelos son evaluados constantemente y sometidos a procesos de mejora continua.

La clasificación emocional de la música puede ser utilizada, como una base para los sistemas recomendadores. En donde, dependiendo de la clasificación de la música, dichos sistemas puedan sugerir a los oyentes piezas musicales de acuerdo a una emoción que se quiera transmitir; y con ello facilitar el consumo de la música desde plataformas digitales. El estudio comparativo desarrollado entre las librerías *Spotify API, jAudio* y *AcousticBrainz*, es un primer paso para analizar los parámetros musicales que se relacionan con las emociones. Aunque muchas de estas librerías de MER actualmente

ofrecen información interesante, los sistemas MER todavía no están lo suficientemente desarrollados para ofrecer una solución universal, y que además genere un aporte altamente confiable a los sistemas recomendadores, lo que motiva la continuidad de la investigación en esta línea.

Respecto a las bases de datos musicales: La revisión se realizó teniendo en cuenta el campo de los sistemas de recomendación musical, encontrando limitaciones muy importantes, como lo son: la falta de canciones originales de artistas noveles, la falta de etiquetado completo de la estructura musical, la falta de una licencia que permita el acceso a los archivos de sonido, y la falta de etiquetado emocional para cada parte de la estructura de una canción por parte del artista y los oyentes. Todas estas limitaciones, muy relacionadas también con sesgos preexistentes, muestran la importancia de diseñar y compartir nuevas bases de datos (*datasets*) que le permitan a la comunidad científica seguir avanzando en el reconocimiento de emociones de la música, y el diseño de mejores sistemas recomendadores musicales.

Con relación a MediaEval: MediaEval es quizás uno de los *datasets* más completos en cuanto a etiquetado emocional, por lo que en principio se considera relevante para esta Tesis. Sin embargo, también es importante mencionar sus limitaciones:

- Se evidencia una alta variabilidad de las anotaciones emocionales entre diferentes anotadores, lo que genera una gran preocupación sobre las condiciones en que se definió el proceso de etiquetado y las razones por las cuales esta situación se presenta.
- Existe un evidente desequilibrio en la distribución de los datos de las canciones entre los cuadrantes, lo que debe considerarse en cualquier experimentación con este *dataset*.
- Al igual que los demás *datasets* revisadas en el capítulo 3.2, MediaEval no involucra el artista de ninguna manera en procesos de etiquetado, y los fragmentos de sonido no cuentan con una estructura musical definida.

Con relación a los sistemas de predicción: Se revisaron algunos trabajos que implementan modelos predictivos para el reconocimiento de emociones en la música. En principio, se identifica que las diferentes configuraciones permitidas por las técnicas utilizadas, tienen un impacto relevante sobre el desempeño, así como también los *datasets* musicales utilizados. Se resaltan los siguientes aspectos:

- La comparación de trabajos en el campo de MER es realmente complicado, considerando que los experimentos son ejecutados con diferentes configuraciones, diferentes algoritmos, y diferentes *datasets*.
- En los trabajos revisados no se realiza un análisis de *features* (conocido como un análisis de sensibilidad), que permita comprender cuales son las *features* más relevantes, y si algunas de ellas podrían ser descartadas. Tampoco se evidencia con tanta frecuencia la reducción de componentes a través de un análisis de componentes principales (*Principal Component Analysis* - PCA).

- En los trabajos revisados el análisis se concentra en *features* de bajo nivel de la música, lo que representa el sonido, quedando pendiente un análisis y/o discusión más al nivel de la música, esto con el objetivo de considerar la participación de compositores y plantear una discusión en una terminología que sea cercana a su disciplina (la música como arte).

Con relación a los sistemas de clasificación: Se revisaron algunos trabajos que implementan modelos de clasificación para el reconocimiento de emociones en la música. Adicionalmente, nuevamente se referenciaron algunos trabajos de predicción, para resaltar que algunos de ellos son utilizados como punto de partida para el diseño e implementación de modelos de clasificación. Dentro de la revisión de los trabajos también se ha analizado las características de los *datasets*, las técnicas utilizadas en los modelos, y las métricas consideradas para evaluar el desempeño de estos mismos. Un hallazgo importante es que en muchos casos existen problemas con *datasets* musicales que se encuentran desbalanceados, lo que tiene un impacto negativo en la capacidad de generalización de la mayoría de los algoritmos de *machine learning*; a partir de este hallazgo se considera que es importante definir una fase de preprocesamiento de datos, que como mínimo implemente estrategias de balanceo de datos y estratificación, por lo que dicha fase se considera más adelante en los aportes de esta Tesis.

También se resalta la complejidad para comparar el desempeño entre los diferentes trabajos debido a sus diferencias metodológicas en la representación de los datos, procesos de etiquetado, la selección de las características, y los modelos emocionales, que conduce a diferentes métricas de evaluación, lo que hace realmente difícil comparar la precisión de los algoritmos aplicados [130]. Además, los métodos y experimentos propuestos en los diferentes trabajos son muy difíciles de replicar o comparar, ya que la mayoría de ellos utilizan conjuntos de datos privados o diferentes conjuntos de datos públicos, con un número insuficiente de canciones y diferentes características de bajo y alto nivel [108].

Con relación a los sistemas difusos (*fuzzy*): En los procesos de clasificación, la lógica difusa podría generar mayores tasas de acierto en comparación con las obtenidas mediante algoritmos de *machine learning*, debido a la especificación detallada de sus reglas por parte de un experto, aunque esto implica un enorme esfuerzo que quizás no resulta muy práctico para la dinámica que siguen los sistemas recomendadores musicales, en su interés por aprender del comportamiento de los usuarios para ofrecer una experiencia lo más personalizada posible. En definitiva, se resalta especialmente el proceso de etiquetado, en el que la teoría de conjuntos difusos se muestra muy potente para tratar con la subjetividad de dominios específicos, como es el caso de las emociones, y de algunas características musicales.

Con relación a los sistemas recomendadores musicales: Se realizó una revisión del estado del arte enfocada en la comprensión de las diferentes estrategias de recomendación utilizadas por los sistemas recomendadores musicales. Se resalta la importancia que tienen las estrategias de filtrado basado en contenido y emociones en los trabajos más recientes. También se identifica un interés especial en generar procesos de recomendación personalizados para mejorar la experiencia del usuario.

Con relación a los sesgos: Se identificaron y analizaron los sesgos más típicos en

los MRS. Para ello, se realizó una revisión bibliográfica, teniendo en cuenta diferentes enfoques como el análisis de las estrategias de recomendación, los estudios de sesgos y los *datasets* musicales. El análisis permitió identificar las estrategias de recomendación más recientes utilizadas en MRS, el tipo de sesgos (preexistentes, técnicos, emergentes) que se presentan en cada estrategia de recomendación, así como los sesgos más comunes en los conjuntos de datos musicales. El análisis también reveló importantes hallazgos que ayudan a entender cómo y por qué están presentes los sesgos en MRS, como por ejemplo, que el filtrado colaborativo y el basado en la popularidad son algunas de las estrategias más afectadas por los sesgos preexistentes, y los sesgos técnicos están más relacionados con las condiciones de los datos y los algoritmos. También se incluyeron algunas recomendaciones para el tratamiento de sesgos.

Considerando las diferentes limitaciones identificadas a partir de la revisión del estado del arte, esta Tesis presenta sus aportes en los siguientes capítulos, generando contribuciones a través del diseño de un prototipo de sistema de predicción (Apartado 4.1), un prototipo de sistema de clasificación no determinístico (Apartado 4.2), un prototipo de sistema de clasificación determinístico (Apartado 4.3), el diseño de un nuevo *dataset* musical (Capítulo 5), y un prototipo de sistema recomendador (Capítulo 6) que busca mitigar sesgos preexistentes y técnicos.

Capítulo 4

Implementación de sistemas para el reconocimiento de emociones en la música

En este capítulo se presenta el diseño y el desarrollo de tres prototipos para el reconocimiento de emociones en la música, el primero desde un enfoque predictivo (Apartado 4.1), el segundo desde un enfoque de clasificación no determinístico (Apartado 4.2), y el tercero desde un enfoque de clasificación determinístico (Apartado 4.3). Para cada uno de los prototipos se incluye una descripción del proceso de su diseño, como también una sección dedicada a la validación de los resultados obtenidos por los diferentes experimentos. De manera general, en el diagrama de bloques de la Figura 4.1 se muestran las estrategias de solución utilizadas para el diseño y desarrollo de estos prototipos. Para finalizar, se presentan las conclusiones generales del capítulo en relación a los tres prototipos diseñados (Apartado 4.4). Estos prototipos y el análisis de su comportamiento que se presenta aquí, se constituyen en los aportes centrales de esta Tesis, y serán en gran medida, la base para el diseño de la estrategia de recomendación implementada en el prototipo del sistema de recomendación musical basado en emociones que se describe en el capítulo 6.

4.1 Sistema de predicción de emociones

En este apartado se desarrolla un sistema de predicción de emociones para la música [131]. El objetivo principal de este sistema consiste en determinar el valor de *valence* y *arousal* para una canción a partir de sus características de sonido de bajo nivel. Una vez obtenido ese valor, la canción correspondiente es ubicada a través de una coordenada *(X,Y)* en un plano bidimensional con el nivel de *valence* en el eje X y el de *arousal* en el eje Y. Posteriormente, se evalúa la tasa de éxito de la ubicación del valor predicho en el plano V/A, y con ello se analizan posibles acciones de mejora. Inicialmente se describe el diseño del sistema (Sección 4.1.1), luego se presentan los experimentos con la respectiva discusión de resultados (Sección 4.1.2). Y finalmente, se presentan las consideraciones más relevantes de este sistema (Sección 4.1.3).

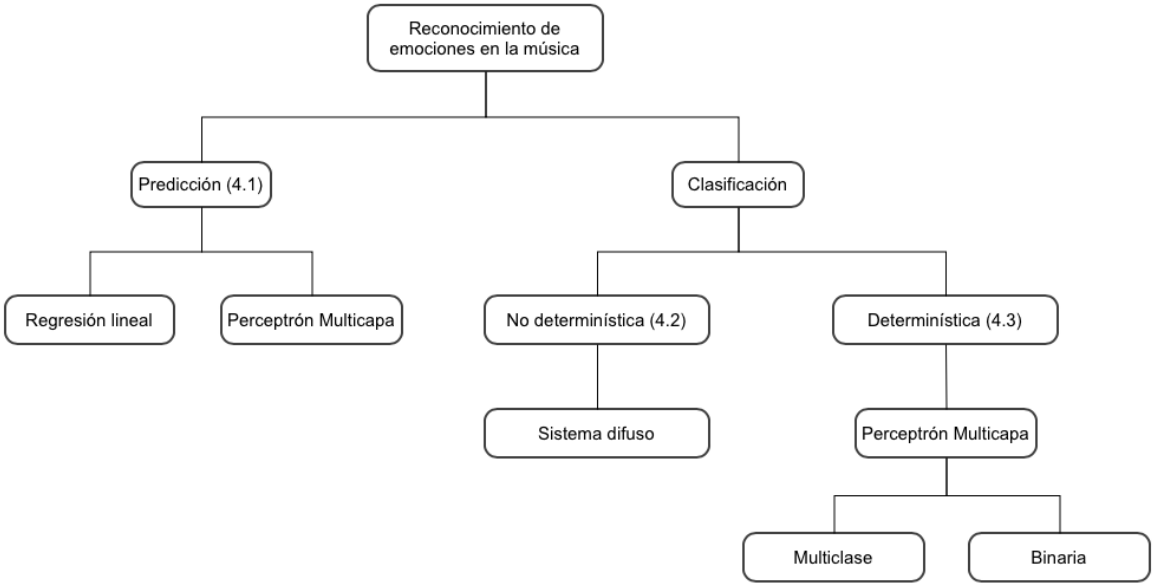


Figura 4.1: Estrategias de solución utilizadas para el diseño y desarrollo de los prototipos MER.

4.1.1 Diseño del sistema

Para el diseño del sistema se implementa, en primer lugar, un modelo de regresión lineal con múltiples características de bajo nivel para probar una técnica de predicción simple. Debido al gran número de características del *dataset* de MediaEval, se encuentra una alta complejidad en la construcción de la función hipótesis, trayendo como consecuencia un error cuadrático medio (RMSE) aproximado de 0.8 para los modelos de *valence* y *arousal*. En este caso, las predicciones son difíciles de realizar mediante una única función hipótesis debido al gran número de variables con relaciones no lineales. Generalmente, las redes neuronales consiguen mejores resultados en comparación con los modelos de regresión lineal, por lo que, para mejorar los resultados obtenidos, se diseña un nuevo modelo de predicción bajo el enfoque de perceptrón multicapa (MLP) y se implementa a través de la librería *tf.keras*[1] incluida en *TensorFlow*. La estructura de la red neuronal se presenta en la Figura 4.2. En esta Figura se pueden observar las neuronas de entrada, que representan las características de sonido de las canciones; el MLP requiere de 260 neuronas de entrada para las 260 características de sonido disponibles en el conjunto de datos de MediaEval (ver apartado 3.3). Es importante mencionar que la implementación del PCA (Análisis de componentes principales) podría reducir el número de características; en consecuencia, se necesitaría un número menor de neuronas de entrada. También se incluye una capa oculta, cuya arquitectura interna es un parámetro de diseño con el cual se puede experimentar para optimizar la capacidad de generalización de la red neuronal. Por último, es importante señalar que el MLP únicamente tiene una neurona de salida, por lo que son necesarias al menos

[1]https://www.tensorflow.org/api_docs/python/tf/keras

dos redes neuronales para implementar el sistema de predicción de emociones, una para *valence*, y otra *arousal*.

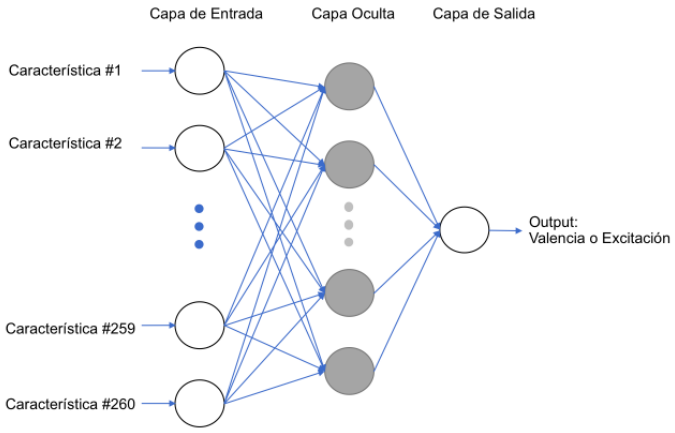


Figura 4.2: Perceptrón Multicapa

.

Las principales fases propuestas para el diseño del sistema de predicción se presentan en la Figura 4.3, y se explican a continuación.

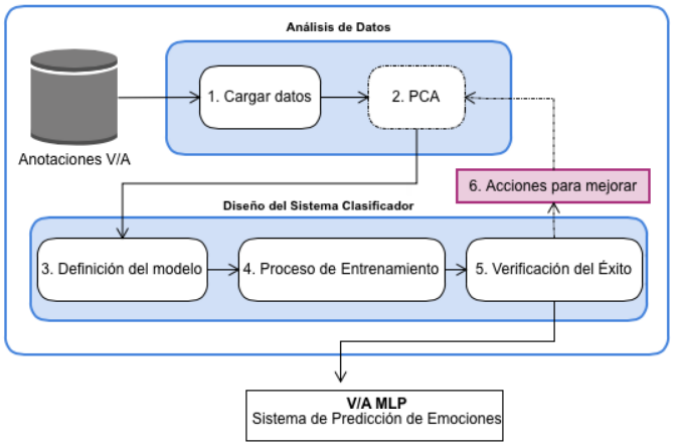


Figura 4.3: Proceso de predicción de las emociones - Fases principales. Elaboración propia.

Fase 1: Carga de datos

Los pasos para la carga de datos son los siguientes:

- Transformación de los valores de las anotaciones V/A disponibles en el *dataset* de MediaEval a una nueva escala con valores entre 1 y -1. Esta escala suele ser la más utilizada en la literatura de referencia para facilitar la comparación de resultados entre diferentes trabajos.

- Definición de la proporción de datos seleccionados para realizar el entrenamiento y las pruebas. Por recomendación el 80 % para entrenamiento y el 20 % para pruebas, teniendo en cuenta que estos valores son sugeridos por librerías como *TensorFlow* y *Scikit-learn*, y además, considerando el tamaño del *dataset* de MediaEval, el cual no es muy grande, y por tal razón no es conveniente definir una menor proporción de datos de entrenamiento debido a que afectaría su capacidad de generalización.
- Carga de los datos seleccionados por cuadrante.
- Normalización de las características de bajo nivel para facilitar la convergencia del proceso de entrenamiento.

Fase 2 (Opcional): PCA

El PCA se aplica para reducir la dimensión del conjunto de datos de MediaEval. En general, esta técnica mejora el rendimiento de los algoritmos de *machine learning* en lo que respecta al tiempo de convergencia y las tasas de éxito de las predicciones [132]. Esta fase es opcional y permite comparar las tasas de éxito con y sin PCA.

Fase 3: Definición del modelo

El proceso de definición del modelo implica las siguientes acciones:

- Definición de la arquitectura de capas.
- Configuración de la función de activación utilizada por las neuronas.
- Configuración de la tasa de aprendizaje (*learning rate (LR)*) para el proceso de entrenamiento, el cual controla la longitud de salto sobre la curva de error para luego generar la actualización de los pesos sobre las conexiones de las diferentes neuronas.
- Definición del modo de entrenamiento: entrenamiento convencional, validación cruzada o parada temprana (*early stop*).
- Configuración de las métricas para luego analizar las tasas de éxito. Las métricas incluidas en el monitoreo incluyen el RMSE y el error absoluto medio (MAE), que corresponden a las más utilizadas en sistemas de predicción de valores aproximados.

Fase 4: Proceso de entrenamiento

Es importante mencionar los siguientes pasos del proceso:

- Las épocas de entrenamiento pueden establecerse manualmente o en modo de parada anticipada. El modo de parada anticipada permite que el proceso de entrenamiento se detenga automáticamente cuando se alcanza el error absoluto medio (MAE) más bajo.

- Creación de un modelo para las predicciones de *arousal* y otro para las de *valence*; la independencia de cada modelo permite mejorar y especializar su capacidad de predicción.
- Para cada modelo hay que definir: el conjunto de datos de entrenamiento, el conjunto de datos de pruebas, las épocas y el parámetro de parada temprana (opcional).
- Para el proceso de entrenamiento con validación cruzada, se debe especificar el conjunto de datos de validación [133].

Fase 5: Verificación del éxito

En esta fase se calculan las métricas de rendimiento del sistema. Los resultados obtenidos de las métricas son útiles para establecer nuevos valores en los parámetros del sistema (Fase 3 y 4) con el fin de reducir el nivel de error en las predicciones. Las métricas MAE y RMSE utilizan una escala numérica entre 0 y 1, en donde los valores más cercanos a 1 indican un mayor nivel de error, mientras que los valores más cercanos a 0 indican un menor nivel de error.

Fase 6: Acciones para mejorar

Según los resultados obtenidos, se configura el modelo predictivo con nuevos ajustes y se repite el proceso de entrenamiento hasta conseguir una mejora [134].

4.1.2 Experimentos y resultados de los modelos

En esta sección se presenta un resumen de los resultados obtenidos en los diferentes escenarios considerados para la experimentación. La tasa de éxito se analizó mediante las métricas MAE y RMSE. Con respecto a la configuración de los parámetros del modelo, estos fueron probados con las siguientes tasas de aprendizaje (LR): 0.001, 0.010, 0.020, 0.030, 0.040, 0.050, 0.060, 0.070. Además, el modelo se probó con una y dos capas ocultas (HL) en experimentos independientes; del mismo modo, se establecieron 64 y 128 neuronas en las capas ocultas para evaluar las tasas de éxito.

Se diseñaron 3 experimentos en donde se ajustaron los parámetros relacionados con la modalidad de entrenamiento y la activación del PCA, estos experimentos incluyeron ambos modelos, *valence* y *arousal*. Posteriormente, los mejores escenarios de prueba, en consideración a MAE y RMSE más bajos, se presentan en la Tabla 4.1, en donde se muestra el mejor escenario para cada uno de los 3 experimentos. Para todos los casos, el conjunto de datos de entrenamiento y el conjunto de datos de prueba fueron los mismos con el objetivo de evaluar y comparar diferentes configuraciones del MLP. El Experimento 1 implementa la parada temprana, y por esa razón tiene un conjunto de datos adicional para la validación, que es un subconjunto de datos de entrenamiento. Los experimentos 1 y 2 utilizan las 260 características de bajo nivel para el proceso de entrenamiento. En el experimento 3, después de aplicar el PCA con el 95 % para la

retención de la varianza, se obtienen dos componentes, que se utilizan como características en el proceso de entrenamiento. Los mejores MAE/RMSE para el *valence* y el *arousal* se obtuvieron después de aplicar PCA; respectivamente 0.18/0.23 y 0.20/0.24.

Tabla 4.1: Mejores escenarios de pruebas para los experimentos

Configuración de experimentos			Valence		Arousal	
Exp #	Modo	PCA	MAE	RMSE	MAE	RMSE
1	Parada temprana	No	0.21	0.25	0.22	0.27
2	Convencional	No	0.20	0.25	0.23	0.28
3	Convencional	Sí	**0.18**	**0.23**	**0.20**	**0.24**

Como se mencionó en la sección 3.3, el conjunto de datos de MediaEval no se encuentra perfectamente balanceado por cuadrantes, lo que podría tener un impacto en la tasa de éxito de la predicción, ya que el MLP terminaría especializándose más en los cuadrantes con mayor cantidad de datos.

Teniendo en cuenta lo anterior, el conjunto de datos se dividió por cuadrantes en un nuevo experimento y cada uno de ellos se utilizó para entrenar un modelo completamente independiente. La Tabla 4.2 y la Tabla 4.3 muestran diferentes posibilidades de configuración de la red neuronal que permiten obtener el mejor MAE/RMSE para el proceso de predicción de *valence* y *arousal*, respectivamente. También se específica la tasa de aprendizaje (*learning rate - LR*), la cantidad de capas ocultas (*hidden layers*), y el número de neuronas por capa.

Tabla 4.2: Entrenamiento convencional con PCA para Valence

Cuadrante	LR	HL	Neuronas	MAE	RMSE
Q1	0.070	1	128	**0.18**	**0.14**
Q2	0.050	2	64	**0.09**	**0.11**
	0.060	2	64	0.10	0.11
	0.070	2	64	**0.09**	**0.11**
	0.030	1	64	**0.09**	**0.11**
Q3	0.070	1	64	0.12	0.14
	0.030	2	64	0.11	0.14
	0.070	2	64	**0.10**	**0.14**
Q4	0.010	1	128	**0.11**	**0.14**
	0.050	1	128	**0.11**	**0.14**

Las métricas de la tasa de éxito de cada cuadrante se analizaron para ambos modelos predictivos: *valence* y *arousal*. Se aplicó el PCA con el 95 % para la retención de la varianza y se ejecutó el proceso de entrenamiento sin parada anticipada. En el mejor de los casos, fue posible obtener un RMSE de 0.11 (Q2 para *valence* y Q4 para *arousal*), lo que representa una tasa de éxito considerablemente buena teniendo en cuenta los trabajos analizados en la Tabla 3.5 (Apartado 3.4), en donde se evidencia una RMSE mínima de 0.13.

Tabla 4.3: Entrenamiento convencional con PCA para Arousal

Quadrant	LR	HL	Neurons	MAE	RMSE
Q1	0.040	1	128	0.14	0.16
	0.060	1	128	**0.13**	**0.16**
	0.040	1	64	**0.13**	**0.16**
	0.070	1	64	**0.13**	**0.16**
Q2	0.040	2	64	**0.11**	**0.14**
	0.060	1	128	**0.11**	**0.14**
Q3	0.040	1	64	**0.12**	**0.14**
	0.060	1	64	**0.12**	**0.14**
	0.050	2	64	**0.12**	**0.14**
Q4	0.040	1	128	**0.09**	**0.11**
	0.070	2	64	**0.09**	**0.11**

Estos resultados numéricos parecen prometedores, por lo que se consideraron los valores V/A predichos para hacer una prueba de clasificación emocional por cuadrantes. Como extensión del sistema de predicción, se implementó un motor de reglas condicionales muy básico. De acuerdo con la coordenada obtenida por los mejores modelos presentados anteriormente en las Tablas 4.2 y 4.3, el sistema de reglas determinaría el cuadrante. Los resultados se muestran en la Figura 4.4, en donde los punto azules corresponden a la ubicación real, y los puntos de color naranja corresponden a los valores previstos. Es importante analizar que los valores predichos no están debidamente agrupados dentro de los cuadrantes que les corresponde. En particular, la tasa de precisión de la clasificación obtenida por cuadrante es: Cuadrante 1 (Q1) = 75 %, Cuadrante 2 (Q2)= 5 %, Cuadrante 3 (Q3) = 35 %, y Cuadrante 4 (Q4)= 21 %.

4.1.3 Consideraciones acerca del primer prototipo

Este apartado se ha centrado en el diseño de un sistema para la predicción de una coordenada V/A (modelo dimensional emocional) partiendo del entrenamiento de una red neuronal, la cual es entrenada a partir de 260 características de bajo nivel con valores V/A previamente etiquetados a través del proyecto MediaEval. Los resultados experimentales mostraron que la definición de un modelo de predicción para cada cuadrante y para cada dimensión del modelo emocional (*valence* y *arousal*) mejora las tasas de éxito en comparación con la definición de un solo modelo de predicción para todo el conjunto de datos. Sin embargo, aunque los valores de RMSE obtenidos por el sistema de predicción emocional parecen ser bastante buenos, no fue posible extender la funcionalidad de este sistema para lograr el éxito en las clasificaciones emocionales en cuatro o más cuadrantes mediante reglas lógicas. La extensión de un sistema de predicción a un sistema de clasificación obteniendo una predicción de los valores de *valence* y *arousal* no es suficiente para realizar una clasificación exitosa. Por ello, se encuentra necesario diseñar e implementar un nuevo sistema de clasificación emocional, lo que se realizará en los dos siguientes apartados, desde una perspectiva no determinística, y luego, determinística.

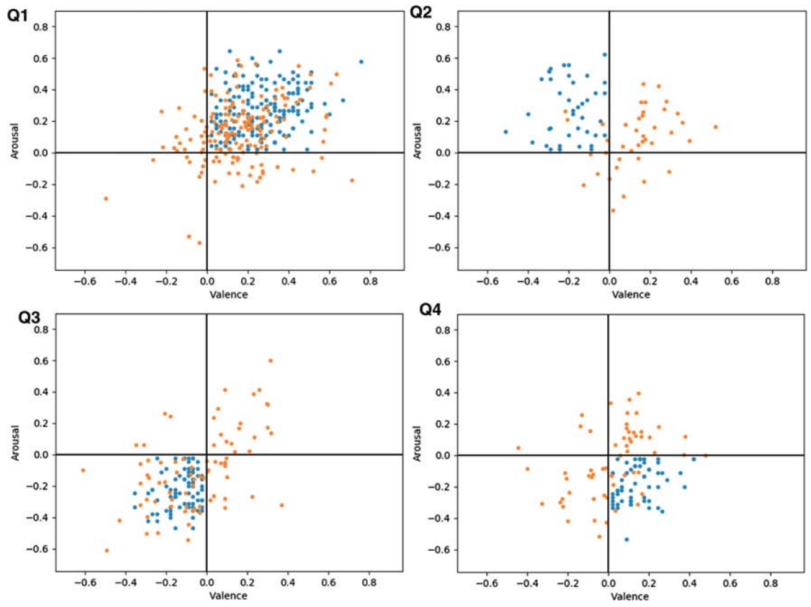


Figura 4.4: Clasificación basada en la predicción. Q1 arriba a la izquierda, Q2 arriba a la derecha, Q3 abajo a la izquierda y Q4 abajo a la derecha. La ubicación en cuadrantes de los valores predichos dista mucho de ser adecuada para un sistema de clasificación de emociones (especialmente Q2 y Q4). Valores reales en puntos azules, valores predichos en puntos naranjas. Elaboración propia.

4.2 Sistema de clasificación emocional no determinística

En este apartado se presenta el diseño de un sistema difuso, detallando el proceso completo de fusificación de la entrada, desfusificación de la salida, y el diseño de las reglas de inferencia, teniendo en cuenta la arquitectura definida anteriormente en la Figura 2.5. El sistema difuso fue diseñado para reconocer el nivel de *arousal* (alto, medio, bajo) a través de las pulsaciones por minuto de una canción. Este sistema se ha implementado utilizando la librería *skfuzzy.control*[2] que se ejecuta sobre lenguaje *Python*[3]; este subpaquete proporciona un *API* de alto nivel para el diseño de sistemas *fuzzy*.

La estrategia de etiquetado de un sistema MER bajo un enfoque de lógica difusa se presenta en los procesos de fusificación de los antecedentes (Sección 4.2.1) y desfusificación de los consecuentes (Sección 4.2.2). En cuanto a la estrategia de clasificación MER, ésta se implementa a través de reglas basadas en un sistema de inferencia (Sección 4.2.3). En la Sección 4.2.4 se presentan diferentes escenarios de prueba para estudiar y discutir los hallazgos encontrados en los valores *crisp* y *fuzzy* tanto en la entrada como en la salida del sistema. Finalmente, se presentan las consideraciones más importantes sobre este sistema (Sección 4.2.5).

[2]https://pythonhosted.org/scikit-fuzzy/api/skfuzzy.control.html
[3]https://www.python.org

4.2.1 Fusificación de los antecedentes (Entradas)

Inicialmente, hay que definir las variables del sistema. Para este experimento en particular, la única variable de entrada es la velocidad percibida de la canción (*beats per minute*), que se explica a continuación desde el punto de vista de los valores *crisp* y *fuzzy*. El sistema se podría escalar fácilmente para considerar variables adicionales, aunque la base de conocimiento relacionada con estas variables es indispensable para formular nuevas reglas e integrarlas en el motor de inferencia (se explica en la sección 4.2.3)

- Universo (rango del valor *crisp*): ¿Qué tan rápida es la canción en términos de bpm en una escala de 0 a 200?
- Conjunto difuso (rango del valor *fuzzy*): lento (*slow*), moderado (*moderate*), rápido (*fast*).

Para cada uno de los 3 conjuntos difusos, se implementa una función de pertenencia trapezoidal. Esta función se eligió porque se ajusta a la descripción de la percepción del oyente [9][8], manteniendo unos intervalos de valores constantes con grado de pertenencia 1 a los respectivos conjuntos difusos. Las funciones se muestran juntas en la Figura 4.5, donde es posible analizar sus puntos de intersección y las áreas de solapamiento.

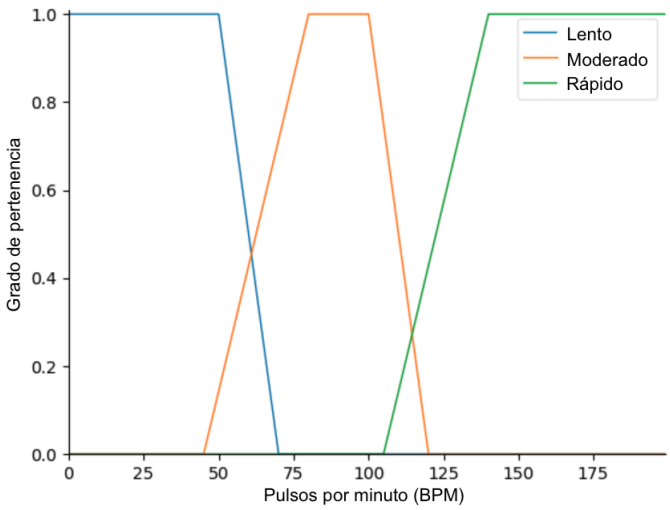


Figura 4.5: Funciones de pertenencia para las canciones con respecto a las categorías de velocidad. Elaboración propia.

4.2.2 Desfusificación de los consecuentes (Salidas)

En el consecuente, se determina la salida del sistema, en este caso el nivel de *arousal*. A continuación se ofrece la descripción *crisp* y *fuzzy*.

- Universo (rango del valor *crisp*): ¿Cuál es el nivel de excitación que percibo a través de la canción que escucho en una escala de 0 a 10?

- Conjunto difuso (rango del valor *fuzzy*): bajo (*low*), medio (*medium*), alto (*high*)

Las funciones de pertenencia se muestran en la Figura 4.6.

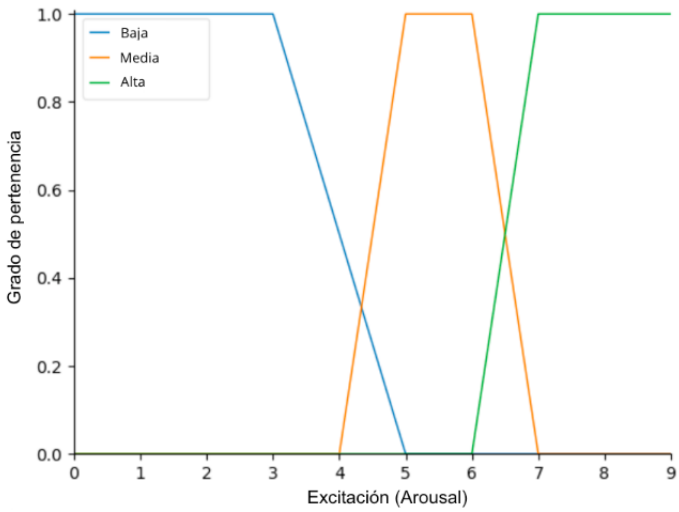


Figura 4.6: Funciones de pertenencia para las canciones con respecto a las categorías de *arousal*. Elaboración propia.

4.2.3 Reglas *fuzzy* (Inferencia)

Algunos estudios apoyan la relación entre el tempo de la canción y la percepción emocional [9] [8]. Normalmente, se formula una relación directa, en la que un nivel de tempo alto genera un nivel de percepción de excitación (*arousal*) alto, y un nivel de tempo bajo genera un nivel de excitación bajo. En este sentido, se proponen las siguientes reglas para el sistema de inferencia.

- SI el tempo es lento (*slow*), LUEGO la excitación será baja (*low*)
- SI el tempo es moderado (*moderate*), LUEGO la excitación será media (*medium*).
- SI el tempo es rápido (*fast*), LUEGO la excitación será alta (*high*).

Estas reglas se definen a través del módulo *skfuzzy.control.Rule*[4].

4.2.4 Experimentos y resultados del sistema

A través del módulo *ControlSystemSimulation*, el sistema de inferencia con el enfoque de *Mamdani* se coloca en funcionamiento, este modelo define 4 pasos: fusificación, evaluación de reglas, agregación de las salidas de las reglas, y desfusificación [135]. Inicialmente, se suministra un valor de entrada (antecedente), luego se ejecuta el sistema de control y, finalmente, se muestra el valor de salida (consecuente).

[4]https://pythonhosted.org/scikit-fuzzy/api/skfuzzy.control.html

En la Tabla 4.4, se presenta un conjunto de 8 casos de prueba. Para cada caso, se muestran los valores *crisp* para los valores de entrada y salida, así como los diferentes niveles de pertenencia para cada conjunto difuso (clase) dependiendo del dominio analizado. Para el caso del tempo, en el vector de izquierda a derecha se muestran los niveles de pertenencia de las clases: lento, moderado y rápido. Para el caso del *arousal*, en el vector de izquierda a derecha se muestran los niveles de pertenencia de las clases: bajo, medio y alto.

Tabla 4.4: Casos de prueba con valores *crisp* y vectores de valores de pertenencia para la entrada y la salida del prototipo de sistema difuso.

	Entrada (Bpm)		Salida (*Arousal*)	
Caso #	Valores *crisp*	Valores de pertenencia	Valores *crisp*	Valores de pertenencia
1	25	[1.0, 0.0, 0.0]	2.0	[1.0, 0.0, 0.0]
2	55	[0.75, 0.28, 0.0]	2.74	[0.74, 0.28, 0.0]
3	75	[0.0, 0.86, 0.0]	5.5	[0.0, 0.85, 0.0]
4	100	[0.0, 1.0, 0.0]	5.5	[0.0, 1.0, 0.0]
5	125	[0.0, 0.0, 0.57]	7.6	[0.0, 0.0, 0.57]
6	150	[0.0, 0.0, 1.0]	7.7	[0.0, 0.0, 1.0]
7	175	[0.0, 0.0, 1.0]	7.7	[0.0, 0.0, 1.0]
8	199	[0.0, 0.0, 1.0]	7.7	[0.0, 0.0, 1.0]

Es muy interesante observar que a medida que aumenta el valor *crisp* de la entrada, los valores de pertenencia en los vectores de pertenencia para las categorías de entrada y salida muestran un cambio, variando los valores de forma incremental de izquierda a derecha. Este comportamiento en los valores de pertenencia se debe a las funciones de pertenencia que se están utilizando, en las que la función de la clase central aumenta inicialmente, luego alcanza una meseta y finalmente disminuye. Por otro lado, también es importante destacar que los valores *crisp* de salida tienden a tomar valores centrales respecto al rango en el que se ha definido la función de pertenencia correspondiente; esto se debe al método del centroide implementado en la técnica de desfusificación.

En los sistemas difusos es importante entender los valores de salida, teniendo en cuenta que pueden ser expresados como valores *crisp*, o como valores de pertenencia a las diferentes clases existentes en la función de pertenencia de salida. Para entender mejor esta parte, se analiza a continuación el caso de prueba # 2, en el que el valor *crisp* de entrada es 55. Como se muestra en la Figura 4.7, las reglas de inferencia sólo activan dos zonas, que corresponden al *arousal* baja (*low*) y medio (*medium*); esto se conoce como agregación de salida. En la zona de activación de bajo *arousal*, hay un grado de pertenencia de 0.74, que es mucho mayor que el grado de pertenencia de 0.28 en la zona de *arousal* medio. Este análisis inicial revela que esta canción es predominantemente de un nivel bajo de *arousal*. El resultado final debe expresarse mediante un valor *crisp*, y para ello se aplica una técnica de desfusificación con el método del centroide[5], que calcula el punto en el que una línea vertical divide el conjunto en dos áreas con igual superficie. El valor *crisp* final de la salida es 2.74.

[5]https://pythonhosted.org/scikit-fuzzy/api/skfuzzy.defuzzify.html

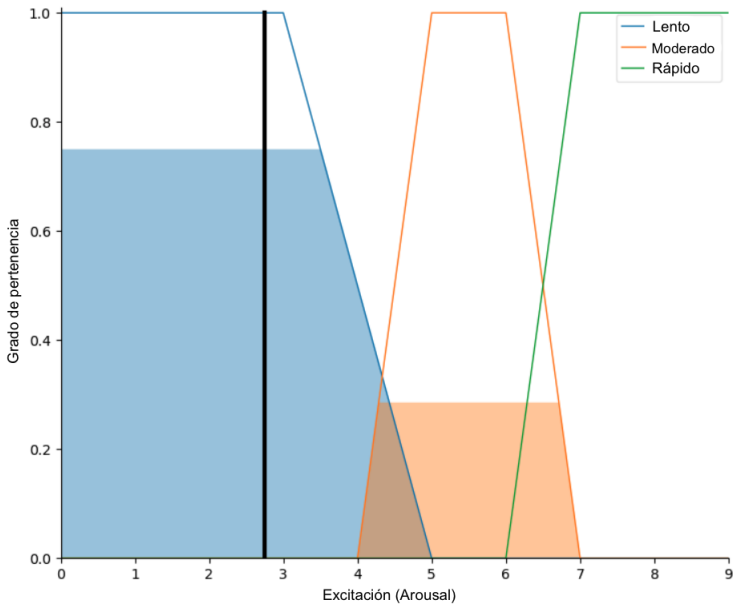


Figura 4.7: Salida del *Arousal* para el caso de prueba 2. Elaboración propia.

4.2.5 Consideraciones acerca del segundo prototipo

En este apartado se ha presentado el diseño un sistema para la clasificación emocional de la música a través de un enfoque no determinístico. Este sistema es una aproximación que permite comprender el nivel de esfuerzo para diseñar un sistema difuso con una sola variable de entrada y un motor de inferencia que se encarga del proceso de clasificación. Como se ha podido comprobar, todo el sistema de clasificación implica la elaboración de reglas de inferencia que deben ser validadas directamente con un experto. A medida que el sistema de clasificación se vuelve más sofisticado, el esfuerzo de diseño y la incorporación de nuevas reglas exige un trabajo altamente operativo, lo que no es muy conveniente en el ámbito de MER debido al dinamismo con el que se identifican cada vez más usuarios, factores de contexto, y también características musicales; así como el interés por la personalización en las estrategias de clasificación de los sistemas de recomendación [112]. Por estas razones, se decide no profundizar en la linea de sistemas difusos aplicados a MER, y se concentra el esfuerzo en el siguiente apartado 4.3 en soluciones basadas sobre *machine learning*. Sin embargo, con respecto a los resultados obtenidos por la estrategia de etiquetado MER bajo el enfoque difuso, hay que destacar cosas que son realmente muy positivas y que podrían considerarse en trabajos futuros:

- Las funciones de pertenencia de cada clase son un instrumento muy importante en el campo de las emociones, considerando que facilitan el proceso de asociar un valor dentro de una escala numérica, con una categoría; lo cual es muy útil para los sistemas MER que inicialmente aplican un modelo de predicción con un enfoque dimensional, y posteriormente requieren la adopción de un modelo de clasificación con un enfoque categórico [83].

- En cuanto a las características musicales, algunas de ellas también suelen des-

cribirse mediante cuantificadores sobre adjetivos, como la velocidad (lento, moderado, rápido, muy rápido), la posibilidad de bailar, entre otros casos. Así que la descripción de muchas de estas características también puede ser tratada con estrategias de fusificación y desfusificación.

4.3 Sistema de clasificación emocional determinística

En este apartado se presenta el diseño de un sistema de clasificación emocional para la música, que en lugar de predecir valores, predice directamente clases, particularmente, cuadrantes. Las principales tareas metodológicas del sistema de clasificación se muestran en la Figura 4.8. El sistema considera dos bloques de tareas: el pre-procesado de datos y la clasificación. En el pre-procesado (sección 4.3.1) se tienen las fases de etiquetado, selección de datos, y balanceo de datos. En el proceso de clasificación (sección 4.3.2) se detallan los tipos de clasificadores implementados, posteriormente, se explican las principales diferencias entre las estrategias de clasificación basadas en multiclase y en clasificación binaria (sección 4.3.3). Adicionalmente, se explica el enfoque de clasificación utilizando ventanas de tiempo (sección 4.3.4) y se analizan los resultados obtenidos (sección 4.3.5). Para finalizar, se presentan las principales consideraciones de este sistema (sección 4.3.6).

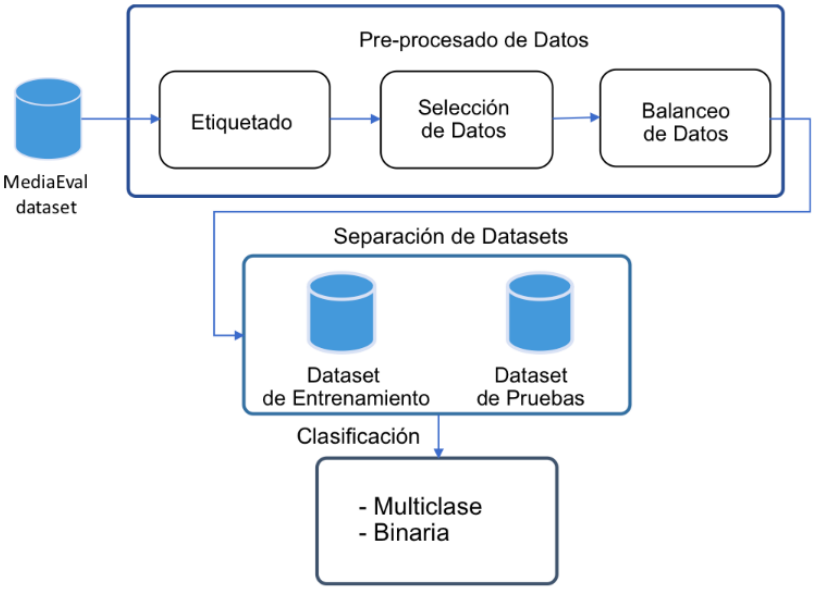


Figura 4.8: Diagrama de bloques para llevar a cabo un proceso de clasificación. Elaboración propia.

Este sistema ha sido desarrollado en lenguaje *Python*, y las principales librerías utilizadas se detallan en la Tabla 4.5.

Tabla 4.5: Librerías implementadas

Fase	Librerías
Pre-procesado de datos	model_selection.StratifiedShuffleSplit imblearn.over_sampling imblearn.under_sampling imblearn.combine
Clasificación	klearn.neural_network sklearn.ensemble.RandomForestClassifier sklearn.svm

4.3.1 Pre-procesamiento de datos

En esta fase inicial se realizan algunas transformaciones sobre el conjunto de datos original de MediaEval para obtener un conjunto de datos de entrenamiento y un conjunto de datos de prueba debidamente separados.

- **Etiquetado.** Cada canción se etiqueta con el cuadrante o clase binaria a la que pertenece.
- **Selección de datos.** Se define la proporción de datos seleccionados para el entrenamiento, la validación y la prueba. Por defecto, el 80 % para el entrenamiento y la validación, y el 20 % para las pruebas. El conjunto de datos de validación se establece mediante el parámetro de parada temprana, y suele ser el 20 % del conjunto de datos de entrenamiento. La división de los datos entre el entrenamiento y las pruebas se realiza con un enfoque estratificado, lo que permite separar los conjuntos de datos manteniendo las proporciones de las clases [136].
- **Balanceo de datos.** Se aplican diferentes estrategias para evaluar la distribución de las clases en el conjunto de datos de entrenamiento. Es importante destacar que los conjuntos de datos desbalanceados tienen un impacto negativo en las métricas de clasificación de las clases minoritarias [84] [137]. Para balancear los datos, se prueban y evalúan tres estrategias: *over-sampling*, *under-sampling* y una combinación de *over-sampling* con *under-sampling* [138]. Cada una de las estrategias anteriores tiene diferentes algoritmos que también se evalúan según los resultados obtenidos con el sistema de clasificación en cada uno de sus experimentos, como se presenta en la sección 4.3.5.

4.3.2 Clasificación con *Linear SVM*, *RandomForest* y *MLP*

Para realizar el proceso de clasificación se implementan y evalúan tres clasificadores: *Linear SVM* [139], *RandomForest* [140] y un *Multilayer Perceptron Classifier* (MLP) [141]. Es importante señalar que para todos los experimentos el conjunto de datos de entrenamiento se encuentra completamente separado del conjunto de datos de prueba de forma estratificada y aleatoria. Este punto es muy importante para garantizar la eficacia generalizadora del clasificador.

En un primer paso, los tres clasificadores propuestos se evalúan con el conjunto de datos anotado en 4 cuadrantes. Luego, se considera el clasificador MLP con una estrategia de clasificación binaria. Por último, se analiza la influencia del promedio temporal con diferentes ventanas de tiempo en la clasificación emocional dinámica. Los valores de los principales ajustes para cada uno de los clasificadores implementados se presentan en la Tabla 4.6.

Tabla 4.6: Valores de configuración para cada clasificador.

Técnica de clasificación	Parámetro	Valor
SVM	class_weight	balanced
RandomForest	n_estimators	100
	max_depth	2
	random_state	0
	class_weight	balanced
MLP	hidden_layer_sizes	(160)
	max_iter	500
	verbose	True
	activation	relu
	early_stopping	True
	validation_fraction	0.2
	tolerance	0.0001
	n_iter_no_change	20

Para el caso de *Linear SVM* y *RandomForest* no hay muchas posibilidades de parametrización en comparación con MLP. Sin embargo, uno de los parámetros más relevantes para el objetivo de este apartado, que también está disponible en ambos clasificadores, es la posibilidad de asignar pesos a las clases. Como se ha mencionado anteriormente, el conjunto de datos de MediaEval se encuentra desbalanceado, por lo que las técnicas de balanceo disponibles para cada clasificador son utilizadas y evaluadas, encontrando que no son estrategias suficientemente efectivas para mejorar el desempeño de la clasificación sobre las clases minoritarias (cuadrantes con menor cantidad de datos). Con respecto al MLP los parámetros más relevantes para evitar el sobreajuste y mejorar la capacidad de generalización son: *early_stopping*, *tolerance* y *n_iter_no_change*.

Algunas de las consideraciones presentadas en [142] fueron aplicadas para la definición de la arquitectura óptima de la red neuronal, en particular con respecto a la determinación del número de capas ocultas y su respectivo número de neuronas.

4.3.3 Clasificación considerando *One vs rest scheme*

Existen dos posibles enfoques en el diseño de sistemas de clasificación: clasificadores binarios y clasificadores multiclase. Aunque el uso de clasificadores binarios puede requerir más recursos computacionales (varios clasificadores se ejecutan en paralelo), en general permite mayores tasas de éxito en el proceso de clasificación [143] [144], debido a que cada clasificador se especializa en una clase en particular. Además, también se considera el uso de clasificadores binarios para reducir el impacto de la complejidad de la clasificación debido a los datos desbalanceados [145]. El enfoque *One vs rest scheme*

se incluye dentro de los experimentos basados en la clasificación emocional a lo largo del tiempo (presentado en el siguiente apartado), con el objetivo de analizar los resultados obtenidos con clasificadores binarios por cuadrante y clasificadores multiclase para *valence* y *arousal*, buscando identificar cual es el enfoque más adecuado para el problema que en particular se aborda en esta Tesis.

4.3.4 Clasificación emocional a lo largo del tiempo

En el apartado 3.3, se menciona que el *dataset* de MediaEval fue etiquetado dinámicamente. Este proceso se dio para cada canción, generando una anotación continua de 45 segundos. A partir de esta anotación, resulta importante analizar cual es la longitud de ventana adecuada para obtener una anotación emocional más precisa. Por esta razón, el sistema de clasificación se entrena con ventanas de tiempo que varian entre 0.5 y 10 segundos, con 0.5 segundos de incremento. Se genera un nuevo conjunto de datos para cada una de las ventanas de tiempo de diferente duración. Los nuevos datos se obtienen promediando en el tiempo los valores correspondientes tanto de las características del sonido como de las anotaciones emocionales. Finalmente, para cada caso, se describe el valor *F-measure* y se determina su relación con la longitud de la ventana. Para este proceso de clasificación dinámica es necesario personalizar el algoritmo de estratificación para garantizar una correcta separación entre los datos de entrenamiento y los de prueba. En primer lugar, se realiza una estratificación a partir de la base de datos promediada y, a continuación, se separan completamente todas las ventanas de tiempo de cada canción y se añaden al conjunto de datos correspondiente (de entrenamiento o de prueba).

4.3.5 Experimentos y resultados de los modelos

Esta sección presenta los resultados obtenidos para una serie de experimentos realizados de acuerdo a los enfoques explicados en la secciones anteriores (4.3.2, 4.3.3 y 4.3.4). Estos resultados son presentados desde dos perspectivas: sistema de clasificación con valores promediados por canción, y sistema de clasificación analizado por la longitud de la ventana de tiempo.

Sistema de clasificación con valores promediados por canción

A continuación se presentan los resultados de clasificación obtenidos al trabajar con el conjunto de datos MediaEval de 1802 canciones y 260 características de bajo nivel. Los valores de las características de anotación emocional y de sonido se promediaron a lo largo de la duración de la canción obteniendo un conjunto de valores únicos por canción.

Este trabajo ha considerado la implementación de tres clasificadores diferentes. Como se ha indicado en la sección 4.3.2, SVM y RandomForest tienen la posibilidad de ajustar las ponderaciones de las clases para tratar el problema del desbalanceo de clases.

En el caso de MLP, este parámetro no está disponible, por lo que la librería *imbalanced-learn*[6] se ha utilizado para tratar el desbalanceo de los datos. Esta librería cuenta con diferentes estrategias, que han sido evaluadas para identificar la mejor alternativa para el clasificador MLP.

Los resultados obtenidos se muestran en la Tabla 4.7. Estos resultados muestran que, sin aplicar el balanceo de datos, las clases menos representadas en el conjunto de datos no se clasifican correctamente. Aplicando alguna estrategia de balanceo de datos, el rendimiento de la clasificación mejora sustancialmente, aunque sigue siendo bajo.

Tabla 4.7: Clasificador multiclase con estrategias de equilibrio para el clasificador MLP.

Estrategia de balanceo.	Algoritmo	F-measure	AVG F
None	None	[0.74, 0.19, 0.60, 0.07]	0.40
OverSampler	None	[0.74, 0.19, 0.60, 0.07]	0.40
	resample	[0.73, 0.35, 0.62, 0.25]	0.49
	SMOTE	[0.73 ,0.34, 0.69, 0.21]	0.49
	ADASYN	[0.71, 0.33, 0.60, 0.30]	0.48
	BorderlineSMOTE	[0.73, 0.29, 0.59, 0.29]	0.47
UnderSampler	RamdomUnderSampler	[0.62, 0.25, 0.58, 0.32]	0.44
	ClusterCentroids	[0.51, 0.29, 0.53, 0.25]	0.39
Combinación	SMOTEEN	[0.64, 0.36, 0.57, 0.32]	0.47
	SMOTETomek	[0.72, 0.31, 0.62, 0.34]	**0.50**

Una vez identificada la mejor estrategia de equilibrio para el clasificador MLP, se han comparado los tres clasificadores propuestos. Se han calculado las medidas *F-measure* para cada cuadrante, y el rendimiento de clasificación de los clasificadores se resume en la Tabla 4.8. El experimento en el que se obtienen los mejores *F-measures* promediados para todas las clases es el que implementa MLP con SMOTETomek [146]. Sin embargo, se puede observar que, para las clases Q1 y Q3 otros clasificadores podrían clasificar con mejores resultados (Random Forest con la función class_weight), pero con las clases menos representadas en el conjunto de datos se obtienen valores de *F-measure* muy bajos.

Tabla 4.8: Comparación del clasificador multiclase implementando SVM, RandomForest y MLP.

Clasificador	Estrategia de balanceo	*F-measure*	AVG F
SVM	None	[0.75, 0.12, 0.61, 0.17]	0.41
	class_weight	[0.69, 0.30, 0.62, 0.24]	0.46
RamdonForest	None	[0.77, 0.04, 0.60, 0.21]	0.40
	class_weight	[0.78, 0.19, 0.64, 0.17]	0.44
MLP	None	[0.74, 0.19, 0.60, 0.07]	0.40
	SMOTETomek	[0.72, 0.31, 0.62, 0.34]	**0.50**

Para mejorar los resultados anteriores, se implementa un conjunto de clasificadores binarios. La Tabla 4.9 presenta cuatro clasificadores binarios aplicando MLP como algoritmo de clasificación, así como todas las estrategias de balanceo de datos. Estos resultados se obtienen mediante el enfoque de uno contra el resto (*one vs. rest*), en

[6]https://imbalanced-learn.readthedocs.io/en/stable/

el que se diseñan clasificadores independientes para identificar cada clase (cuadrante). Los mejores valores de *F-measure* para cada cuadrante se indican en negrita. Teniendo en cuenta que el cuadrante Q1 representa la clase mayoritaria, podría considerarse realizar un *Undersampling* para reducir su tamaño, sin embargo, esto generaría una pérdida de datos reales en los procesos de entrenamiento, por lo que finalmente se decide no aplicar ninguna estrategia de balanceo a este cuadrante. Con respecto los demás cuadrantes (Q2, Q3 y Q4), se aplican las diferentes estrategias y algoritmos disponibles para realizar balanceo, encontrando que la mejor estrategia de balanceo común para todos los cuadrantes es *Oversampling* con *BorderlineSMOTE*, en lugar de la estrategia de balanceo combinada SMOTETomek del clasificador anterior que incluía los cuatro cuadrantes (enfoque multiclase).

Tabla 4.9: Clasificadores binarios por cuadrante

Cuadrante	Estrategia de balanceo	Algoritmo	F-measure	AVG F
Q1	None	None	[0.72, 0.76]	**0.74**
Q2	None	None	[0.93, 0.04]	0.49
	OverSampler	resample	[0.91, 0.32]	0.62
		SMOTE	[0.91, 0.37]	0.64
		ADASYN	[0.89, 0.30]	0.60
		BorderlineSMOTE	[0.91, 0.39]	**0.65**
	UnderSampler	RamdomUnderSampler	[0.72, 0.24]	0.48
		ClusterCentroids	[0.54, 0.24]	0.39
	Combination	SMOTEEN	[0.83, 0.23]	0.53
		SMOTETomek	[0.89, 0.33]	0.61
Q3	None	None	[0.89, 0.60]	0.75
	OverSampler	resample	[0.88, 0.64]	**0.76**
		SMOTE	[0.87, 0.63]	0.75
		ADASYN	[0.86, 0.61]	0.74
		BorderlineSMOTE	[0.88, 0.64]	**0.76**
	UnderSampler	RamdomUnderSampler	[0.83, 0.63]	0.73
		ClusterCentroids	[0.75, 0.56]	0.66
	Combination	SMOTEEN	[0.83, 0.65]	0.74
		SMOTETomek	[0.86, 0.61]	0.74
Q4	None	None	[0.92, 0.04]	0.48
	OverSampler	resample	[0.89, 0.22]	0.56
		SMOTE	[0.89, 0.28]	**0.59**
		ADASYN	[0.89, 0.18]	0.54
		BorderlineSMOTE	[0.89, 0.28]	**0.59**
	UnderSampler	RamdomUnderSampler	[0.75, 0.30]	0.53
		ClusterCentroids	[0.48, 0.27]	0.38
	Combination	SMOTEEN	[0.81, 0.28]	0.55
		SMOTETomek	[0.88, 0.23]	0.56

Por último, la Tabla 4.10 presenta los clasificadores binarios para *valence* y *arousal* y sus respectivos *F-measures* obtenidos para el mismo conjunto de datos de prueba. Estos resultados son bastante buenos, debido a que los datos del *dataset* se distribuyen más uniformemente entre los medio-planos del plano V/A.

Tabla 4.10: Clasificadores binarios para *valence y arousal*

Clasificador	F-measure
[V-,V+]	[0.69, 0.77]
[A-,A+]	[0.66, 0.72]

En la Figura 4.9, a partir del mismo conjunto de datos de prueba de 301 canciones, se puede observar que 267 canciones (74 %) se clasifican correctamente en *valence*, 250 canciones (69 %) se clasifican correctamente en *arousal*, 206 canciones (57 %) se clasifican correctamente en ambas dimensiones, 50 canciones (14 %) no están clasificadas en ninguna de ellas, 61 canciones (17 %) están correctamente clasificadas en *valence* pero no están correctamente clasificadas en *arousal*, 44 canciones (12 %) no están correctamente clasificadas en *valence* pero están correctamente clasificadas en *arousal*. En general, se puede observar que el modelo genera mejores clasificaciones por *valence* que por *arousal*.

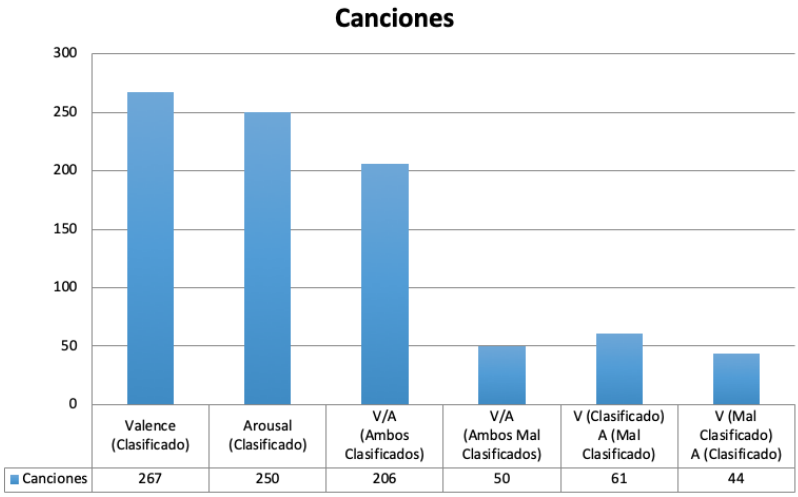


Figura 4.9: Tasa de éxito en el clasificador binario V/A. Elaboración propia.

Sistema de clasificación analizado por la longitud de la ventana

A continuación se analizan las tasas de éxito en el proceso de clasificación a través del proceso de anotación emocional dinámico disponible en el conjunto de datos de MediaEval. Como se presentó en la sección 4.3.4, se ha obtenido un valor promedio de características de bajo nivel y anotaciones emocionales para diferentes longitudes de ventana. En las Figuras 4.10, 4.11 y 4.12 el eje horizontal representa el intervalo de tiempo definido para calcular la ventana promedio en segundos, mientras que el eje vertical representa el comportamiento del *F-measure* frente a la variación de datos temporales en una clasificación emocional dinámica.

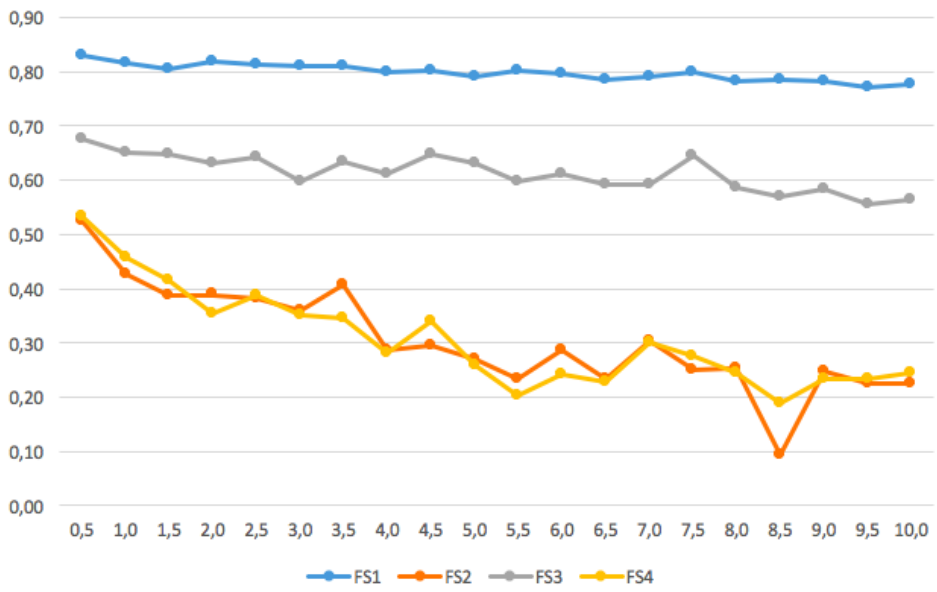


Figura 4.10: Rendimiento para diferentes longitudes de ventana sin estratificación por canción. El eje horizontal es la longitud de la ventana promedio en segundos y el eje vertical es el *F-measure* para cada clase. Elaboración propia.

Aunque en la Figura 4.10 se observan altos valores de *F-measure* para las clases minoritarias con longitudes de ventana cortas, es importante destacar que para este experimento el proceso de estratificación no garantizaba la separación completa de las canciones. El conjunto de datos se elige a partir de ventanas de tiempo separadas aleatoriamente sin tener en cuenta la precaución de separar completamente las canciones en los conjuntos de datos de entrenamiento y de prueba. Esto genera muy buenos resultados, pero se debe a un sobreajuste que pone en duda la capacidad de generalización del clasificador.

En las Figuras 4.11 y 4.12 se realiza un proceso de estratificación antes de promediar. Esto asegura la separación adecuada de todas las ventanas de tiempo de cada canción en el conjunto de datos de entrenamiento y de prueba. La Figura 4.11 muestra los diferentes valores obtenidos de *F-measure* en función de la variación de la longitud de la ventana. Adicionalmente, la Figura 4.12 muestra la misma información pero considerando el balanceo de los datos mediante la técnica *SMOTETomek*. Estos resultados muestran que, en general, la variación de la longitud de la ventana no mejora significativamente las tasas de éxito en el proceso de clasificación. Por otro lado, se puede observar que el balanceo de datos mejora ligeramente los *F-measures* de las clases minoritarias frente a un clasificador multiclase con valores promediados en el tiempo por canción. Esto se puede observar en el mejor vector promedio de *F-measures* [0.72, 0.31, 0.62, 0.34] de la Tabla 4.7, frente al vector de valores más altos de *F-measures* [0.63, 0.30, 0.52, 0.31] presentado en la Figura 4.12.

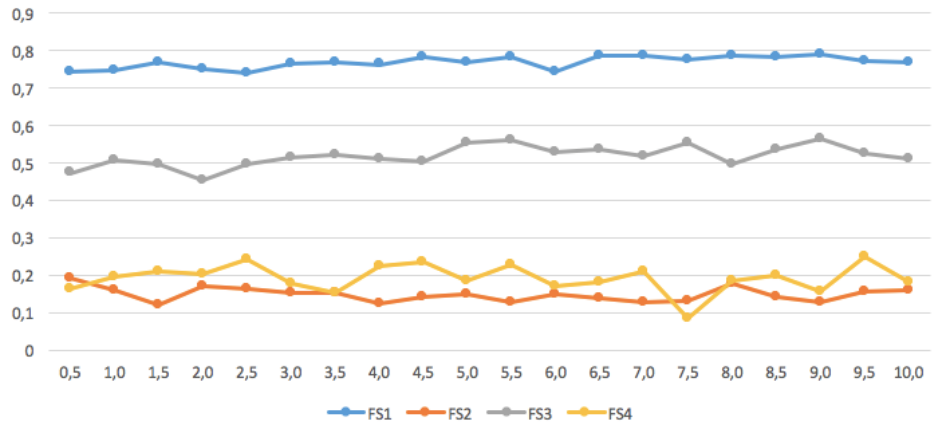


Figura 4.11: Rendimiento para diferentes longitudes de ventana con estratificación por canción. El eje horizontal es la longitud de la ventana promedio en segundos y el eje vertical es el *F-measure* para cada clase. Elaboración propia.

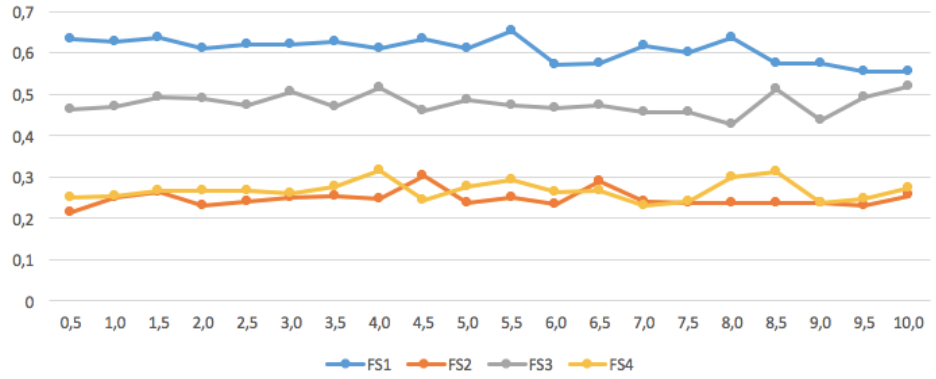


Figura 4.12: Rendimiento para diferentes longitudes de ventana con estratificación y *SMOTETomek* por canción. El eje horizontal es la longitud de la ventana promedio en segundos y el eje vertical es el *F-measure* para cada clase. Elaboración propia.

4.3.6 Consideraciones del tercer prototipo

Este apartado se ha centrado en el desarrollo de un sistema clasificador emocional de la música utilizando el conjunto de datos MediaEval. Inicialmente el sistema clasificador toma como referencia el sistema de predicción previamente diseñado en el apartado 4.1, en concreto, un Perceptrón Multicapa (MLP). Los resultados de la predicción muestran que, a pesar de poder alcanzar valores de RMSE bajos (0.11) en la predicción de *valence* y *arousal*, estos valores no fueron lo suficientemente buenos para extender la funcionalidad del sistema a la clasificación mediante reglas que asocian coordenadas V/A con un cuadrante. Los índices de precisión obtenidos en una clasificación de cuatro cuadrantes dan un valor medio del 34 %. Estos resultados generan la necesidad de desarrollar un nuevo sistema clasificador categórico basado en el mismo conjunto de datos. Partiendo de esta necesidad, se diseña y evalúa tres clasificadores diferentes: *SVM*, *Random Fo-*

rest y *MLP*. También se analiza y optimiza la etapa de pre-procesamiento del conjunto de datos, que es un elemento clave en el desarrollo de este trabajo. En esta etapa, se implementa una estrategia de estratificación dividiendo aleatoria y proporcionalmente el conjunto de datos en los datos de entrenamiento y los datos de prueba, y también se implementa diversas técnicas de balanceo de datos. Los clasificadores se implementaron bajo el enfoque binario, como también bajo el enfoque *one vs. rest*. Y finalmente, también se realizaron varias pruebas variando la duración de las ventanas de tiempo, con el objetivo de analizar si dicha variación podía o no mejorar el desempeño del sistema.

4.4 Conclusiones

En este capítulo se han presentado tres prototipos de sistemas para el reconocimiento de emociones en la música. El primer prototipo utiliza algoritmos de *machine learning* (en particular regresión) para la predicción de valores de *valence* y *arousal*, que luego son interpretados a través de un modelo afectivo dimensional. Aunque dicho prototipo muestra unos valores de error razonables frente a otros trabajos de referencia analizados en el apartado 3.4, no fue posible realizar una extensión de su funcionalidad a la clasificación por cuadrantes, debido a que la tasa de acierto de clasificación por cuadrante resulta ser muy baja en consideración a los trabajos de clasificación estudiados en el apartado 3.6.

Posteriormente, se presentan el segundo y tercer prototipo, ambos prototipos orientados a la clasificación, en donde se suele utilizar un modelo afectivo categórico. El segundo de los prototipos se diseña bajo un enfoque no determinístico, permitiendo realizar la clasificación de piezas musicales con diferentes grados de pertenencia a diferentes categorías. En este prototipo las reglas de inferencia son diseñadas por un experto, lo que puede ser considerado como una desventaja desde el punto de estrategias de recomendación, teniendo en cuenta que la estrategia de *enfoque personalizado* es una de las más exploradas en los más recientes trabajos dentro del campo de sistemas recomendadores, según la revisión de literatura del apartado 3.7; para esta estrategia resulta mucho más conveniente la capacidad de descubrimiento automático de reglas de clasificación que facilita los algoritmos de *machine learning*, que la fijación de reglas estáticas por parte de un experto. Finalmente, el tercer prototipo es diseñado bajo un enfoque determinístico, lo que conlleva a realizar la clasificación de piezas musicales en una sola categoría con grado de pertenencia absoluto, y además implementa algoritmos de *machine learning* para el descubrimiento de las reglas de clasificación (en particular redes neuronales de clasificación). Es importante resaltar, que el desbalanceo de datos por cuadrantes del *dataset* de MediaEval genera una mayor especialización de los modelos de *machine learning* sobre la clase mayoritaria, y al mismo tiempo, altos niveles de error para clasificar correctamente las otras clases. Con base en ello, fue necesario implementar algunas estrategias de balanceo de datos que permitieron mejorar el desempeño general de los modelos, presentando unos valores de precisión destacables frente a la revisión del estado del arte del apartado 3.6. Esta situación también muestra la importancia de desarrollar nuevos *datasets* balanceados y con mayor diversidad en etiquetas, como también, con procesos de etiquetado de calidad.

Desde una visión mucho más general de los resultados obtenidos, es importante resaltar que este capítulo también presenta una discusión muy importante sobre las ventajas y desventajas identificadas sobre cada enfoque, técnica, y algoritmo utilizado. Lo que da lugar a profundizar la investigación y exploración en el campo de MER, y que dependiendo de las necesidades particulares de un nuevo problema de investigación, o de la variación del entorno de estudio, esta discusión puede llegar a plantear nuevos direccionamientos hacía nuevas contribuciones. El siguiente paso que se plantea en esta Tesis (Capítulo 5) es el diseño de un nuevo *dataset* que responda en gran parte a las limitaciones identificadas en la Sección 3.2.2. Parte de los prototipos resultantes, en especial el reconocimiento del *valence* y *arousal* sobre el tiempo que permite realizar el primer prototipo, como también el nuevo *dataset*, serán los principales elementos para el diseño del prototipo de sistema recomendador que se presenta en el capítulo 6.

Capítulo 5

Diseño del *dataset* musical: *Emotional Non-Superstar Artist-Dataset (ENSA)*

Los *datasets* musicales son uno de los elementos más importantes para promover y facilitar el desarrollo de nuevos sistemas recomendadores. Aunque en la actualidad existen algunos *datasets* disponibles, éstos presentan ciertas limitaciones para los objetivos de esta Tesis. Entre estas limitaciones se resalta la importancia de la disponibilidad de características que permitan realizar un tratamiento de los sesgos preexistentes generados por el efecto de la popularidad, como también la falta de características adicionales que permitan modelar nuevos y mejores descriptores musicales y emocionales. Este capítulo presenta en detalle la preparación de un nuevo *dataset* (Apartado 5.1), el contenido del *dataset* (Apartado 5.2), el análisis del *dataset* (Apartado 5.3), y finalmente las conclusiones (Apartado 5.4).

5.1 Preparación del dataset

A partir de las limitaciones encontradas en las bases de datos musicales revisadas en el apartado 3.2, como también del análisis de sesgos realizado en el apartado 3.8, se evidencia la necesidad de diseñar un nuevo *dataset* para cubrir algunos aspectos particulares y novedosos para el campo de los sistemas recomendadores musicales, como lo son: la inclusión de canciones completas y originales por parte de artistas noveles, la inclusión de la estructura musical de la canción a través de una *metadata*, la evaluación emocional por parte del artista, y la evaluación *like/dislike* en complemento a la evaluación emocional por parte del oyente. La base de datos que se diseña y explica en este apartado lleva como nombre ENSA, que hace alusión al nombre completo *Emotional Non-Superstar Artist-Dataset*, y sus características novedosas pueden evidenciarse en la Tabla 5.1, en donde se incluyen los datasets revisados previamente en el apartado 3.2 para facilitar la comparación con ENSA. Este nuevo dataset será utilizado en el diseño del sistema recomendador que se presenta en el capítulo 6, fundamentalmente para mitigar el sesgo preexistente que normalmente promueve las estrategias de recomendación

basadas en la popularidad de canciones que suelen ser muy famosas.

Tabla 5.1: Comparativo de datasets musicales con ENSA

Dataset	Año	Archivos	Duración	Audio Disponible	Estructura Musical	Modelo Afectivo	Artistas No-Famosos	Etiquetado emocional por artista	Etiquetado emocional por oyente
GTZAN [85]	2002	1000	30s	✓	✗	None	✗	✗	✗
Ballroom [86]	2006	698	≈30s	✓	✗	None	✗	✗	✗
MagnaTagATune [87]	2009	25,850	≈30s	✓	✗	None	✗	✗	✗
Million Song Dataset [88]	2011	1M	-	✗	✗	Categorical	✗	✗	✓
UrbanSound8k [89]	2014	8732	⩽4s	✓	✗	None	✗	✗	✗
ESC-50 [90]	2015	2000	5s	✓	✗	None	✗	✗	✗
TUT Acoustic Scene [91]	2016	1560	30s	✓	✗	None	✗	✗	✗
Mediaeval [81]	2016	1744 58	≡45s [46s, 627s]	✓	✗	Dimensional	✗	✗	✓
AudioSet [92]	2017	≈2.1M	10s	✗	✗	Categorical	✗	✗	✗
ENSA-Dataset	2021	60	≈234s	✓	✓	Categorical	✓	✓	✓

Para la preparación del conjunto de datos se llevaron a cabo los siguientes pasos:

- Invitación para los artistas: Para obtener las canciones originales, se compartió una invitación a través del correo electrónico y de algunas redes sociales en la que se preguntaba a los artistas noveles del Valle del Cauca en Colombia si querían formar parte de esta investigación. Los artistas interesados en participar aportaron sus canciones originales bajo una Licencia *Creative Commons No Comercial* que permite el libre acceso a los datos de las obras sin ánimo de lucro y que, además, exige siempre referenciar la autoría de las piezas musicales de cada artista. Las canciones en su formato digital fueron publicadas en el GitHub del tesista[1], para que posteriormente la comunidad científica tenga acceso a ellas.

- Entrevista con los artistas: Se realizaron entrevistas individuales con cada uno de los artistas para conocer sus **estrategias de composición** y la **estructura musical** que definían en sus canciones. Todas estas entrevistas fueron realizadas por videoconferencia, y tuvieron una duración aproximada de 30 minutos. En la parte inicial de la entrevista se explicó a los artistas en más detalle el objetivo del estudio de investigación y su participación, dejando claro que podrían revocar su voluntad de participación en cualquier momento. Las entrevistas se desarrollaron de manera semi-estructurada, en la cual se abordaba con cada artista su forma de componer, la descripción de las canciones, y las estructuras musicales que más utilizaba. La información obtenida de estas entrevistas fue muy importante porque permitió comprender que no todos los artistas utilizan los mismos métodos de composición, algunos de ellos aplican métodos muy organizados mientras que otros hacen uso de la improvisación y son menos estructurados en cuanto a la metodología, pero en todos los casos, el punto de partida es la inspiración personal, la cual tiene una conexión directa con las emociones que el artista quiere transmitir con sus composiciones. En cuanto a la estructura de la canción, los artistas destacan que con ella pueden crear la experiencia emocional que quieren comunicar [34]. Muchos de ellos utilizan una estructura bien conocida que incluye una introducción, un puente, un verso, un puente, un coro, un verso, un solo, un coro. Esto demuestra la importancia de comprender la estructura de una canción y analizar los cambios emocionales entre las partes, y por este motivo se pidió a cada artista que indicara el principio y el final del verso y el coro de cada canción, así como

[1]Disponible en: https://github.com/yesidospitiamedina/ENSA/

la etiqueta emocional para el verso, el coro, y una aproximación para la canción completa (etiqueta emocional global); este etiquetado emocional del artista se encuentra orientado a su interés por evocar esas emociones. Finalmente, los artistas también suministraron el **género musical** de cada una de las obras musicales, explicando las razones de dicha clasificación. Toda la información referente al etiquetado por parte de los artistas también se encuentra publicada en el GitHub del tesista[2].

Los artistas utilizaron 8 adjetivos para etiquetar su percepción emocional (excitado, feliz, alarmado, enfadado, deprimido, aburrido, tranquilo, satisfecho) adoptados del modelo circumplejo afectivo [4] de James A. Russell, el cual fue presentado en el apartado 2.3. El modelo afectivo adoptado para el experimento se muestra en la Figura 5.1. En el proceso de etiquetado de canciones por género musical, los artistas asignaron en muchos casos géneros muy específicos que podrían considerarse subgéneros. Para generar una agrupación más general de las canciones por género musical, algunos de estos géneros se indicaron de forma más general, como en el caso del Metal, que incluye subgéneros como el Heavy Metal y el Death Metal, esto permitió reducir el número de géneros musicales de 21 a 16. Además, a veces el artista definía una mezcla de géneros para algunas canciones, como Funk Blues, Blues Folk, Country Blues. Para estos casos, cada canción fue clasificada con el género musical predominante, considerando el que tenía mayor influencia en los diferentes elementos musicales de cada canción.

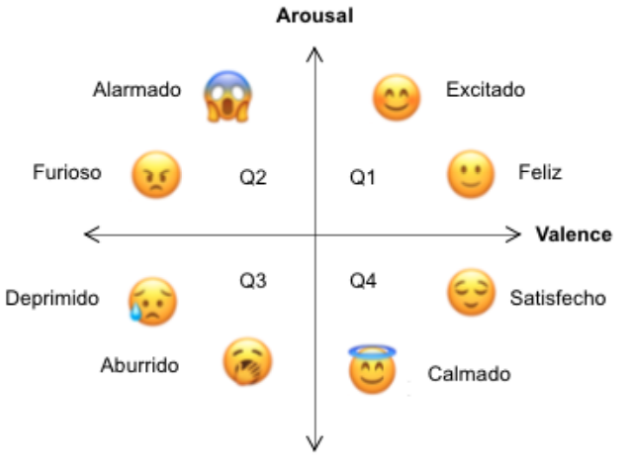


Figura 5.1: Modelo Afectivo. Elaboración propia.

[2]Disponible en: https://github.com/yesidospitiamedina/ENSA/tree/main/Emotions_evoked_by_artists

En la Figura 5.2 se muestra una captura de pantalla del instrumento en *Excel* diligenciado por uno de los artistas. Para este caso es importante destacar que este artista en particular tiene 3 canciones sin estructura musical, por lo que solo suministra etiqueta global. Para los otras 2 canciones, que si cuentan con estructura musical, el artista indica la etiqueta emocional en verso, coro, y global.

Artista	Canción	Verso	Coro	Global
Lina Cardona Herrera	Currulao de nieves	Excitado	Calmado	Feliz
Lina Cardona Herrera	Mi casita de caña			Feliz
Lina Cardona Herrera	Opening - Marianela y la caperucita roja			Feliz
Lina Cardona Herrera	Pessoa - Antología de Poemas			Excitado
Lina Cardona Herrera	El swing de Caperucita	Feliz	Feliz	Feliz

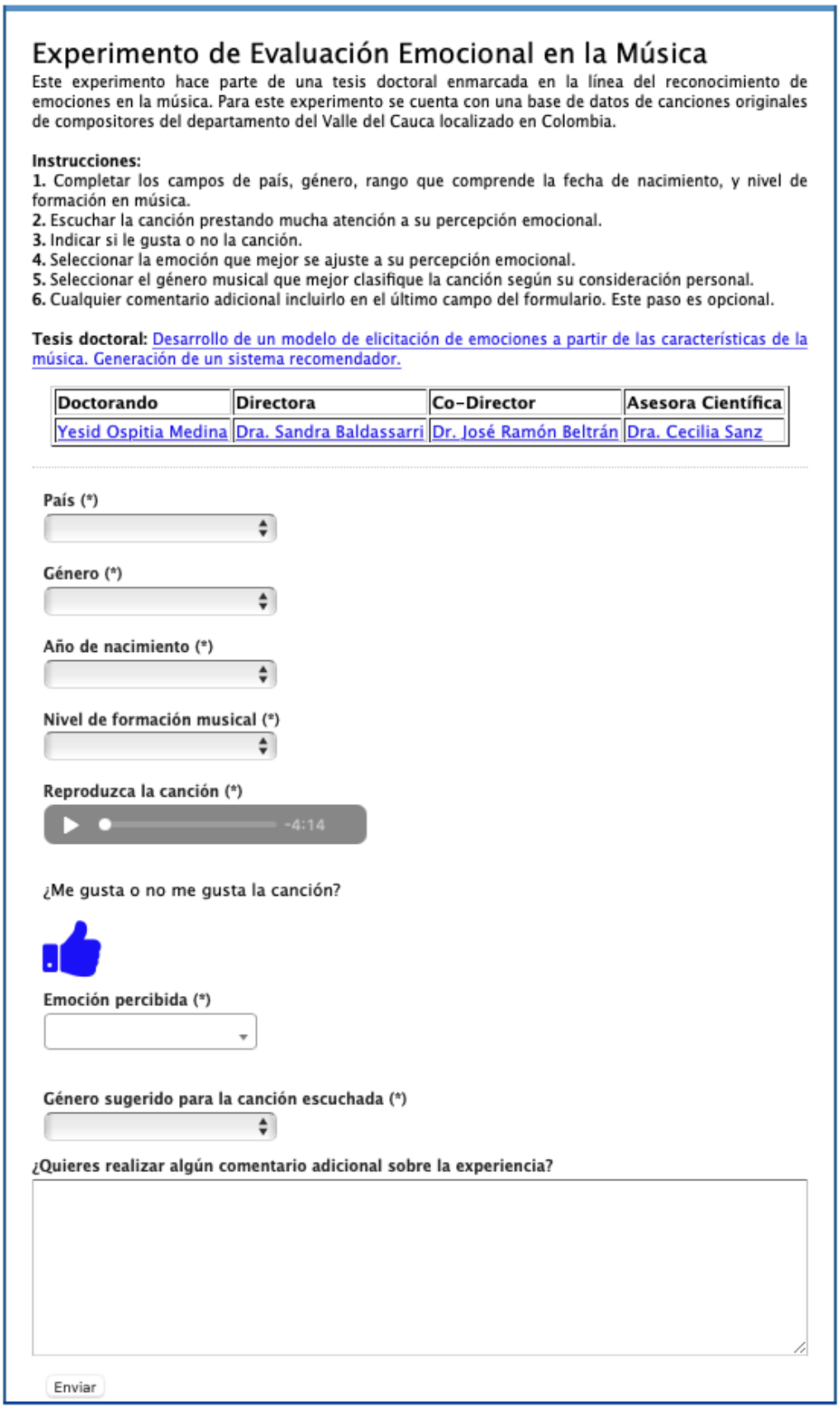
Figura 5.2: Etiquetado emocional por parte del artista. Captura de pantalla.

- Cuestionario para los oyentes[3]**:** Con el fin de validar si la emoción que el artista pretende transmitir en su canción es la misma que percibe cualquier oyente, se elaboró un cuestionario web a través del cual el oyente podía escuchar una canción e indicar (con los mismos adjetivos utilizados por el artista) cuál es su percepción emocional de toda la canción. Además, el oyente también puede especificar si le gusta o no la canción, seleccionar el género musical que considera apropiado para la canción, y escribir a través de un campo complementario cualquier observación o detalle que considere importante explicar sobre su experiencia. También hay otros datos demográficos importantes que se solicitan para caracterizar al usuario, como el género, el rango de edad, y el nivel de conocimiento musical (oyente de música, practicante amateur, experiencia intermedia, experiencia avanzada), los campos completos se pueden ver en la Figura 5.3. Este cuestionario fue diligenciado de manera anónima y bajo el cumplimiento de la ley de protección de datos dispuesta en el decreto 1377 de 2013 del Gobierno de Colombia.

5.2 Definición del contenido del dataset

El conjunto de datos se encuentra conformado por un total de 234 minutos de sonido, lo que equivale a 3.91 horas o 468 clips de audio de 30 segundos. Desde el punto de vista musical, el conjunto de datos incluye 60 canciones completas, lo que significa que se incluye la obra original del artista sin ningún recorte. Las especificaciones y cada uno de los atributos relevantes del conjunto de datos se detallan en la Tabla 5.2. Es importante señalar que no todas las canciones siguen una estructura musical con verso y coro; sólo 39 canciones tienen esta característica. Además, se han marcado las canciones con voz femenina, masculina y sin voz (instrumentales).

[3]Disponible en: http://104.237.5.250/evaluacionensa/form.php



Experimento de Evaluación Emocional en la Música

Este experimento hace parte de una tesis doctoral enmarcada en la línea del reconocimiento de emociones en la música. Para este experimento se cuenta con una base de datos de canciones originales de compositores del departamento del Valle del Cauca localizado en Colombia.

Instrucciones:
1. Completar los campos de país, género, rango que comprende la fecha de nacimiento, y nivel de formación en música.
2. Escuchar la canción prestando mucha atención a su percepción emocional.
3. Indicar si le gusta o no la canción.
4. Seleccionar la emoción que mejor se ajuste a su percepción emocional.
5. Seleccionar el género musical que mejor clasifique la canción según su consideración personal.
6. Cualquier comentario adicional incluirlo en el último campo del formulario. Este paso es opcional.

Tesis doctoral: Desarrollo de un modelo de elicitación de emociones a partir de las características de la música. Generación de un sistema recomendador.

Doctorando	Directora	Co-Director	Asesora Científica
Yesid Ospitia Medina	Dra. Sandra Baldassarri	Dr. José Ramón Beltrán	Dra. Cecilia Sanz

País (*)

Género (*)

Año de nacimiento (*)

Nivel de formación musical (*)

Reproduzca la canción (*)

-4:14

¿Me gusta o no me gusta la canción?

Emoción percibida (*)

Género sugerido para la canción escuchada (*)

¿Quieres realizar algún comentario adicional sobre la experiencia?

Enviar

Figura 5.3: Cuestionario para los oyentes. Elaboración propia.

Tabla 5.2: Especificaciones del dataset

Atributo	**Detalle**
Licencia	BY-NC
Codificador de sonido	MPEG layer 3 (MP3)
Tiempo total	234 minutes
Clips de 30 segundos	468 clips
Cantidad de canciones	60 canciones
Cantidad de artistas	10 noveles
Características de bajo nivel	260 extraídas cada 500 ms
Versos identificados	39 versos
Coros identificados	39 coros
Estructura musical sin definir	21 canciones
Géneros musicales especificados	21 géneros
Géneros musicales agrupados	16 géneros
Canciones con voz femenina	15 canciones
Canciones con voz masculina	39 canciones
Canciones sin voz (instrumental)	4 canciones
Canciones con voz de niño	2 canciones
Modelo afectivo	Categórico (8 emociones)
Etiquetado emocional por el artista	39 versos
	39 coros
	60 Canciones nivel global
Etiquetado emocional por el oyente	106 anotaciones
(Like / dislike) marcado por el oyente	106 anotaciones

La Tabla 5.3 presenta la distribución de las canciones según la percepción emocional del artista sobre sus propias obras musicales. Como se ha mencionado anteriormente, algunas canciones definen en su estructura el verso y el coro, 39 canciones siguen esta estructura y por esta razón, se ha etiquetado la percepción emocional en verso, coro, y la canción completa. Para las otras 21 canciones, los artistas indicaron que no siguen ninguna estructura en particular, y por esta razón, sólo se han etiquetado como canción completa. La Tabla 5.4 muestra la distribución de las canciones por género musical, adicionalmente, se específica el tipo de voz (femenina, masculina, infantil, sin voz). Además de los archivos de sonido, también se incluyen archivos de *metadatos* que contienen el etiquetado emocional y de género musical por artista y oyente, el etiquetado de los versos y los coros, las características de bajo nivel extraídas para cada canción completa cada 500 ms a través de la herramienta *OpenSmile*[4], y la información complementaria proporcionada a través del cuestionario de los oyentes.

[4]http://opensmile.sourceforge.net/

Tabla 5.3: Distribución de canciones por emociones y estructura musical según el artista

Cuadrante	Emoción	Verso	Coro	Global
Q1	Excitado	8	6	11
	Feliz	2	8	8
Q2	Alarmado	6	5	2
	Furioso	7	4	16
Q3	Deprimido	5	11	7
	Aburrido	0	0	0
Q4	Calmado	9	4	8
	Satisfecho	2	1	8

Tabla 5.4: Distribución de canciones por género musical

Género musical	Femenino	Masculino	Niño	Sin Voz
Balada	-	4	-	-
Balada Rock	-	2	-	-
Blues	-	4	-	-
Bossa nova	1	-	-	-
Música clásica	-	-	-	1
Currulao	1	-	-	-
Country	-	1	-	-
Folk	11	-	-	-
Funk	-	2	-	-
Jazz	1	-	-	-
Música latina	-	-	1	-
Metal	-	21	-	-
Pop	-	-	-	2
Pop Rock	1	-	-	-
Rock	-	5	-	1
Swing	-	-	1	-

5.3 Análisis del dataset

En este apartado se presenta un análisis del ENSA-Dataset para comprobar el nivel de coincidencia entre las emociones y los géneros musicales indicados por los artistas y por los oyentes. Con respecto a las emociones, el análisis se realizó a nivel global de toda la canción tratando de identificar los cuadrantes y semiplanos en los que hay coincidencias; en esta comparación, fue realmente útil considerar también la percepción emocional según el sesgo que puede generar la evaluación *me gusta (Like)* o *no me gusta (Dislike)* indicada por el oyente. En cuanto al género musical, comparamos el género indicado por el artista y por el oyente, analizando el nivel de coincidencia, y teniendo en cuenta que algunos géneros pueden ser muy parecidos y para el oyente más común es difícil diferenciarlos.

Para este estudio, se generaron 106 anotaciones a partir del cuestionario que se

puede ver en la Figura 5.3. El cuestionario está diseñado para seleccionar una canción de manera aleatoria para ser evaluada cada vez que el usuario accede a él, garantizando que si se trata de la misma sesión de usuario, el usuario pueda seguir calificando sin repetir canción. En este experimento 46 canciones diferentes recibieron entre 1 y 4 evaluaciones por diferentes oyentes.

La Tabla 5.5 presenta las coincidencias exactas de las emociones etiquetadas por los artistas y los oyentes desde dos perspectivas, la primera es el caso en que al oyente le gusta la canción, y la segunda es el caso en que al oyente no le gusta la canción. Uno de los hallazgos más importantes que se pueden evidenciar es que se generan más coincidencias de etiquetado emocional cuando el usuario evalúa canciones que le gustan. Los cuadrantes Q1 (Excitado, Feliz) y Q4 (Calmado, Satisfecho) son los cuadrantes con mayor número de coincidencias (ver Tabla 5.6), de la misma manera, es el semiplano de *valence* el que muestra el mayor número de coincidencias en la Tabla 5.7.

Tabla 5.5: Análisis de las coincidencias por emociones específicas

Cuadrante	**Emoción**	**Concidencias con Like**	**Coincidencias con Dislike**
Q1	Excitado	5	1
	Feliz	7	0
Q2	Alarmado	0	0
	Furioso	2	1
Q3	Deprimido	1	1
	aburrido	0	0
Q4	Calmado	7	0
	Satisfecho	1	0

Tabla 5.6: Análisis de las coincidencias por cuadrantes

Cuadrante	**Coincidencias Like**	**Coincidencias Dislike**
Q1	17	2
Q2	10	4
Q3	2	1
Q4	14	0

Tabla 5.7: Análisis de las coincidencias por valencia y excitación

Dimensión	**Concidencias con Like**	**Coincidencias con Dislike**
Valence	66	11
Arousal	57	10

La Tabla 5.8 presenta las coincidencias por emociones, cuadrante y dimensiones V/A con respecto a los géneros musicales marcados por el artista, en aquellas canciones que han sido evaluadas por los oyentes en este experimento. La idea principal de este análisis es destacar que para algunos géneros musicales, como el Blues, el Folk y el

Metal, se presentaron mayores coincidencias entre el proceso de evaluación emocional del artista y el del oyente (casos de likes). Esto sugiere que tal vez algunos géneros musicales estén más relacionados con ciertos estímulos emocionales que se perciben más claramente. En cuanto a las coincidencias entre los géneros musicales etiquetados por el artista y los géneros musicales etiquetados por los oyentes, se observa en la Tabla 5.9 que los géneros de Metal y Blues son los que presentan mayor número de coincidencias. Al igual que en los análisis anteriores centrados en la percepción emocional, se encuentra que, en cuanto al género musical, el mayor número de coincidencias se encuentra también con las canciones que le gustan al oyente.

Tabla 5.8: Análisis de las coincidencias entre el género musical (especificado por el artista), las emociones, los cuadrantes y las dimensiones V/A. Sólo casos de likes.

Musical genre	**Emotion**	**Quadrant**	**Valence**	**Arousal**
Balada	-	-	2	2
Balada Rock	-	1	1	2
Blues	5	7	7	9
Bossa nova	-	1	2	1
Música Clásica	-	-	3	-
Currulao	1	1	3	1
Country	-	-	-	-
Folk	6	12	17	15
Funk	2	2	3	2
Jazz	-	-	-	-
Música Latina	-	-	-	-
Metal	4	12	14	15
Pop	1	1	6	1
Pop Rock	-	-	-	1
Rock	2	4	6	5
Swing	2	2	2	3
Total	23	43	66	57

También se analizó el perfil del oyente para averiguar si tiene alguna relación con el número de coincidencias entre los géneros musicales que han sido etiquetados por el artista y los géneros musicales que han sido etiquetados por el oyente. Para ello, se ha analizado los resultados que se muestran en la Tabla 5.10, que refleja que un mayor nivel de conocimiento musical en el perfil del usuario aumenta las posibilidades de etiquetar correctamente el género musical.

Tabla 5.9: Análisis de las coincidencias por género musical

Género musical	Coincidencias con likes	Coincidencias con dislikes
Balada	1	-
Balada Rock	2	1
Blues	4	1
Bossa nova	-	-
Música Clásica	2	1
Currulao	3	-
Country	-	-
Folk	1	-
Funk	-	-
Jazz	-	-
Música Latina	-	-
Metal	14	5
Pop	-	-
Pop Rock	-	-
Rock	2	-
Swing	1	-
Total	30	8

Tabla 5.10: Análisis de coincidencias por género musical (Like & Dislike) según el perfil del oyente

Perfil	Coincidencias	Total de evaluaciones	% Coincidencias
Oyente	27	77	35 %
Aficionado	6	18	33 %
Intermedio	3	7	43 %
Avanzado	2	4	50 %
Total	38	106	36 %

5.4 Conclusiones

Este capítulo se ha centrado en el diseño del ENSA-Dataset, un novedoso conjunto de datos de canciones de artistas noveles (*non-superstar*), así como en el análisis de las etiquetas emocionales y de género musical realizadas por artistas y oyentes. Uno de los hallazgos más interesantes corresponde al efecto que tiene el etiquetado *like/dislike* de los oyentes sobre el número de coincidencias de etiquetado de emociones y géneros musicales entre oyentes y artistas; por una parte, en aquellos casos en que el oyente etiqueta las canciones con *like*, se encuentra un mayor número de coincidencias de etiquetado de emociones y géneros musicales con los artistas, por otra parte, en los casos de *dislike*, se encuentran menores coincidencias. Este resultado sugiere que en un proceso de evaluación emocional de piezas musicales, la percepción emocional del oyente puede estar influenciada por sus preferencias musicales, generando situaciones

de evaluación en donde se etiquete una determinada canción con emociones negativas debido a que dicha canción no se ajusta a las preferencias musicales del oyente (no es de su gusto). Para estos casos, las etiquetas emocionales resultantes podrían tener poca relevancia para las estrategias de recomendación, e incluso serían inapropiadas, pues podrían distorsionar el aprendizaje efectivo sobre las preferencias reales del oyente. En cuanto al nivel de coincidencias de etiquetado de género musical entre oyentes y artistas, se evidencia una relación con el nivel de formación musical del oyente. En la medida en que el oyente tiene un mayor nivel de formación musical, las coincidencias de género musical son mayores. Cuando el oyente se caracteriza por un nivel bajo de formación musical, las coincidencias de género musical son menores.

Desde el punto de vista de sistemas recomendadores, el ENSA-Dataset se constituye como una interesante contribución para la comunidad científica, teniendo en cuenta que a pesar de la existencia de algunos *datasets* musicales, estos son limitados, y muchos de ellos no cuentan con las características del *ENSA-Dataset*, en donde se destaca la disponibilidad de piezas musicales completas, originales, y compuestas por artistas noveles; lo que permite abordar varias de las limitaciones en sistemas recomendadores generadas por los sesgos presentados en el apartado 3.8.

El siguiente paso que se plantea en esta Tesis (Capítulo 6) es el diseño de un prototipo de sistema recomendador, a partir de la utilización del *ENSA-Dataset*, como también de algunos de los resultados obtenidos en el capítulo 4.

Capítulo 6

Desarrollo de un sistema de recomendación musical

Este capítulo presenta el diseño y desarrollo de un prototipo de sistema recomendador para el campo de los MRS, partiendo de los diferentes análisis realizados en los apartados 3.7 y 3.8, en donde respectivamente fue abordado el estado del arte de las estrategias de recomendación más conocidas, y los sesgos potenciales relacionados con cada una de ellas. Inicialmente, se presenta el diseño del sistema teniendo en cuenta la importancia de la estructura musical en las canciones, como también, las implicaciones que ésta tiene en las decisiones del diseño (Apartado 6.1). Luego, se detallan los experimentos y el análisis de los resultados obtenidos (Apartado 6.2). Finalmente, se presentan las conclusiones (Apartado 6.3).

6.1 Diseño del sistema

Existen varias formas de clasificar la música, como por ejemplo, el género musical, si es instrumental o no, si es adecuada o no para bailar (y en que grado), si se interpreta en vivo o no, si es apropiada o no para una actividad específica (deporte, sueño, relajación), etc. Es importante resaltar que muchos de estos criterios se encuentran relacionados con estados de ánimo, o con algún interés en elicitar emociones [8]. También es importante resaltar que a pesar de la existencia de muchos trabajos de investigación sobre el reconocimiento de emociones en la música, una gran parte de estos trabajos se centran en modelos categóricos, mientras una menor cantidad lo hacen sobre modelos dimensionales [4]. En el caso de la clasificación, se utilizan adjetivos para las emociones específicas; en cuanto a los modelos dimensionales, se utiliza una coordenada en un plano bidimensional para establecer el *valence* emocional (valencia positiva o negativa), y el *arousal* (nivel de energía, excitación y/o intensidad).

También es importante señalar que las canciones, desde la perspectiva de la ingeniería del sonido, se analizan como señales digitales y, desde este enfoque, hay muchos trabajos que aplican estrategias de filtrado basadas en el contenido y en las características de bajo nivel del sonido. Esto sugiere que la perspectiva artística de la música puede tener cierto sesgo en las investigaciones actuales, ya que aunque la música puede

tratarse como sonido, los artistas la ven como un arte que se basa en sus emociones y se describe a través de la teoría musical. Una forma de transmitir la emoción del artista es a través de la estructura de la canción, teniendo en cuenta que la acción de escuchar música es una experiencia emocional que se produce a lo largo del tiempo [34]. El artista utiliza todas sus herramientas de composición musical para crear experiencias emocionales que pueden fluctuar en *valence* y *arousal*. De este modo, el oyente puede experimentar una energía cambiante cuando la canción pasa del verso al puente, y luego al coro.

Considerando la relevancia de representar la percepción emocional a lo largo del tiempo en la música, se decide utilizar el enfoque de series temporales, en el que se diseña una serie temporal para representar la percepción emocional de cada parte importante de la estructura de una canción teniendo en cuenta el nivel de *valence* y *arousal*. El diseño del sistema y sus principales fases pueden verse en la Figura 6.1.

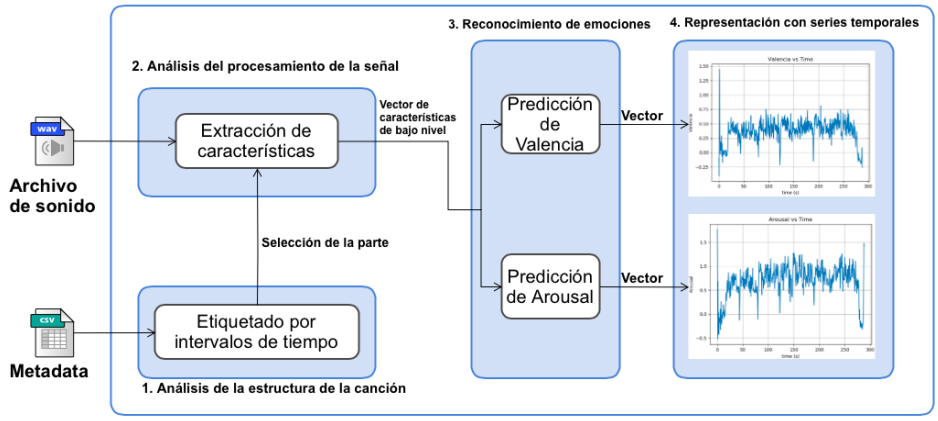


Figura 6.1: Diseño del sistema basado en series temporales. Elaboración propia.

A continuación se explica cada una de las fases:

1. **Análisis de la estructura de la canción:** la estructura de la canción es muy importante porque el compositor la utiliza para crear la experiencia musical a través de la cual el oyente percibe emociones con diferentes niveles de intensidad. Una de las estructuras más conocidas en la música popular occidental incluye una introducción, un puente, un verso, un puente, un coro, un verso, un solo, un coro. Aunque en general suele predominar una emoción (especialmente en las producciones musicales modernas), esta emoción puede presentar una intensidad diferente para cada parte de la estructura, y pueden aparecer casos excepcionales en los que incluso el *valence* es diferente. En esta fase inicial del sistema, con la ayuda de cada compositor, se construye un archivo CSV en el que se definen las diferentes partes de la estructura de sus canciones originales y se utilizan como *metadatos*, indicando el momento en que cada una comienza y termina en la línea de tiempo.

2. **Análisis del procesamiento de la señal:** en esta fase se aplica una herramienta

de extracción de características de sonido. Para este sistema, *OpenSMILE*[1] se ha utilizado para extraer 260 características de bajo nivel cada 500 ms. Los intervalos de extracción vienen determinados por la fase que define la estructura de la canción, por lo que, dependiendo de la parte que se analice, hay que ajustar el tiempo de extracción. Esta fase genera como salida una matriz bidimensional de M x 260, donde M será el número de ventanas de tiempo determinado por la duración total de la parte de la estructura de la canción que interesa analizar. Es importante resaltar que la utilización de *OpenSMILE* y no de otra librería, se debe a que el prototipo resultante del apartado 4.1 fue entrenado con las características disponibles del *dataset* MediaEval, las cuales fueron extraídas a través de *OpenSMILE*.

3. **Reconocimiento de emociones:** para el reconocimiento de emociones se utilizan dos redes neuronales predictivas previamente entrenadas con características de bajo nivel en el apartado 4.1, una para reconocimiento del *valence* y otra para el reconocimiento del *arousal*. Como salida de esta fase se obtienen dos vectores numéricos, en los que cada uno presenta valores entre -1 y 1, y que indican respectivamente el nivel del *valence* y del *arousal* [131].

4. **Representación de series temporales:** la variabilidad del *valence* y del *arousal* a lo largo del tiempo se dibuja en un plano bidimensional, lo que permite obtener una visión emocional de la parte de la estructura de una canción concreta que se desea analizar.

Las matrices finales obtenidas mediante el proceso descrito anteriormente pueden interpretarse como descriptores emocionales de la canción, orientados a sus diferentes partes de la estructura musical. El interés de representar estos descriptores como series temporales se debe principalmente a dos razones:

1. Las series temporales pueden analizarse mediante técnicas de similaridad, lo que es muy útil para determinar el nivel de cercanía de un vector de *valence* o *arousal* de una canción particular con respecto a otros vectores de canciones diferentes.

2. Las mismas métricas de similaridad pueden utilizarse para implementar técnicas de agrupamiento (*clustering*), lo que permite agrupar canciones de acuerdo a niveles de cercanía con respecto a *valence* y *arousal* considerando las distancias de *intra-cluster* e *inter-cluster*. Con la definición de estos grupos, es posible aplicar una estrategia de recomendación basada en la cercanía de las canciones situadas dentro del mismo grupo (*cluster*).

En general, existen dos métricas muy conocidas para analizar la similaridad en las series temporales, la distancia de coincidencia euclidiana y la distancia de coincidencia de deformación temporal dinámica (DTW).

1. **Distancia euclidiana:** la comparación entre las series temporales se realiza punto a punto y siguiendo el orden en que se presentan estos puntos como se muestra

[1] http://opensmile.sourceforge.net/

en la expresión matemática 6.1.1, en la que x y y representan puntos situados en series temporales diferentes pero que se corresponden secuencialmente en el tiempo. Además, tiene una restricción importante en cuanto a que ambas series temporales deben tener el mismo tamaño (número de medidas); lo que desde el punto de vista de la comparación de versos, o coros, no es conveniente porque las duraciones suelen ser diferentes entre canciones, y las variaciones en la percepción emocional difícilmente se presentarán perfectamente sincronizadas en el tiempo.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \tag{6.1.1}$$

2. **DTW:** este algoritmo calcula inicialmente el mejor camino de alineación entre dos series temporales, dicha propiedad se conoce como capacidad de elasticidad. DTW permite alinear puntos desfasados en el tiempo de series temporales de diferentes longitudes, para posteriormente calcular el nivel de similaridad, lo cual resulta realmente conveniente para el caso de la música, en donde normalmente la longitud de tiempo varía entre las diferentes partes de estructura de una canción a otra. En la expresión matemática 6.1.2 x y y representan puntos de diferentes series temporales, y los subíndices i and j varían sobre π, siendo π el camino óptimamente alineado.

$$DTW(x, y) = \sqrt{\sum_{(i,j)\in\pi} |x_i - y_j|^2} \tag{6.1.2}$$

La comparación gráfica de ambas métricas se puede ver en la Figura 6.2, y teniendo en cuenta los argumentos anteriores, DTW es la métrica más conveniente.

Para el desarrollo del sistema se utiliza la librería *TSLEARN*[2], la cual implementa diversos algoritmos de similaridad basados en series temporales [147]. *TSLEARN* se ejecuta sobre lenguaje *Python* y permite implementar la distancia euclidiana, DTW, y softDTW, el último es una variación de DTW que incluye un hiper-parámetro (valor entre 0 y 1) mediante el cual es posible suavizar el proceso de determinación de la mejor alineación entre dos series temporales [148]. Este enfoque es adecuado para aquellos casos especiales en los que la aplicación de DTW genera saltos inadecuados (según el caso de estudio) para asociar los puntos de dos series temporales y determinar su alineación óptima.

[2]https://tslearn.readthedocs.io/en/stable/

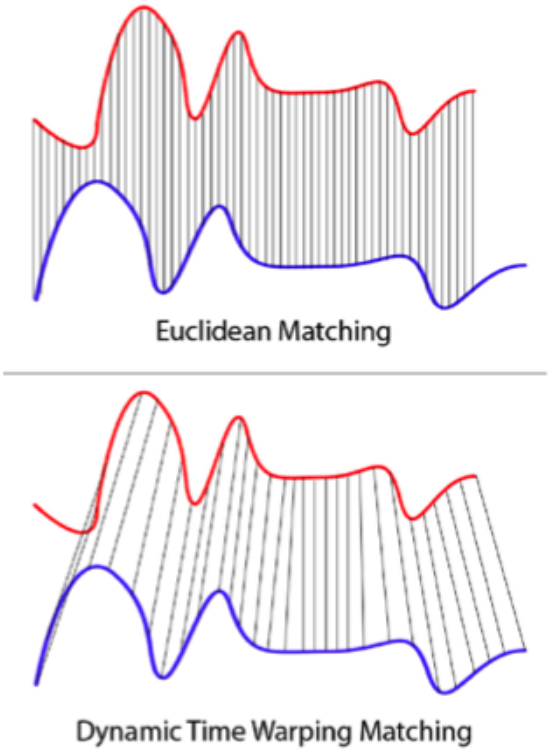


Figura 6.2: Comparación de las medidas de distancia euclidiana y DTW (Figura extraída de [7])

.

6.2 Experimentos y resultados

En este apartado se presentan los resultados obtenidos a partir de dos experimentos realizados con el *dataset* ENSA. En la sección 6.2.1 se calcula el valor DTW para todo el *dataset*, y se analizan las 10 canciones más cercanas a la canción comercial *Hysteria*. Luego, en la sección 6.2.2 se aplican algunas técnicas de agrupamiento sobre el *dataset*, incluyendo nuevamente la canción comercial *Hysteria*, y se presenta un análisis sobre los resultados obtenidos en el proceso de agrupamiento.

6.2.1 Resultados con métricas de similaridad

Para analizar la contribución práctica de las medidas de similaridad, se ha diseñado un experimento para recomendar canciones de artistas noveles (*non-superstar*) basándose en una canción muy famosa de un artista comercialmente muy conocido, teniendo en cuenta la variabilidad de la percepción emocional del verso. En este experimento la canción *Hysteria* de la banda *Def Leppard*[3] es tomada como referencia, para luego ordenar las canciones del *dataset* musical ENSA de acuerdo al nivel de similaridad. Las Tablas 6.1 y 6.2 muestran las distancias de similaridad basadas en el algoritmo DTW

[3]https://www.defleppard.com

para la canción comercial, y sus 10 canciones más similares de artistas noveles de un conjunto total de 39 (las que incluyen el verso en su estructura). Hay que tener en cuenta que un número menor significa que está más cerca, por lo que se puede observar que la canción *Hysteria* tiene una distancia de 0 con respecto a sí misma, con respecto al *valence* la canción no famosa más cercana es *Pensarás en mí* y con respecto al *arousal*, la más cercana es *El swing de Caperucita*. Si la primera prioridad se establece para el *valence*, podría ser una buena opción para este caso sugerir las canciones *pensarás en mí*, *Esperando por ti*, y *Doncella de virgo*, además de estar incluidas en el subconjunto de las canciones más cercanas por *arousal* (Tabla 6.2).

Tabla 6.1: Las 10 canciones no comerciales más cercanas a *Hysteria* según el DTW aplicado por verso y *valence*.

Canción	**Hysteria (DTW para *valence*)**
Hysteria	0
Pensarás en mí	0.61
Esperando por ti	0.69
Doncella de virgo	0.74
Bienvenidos	0.87
Intento de florecer	0.88
Currulao de nieves	0.95
Bajo mi piel acústico	1
Peste de silicio	1.10
Estático ser	1.10
El swing de Caperucita	1.10

Tabla 6.2: Las 10 canciones no comerciales más cercanas a *Hysteria* según el DTW aplicado por verso y *arousal*.

Canción	**Hysteria (DTW para *arousal*)**
Hysteria	0
El swing de Caperucita	1.0
A construir	1.20
Bienvenidos	1.30
Esperando por ti	1.40
Doncella de virgo	1.40
Estático ser	1.50
Currulao de nieves	1.50
Bajo mi piel	1.60
Eterno abrazo	1.60
Pensarás en mí	1.60

6.2.2 Resultados con estrategias de agrupamiento

Las estrategias de *clustering* pertenecen al campo de *machine learning* y se caracterizan especialmente por la realización de un entrenamiento no supervisado, en el que los datos no han sido etiquetados previamente. Los algoritmos de *clustering* tienen como objetivo separar los elementos en diferentes grupos que tienen ciertas características en común, y para ello se minimiza la distancia *intra-cluster* y se maximiza la distancia *inter-cluster*.

El algoritmo *K-means* es uno de los más utilizados y más sencillos de aplicar. El algoritmo se inicializa con K centroides generados aleatoriamente y calcula iterativamente la pertenencia de los patrones a cada grupo en función de la distancia al centroide. A continuación, el centroide se ajusta moviéndolo hacia el punto medio. En muchos casos, el algoritmo *K-means* se utiliza con conjuntos de datos que relacionan dos características para cada elemento del conjunto, lo que permite observar gráficamente la convergencia de estos algoritmos en un plano bidimensional. Este escenario de *K-means* genera un reto en su implementación tradicional sobre la percepción emocional a lo largo del tiempo en la música, en la que para cada elemento (canción) hay muchas mediciones. Sin embargo, existe una variante de este algoritmo desarrollada específicamente para implementaciones de series temporales y está disponible en la librería *TSLEARN*.

Con el fin de mostrar la contribución de la implementación de la librería a los MRS, a continuación se muestran los resultados del *clustering*, siguiendo el mismo experimento de la sección anterior. De nuevo el foco de análisis es la canción famosa *Hysteria*, el algoritmo *TimeSeriesKMeans* se ejecuta para luego identificar y analizar el cluster en el que se encuentra la canción tanto en *valence* como en *arousal*. Para el caso de la agrupación por *valence* (Fig. 6.3), la canción *Hysteria* se ubica en el cluster 1 junto con la canción *Doncella de virgo*, lo que concuerda con los resultados de la Tabla 6.1, en el que aparecen ambas canciones. Con respecto a la agrupación por *arousal* (Fig. 6.4), la canción *Hysteria* se ubica en el cluster 5 junto con las canciones *Currulao de nieves*, *El swing de Caperucita*, y *A construir*. Este resultado es coherente con los resultados de la Tabla 6.2 que también incluye estas canciones.

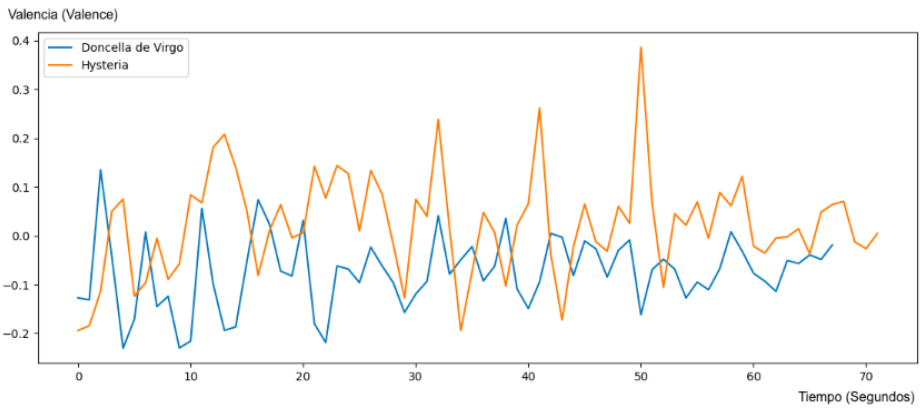


Figura 6.3: Cluster 1 para *valence*. Elaboración propia.

.

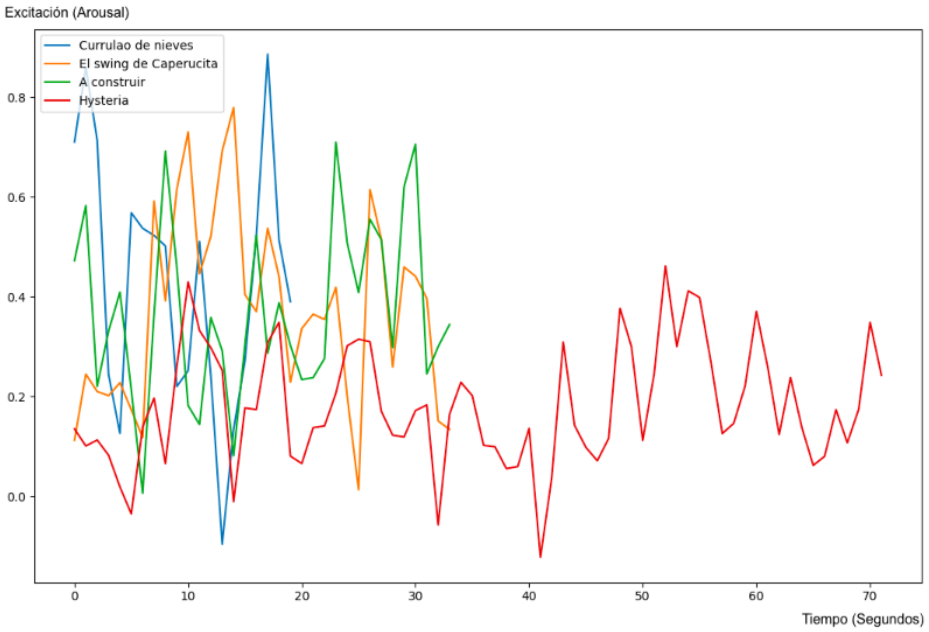


Figura 6.4: Cluster 5 para *arousal*. Elaboración propia.

El entrenamiento para los procesos de *clustering* por *valence* y *arousal* se parametrizó con $K=7$ y, como se puede ver en las Figuras 6.3 y 6.4, el nivel de agrupación por *valence* fue más ajustado (un menor número de canciones) en relación con el *arousal*. El valor de K se incrementó de uno en uno buscando un valor que se ajustara mejor al cluster en el que la canción *Hysteria* fuera ubicada, preferiblemente con el menor número posible de canciones (las más cercanas según similaridad). Es importante destacar que uno de los mayores retos de este experimento es encontrar el valor de K, lo que requiere muchas ejecuciones del mismo experimento, y una revisión manual de los resultados para determinar el valor más conveniente según los intereses del estudio. A pesar de que con *TSLEARN* es posible generar los diferentes *clusters* de series temporales, no es posible generar un gráfico bidimensional para visualizar la dispersión de las canciones de una manera más fácil.

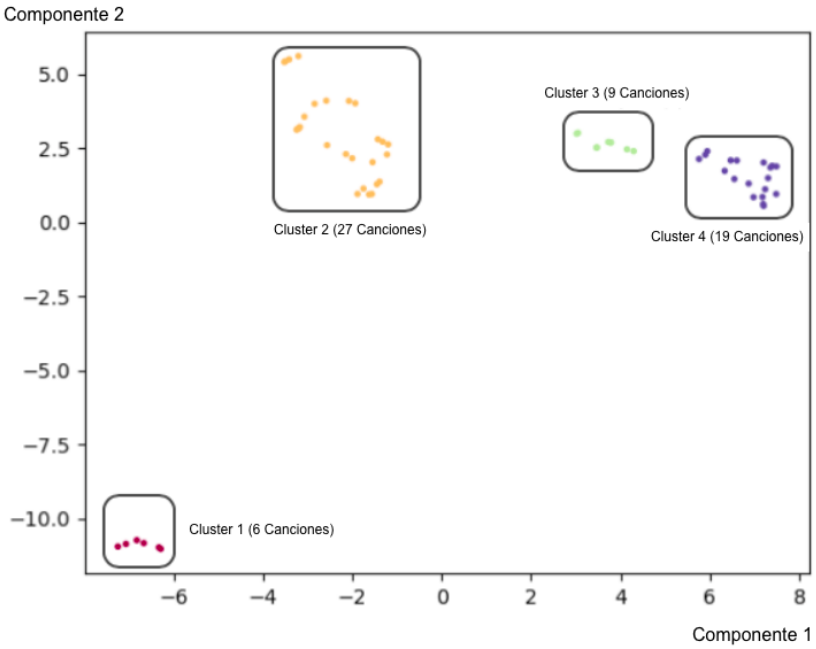


Figura 6.5: Agrupamiento por *valence* con UMAP y HDBSCAN. Elaboración propia.

Teniendo en cuenta las limitaciones anteriores de *TSLEARN*, se diseña un nuevo experimento de agrupación, pero esta vez utilizando dos nuevas librerías: *UMAP*[4] y *HDBSCAN*[5]. La primera permite realizar una reducción de la dimensión del conjunto de datos, y el segundo calcula el número óptimo de clusters. Las Figuras 6.5 y 6.6 muestran la agrupación de las canciones por *valence* y *arousal* respectivamente. Los valores que se pueden ver tanto en el eje x como en el eje y corresponden a los dos componentes resultantes tras aplicar la reducción de dimensión con la librería *UMAP*. Para un análisis más detallado, sería útil revisar las tablas de valores de similitud presentadas en la sección anterior para establecer una relación entre los valores de las métricas de similitud calculadas y las distancias *intra-cluster* e *inter-cluster*. La base de datos completa (60 canciones) fue considerada para generar las dos figuras, y la canción *Hysteria* también fue incluida. La agrupación se ha realizado por versos, y para ello se han analizado las canciones que no siguen una estructura que defina un verso, muestreando una parte representativa de la canción para incluirla en el experimento. Como se puede ver en las Figuras 6.5 y 6.6, la librería *HDBSCAN* determinó 4 clusters para ambos casos de agrupación.

Una vez más centraremos el análisis en el cluster en el que se ha localizado la canción *Hysteria* tanto en la agrupación por *valence* como por *arousal*. Para facilitar el análisis, una comparación entre los diferentes enfoques de los experimentos descritos anteriormente se presenta en la Tabla 6.3. En cuanto al caso de *valence* (Fig 6.5) la canción *Hysteria* se ha situado en el cluster 4 junto con 18 canciones de artistas noveles. De estas canciones nos interesa destacar las siguientes 7 canciones: *Esperando por ti*, *Bajo mi piel acústico*, *Doncella de virgo*, *Intento de florecer*, *Pensarás en mí*, *Currulao*

[4]https://umap-learn.readthedocs.io
[5]https://hdbscan.readthedocs.io

de nieves y *Bienvenidos*. Estas 7 canciones también están presentes en la Tabla 6.1, así como la canción *Doncella de virgo* forma parte del cluster 1 (Fig 6.3) calculado con *TSLEARN* para *valence*. En cuanto a los resultados de la agrupación por *arousal* (Fig 6.6), la canción *Hysteria* se ha situado en el cluster 1 junto con 10 canciones de artistas noveles. De estas canciones es interesante destacar las siguientes 5 canciones: *Esperando por ti*, *Doncella de virgo*, *Estático ser*, *Pensarás en mí* y *Currulao de nieves*. Estas 5 canciones también están presentes en la Tabla 6.2, así como también la canción *Currulao de nieves* es parte del cluster 5 (Fig 6.4) calculado con *TSLEARN* para *arousal*.

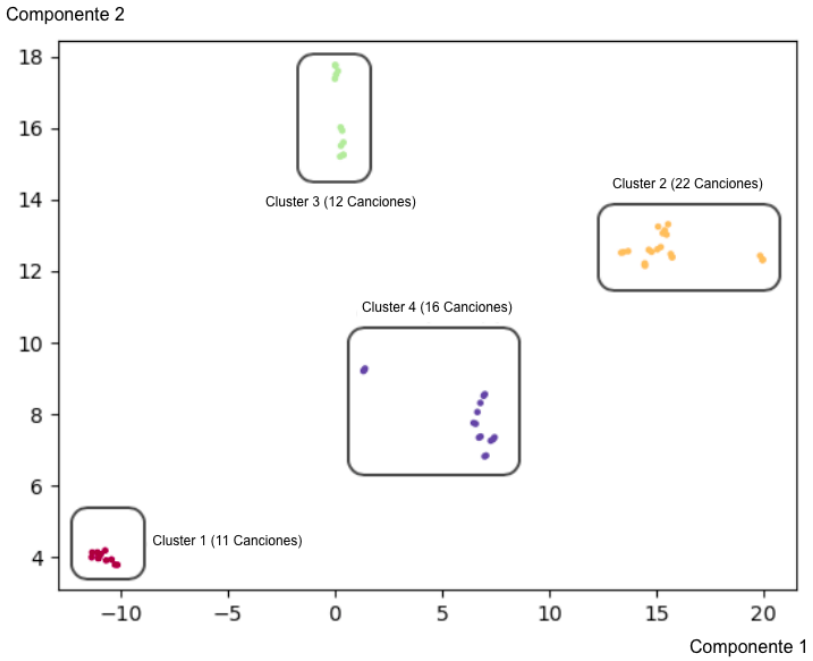


Figura 6.6: Agrupamiento por *arousal* con UMAP y HDBSCAN. Elaboración propia.

Este análisis muestra un importante grado de coherencia entre los diferentes resultados obtenidos a partir de las tablas de valores de similaridad, agrupamiento por *TSLEARN*, y agrupamiento con la utilización de *UMAP* y *HDBSCAN*.

Tabla 6.3: Comparación de las coincidencias de agrupación de canciones en los distintos experimentos

Algoritmo	***Valence***	***Arousal***
DTW	Hysteria (0.0) Pensarás en mí (0.61) Esperando por ti (0.69) Doncella de virgo (0.74) Bienvenidos (0.87) Intento de florecer (0.88) Currulao de nieves (0.95) Bajo mi piel acústico (1.0) Peste de silicio (1.10) Estático ser (1.10) El swing de Caperucita (1.10)	Hysteria (0.0) El swing de Caperucita (1.0) A construir (1.20) Bienvenidos (1.30) Esperando por ti (1.40) Doncella de virgo (1.40) Estático ser (1.50) Currulao de nieves (1.50) Bajo mi piel (1.60) Eterno abrazo (1.60) Pensarás en mí (1.60)
TimeSeriesKMeans con TSLEARN	Hysteria (Cluster 1) Doncella de virgo (Cluster 1)	Hysteria (Cluster 5) Currulao de nieves (Cluster 5) El swing de Caperucita (Cluster 5) A construir (Cluster 5)
UMAP	Hysteria (Cluster 4) Esperando por ti (Cluster 4) Bajo mi piel acústico (Cluster 4) Doncella de virgo (Cluster 4) Intento de florecer (Cluster 4) Pensarás en mí (Cluster 4) Currulao de nieves (Cluster 4) Bienvenidos (Cluster 4)	Hysteria (Cluster 1) Esperando por ti (Cluster 1) Doncella de virgo (Cluster 1) Estático ser (Cluster 1) Pensarás en mí (Cluster 1) Currulao de nieves (Cluster 1)

6.3 Conclusiones

En este capítulo se ha presentado el diseño y desarrollo de un prototipo de sistema recomendador que implementa una estrategia de recomendación híbrida, incluyendo filtrado basado en contenido, filtrado basado en emociones, y filtrado basado en similaridad. Estas estrategias han sido seleccionadas a partir de una revisión de literatura que las resalta como las más utilizadas en la actualidad, y que además, destacan por relacionar las características de la música con las emociones.

La percepción emocional, tanto en *valence* como *arousal*, ha sido representada a través de series temporales, considerando que un proceso de apreciación musical es una actividad que transcurre a lo largo del tiempo, y asimismo, a lo largo de la estructura musical de una canción. Este enfoque permitió calcular grados de similaridad entre los versos de un conjunto de canciones (ENSA *dataset*) utilizando el algoritmo DTW. Este algoritmo permite comparar series temporales de diferentes longitudes, lo que representa una importante ventaja frente al algoritmo de Euclides, y que resulta necesario para comparar versos de diferentes canciones debido a que en general tienen diferentes duraciones.

Adicionalmente, se implementaron diferentes estrategias de *clustering*, mostrando una alta consistencia entre los resultados obtenidos por cada una de ellas. Finalmente,

es importante resaltar la orientación del experimento a la generación de recomendaciones de canciones de artistas noveles a partir de una canción famosa, como es el caso particular de *Hysteria* de la agrupación *Def Leppard*. Este experimento resalta nuevamente la importancia de generar recomendaciones basadas en contenido y no en popularidad, lo que contribuye a la mitigación de los sesgos preexistentes que normalmente se encuentran en la industria musical.

Capítulo 7

Conclusiones y trabajos futuros

El aporte general de esta Tesis ha sido el diseño de un prototipo de un sistema recomendador de piezas musicales, a partir de la relación entre las características intrínsecas de la música y las emociones percibidas por el oyente. Este aporte general se ha logrado a través de la integración de diferentes resultados que fueron obtenidos de manera progresiva durante el desarrollo de esta Tesis. Si bien en los apartados de cierre de los capítulos anteriores se adelantaron algunas conclusiones particulares de cada tema, en este capítulo final se exponen las conclusiones generales de la investigación. Inicialmente, en el Apartado 7.1 se presentan las conclusiones generales analizadas desde el cumplimiento de objetivos específicos (Sección 7.1.1), como también desde el punto de vista de las preguntas de investigación (Sección 7.1.2). El Apartado 7.2 describe las implicaciones prácticas. El Apartado 7.3 presenta la producción científica. El Apartado 7.4 presenta las limitaciones del estudio. Finalmente, en el Apartado 7.5 se sugieren algunas líneas de investigación para explorar en el futuro.

7.1 Conclusiones

Esta Tesis ha tenido como objetivo general la generación de un sistema recomendador de piezas musicales a partir de un modelo de elicitación de emociones vinculado a las características de la música. Para tal fin, se abordaron una serie de objetivos específicos que se indicaron en el apartado 1.3, así como también 5 preguntas de investigación definidas en el apartado 1.4. A continuación, se exponen las conclusiones vinculadas al desarrollo de la Tesis, justificando los objetivos (Sección 7.1.1), y respondiendo las preguntas de investigación (Sección 7.1.2).

7.1.1 Conclusiones en relación a los objetivos

A continuación se retoman y discuten los aportes generados para el cumplimiento de los objetivos de esta Tesis:

- **Objetivo 1:** Determinar el estado actual de la computación afectiva en cuanto a la medición y reconocimiento de emociones a partir de la estimulación musical.

- Los apartados 2.1 y 2.2 aportan la fundamentación teórica que permite abordar los conceptos más relevantes relacionados con la medida y caracterización de las emociones en la música.

- El apartado 2.3 presenta la fundamentación teórica de los modelos computacionales disponibles para la representación de emociones, resaltando en particular, los modelos categóricos y dimensionales. Adicionalmente, se introducen los procesos de etiquetado en la música de manera estática y dinámica.

- El apartado 3.1 presenta una revisión de librerías de alto nivel, en donde se analizan las funcionalidades relacionadas con la extracción de características musicales y el reconocimiento de emociones.

- El apartado 3.3 presenta una descripción del *dataset* de MediaEval. También se incluye un análisis sobre algunos casos puntuales de etiquetado dinámico en un modelo afectivo dimensional (anexo D. Análisis del sistema de etiquetado en MediaEval), que permite comprender con mayor detalle el etiquetado de emociones en la música.

- **Objetivo 2:** Estudiar las características intrínsecas de la música.

  - El apartado 2.2 expone algunas de las características más importantes de la música, como es el caso de la nota musical, la armonía, el acorde, el modo (menor, mayor) y la modulación. Adicionalmente, se introduce la relación entre estas características y la percepción emocional del oyente.
  - El apartado 5.1 presenta una descripción general de los procesos de composición musical de los artistas invitados al diseño del *dataset* ENSA.

- **Objetivo 3:** Determinar cuáles son las estrategias más efectivas para medir y reconocer emociones en la música, realizando experimentos para establecer la adecuación necesaria de estas estrategias.

  - Los apartados 3.4, 3.5, 3.6 presentan la revisión del estado del arte con respecto a los sistemas que realizan predicción y clasificación de canciones en modelos afectivos categóricos, dimensionales, y en algunos pocos casos, categóricos difusos.
  - Los experimentos se realizan a través de los prototipos implementados en el capítulo 4, el diseño del *dataset* ENSA del capítulo 5, y el desarrollo del sistema recomendador del capítulo 6.

- **Objetivo 4:** Construir un modelo que permita establecer una relación entre las características intrínsecas de la música y las emociones percibidas por el oyente.

- En el apartado 3.3 se describe el *dataset* de MediaEval, el cual incluye mediciones sobre características de bajo de nivel de la música, y las relaciona con la percepción emocional del oyente a través de un modelo dimensional afectivo. Esta información que implica una relación entre características de sonido y percepción emocional, es utilizada para posteriormente diseñar los modelos de *machine learning* que hacen parte de la implementación de los prototipos de los apartados 4.1 y 4.3. Los modelos obtenidos permiten para canciones nuevas, y a partir de sus características musicales, determinar que tipo de emociones pueden elicitar en el oyente.

- **Objetivo 5:** Implementar un prototipo para la clasificación emocional de la música en base a sus características intrínsecas.
  - En el apartado 4.1 se implementa un prototipo de predicción de valores de *valence* y *arousal*.
  - En el apartado 4.2 se implementa un prototipo de clasificación no determinístico (*fuzzy*) que define grados de pertenencia a tres categorías de *arousal* (baja, media, alta).
  - En el apartado 4.3 se implementa un prototipo de clasificación determinístico que clasifica en 4 cuadrantes.

- **Objetivo 6:** Estudiar los diferentes tipos de sesgos existentes en las estrategias de recomendación, y proponer algunas medidas para su tratamiento.
  - En el apartado 3.8 se presenta la revisión del estado del arte relacionado con los sesgos, como también una serie de medidas para la mitigación de sus efectos.
  - En el Capítulo 5 se diseña un nuevo *dataset* basado en artistas noveles para contrarrestar el efecto de la popularidad (sesgo preexistente).
  - En el Capítulo 6 se implementa un sistema recomendador que considera una estrategia de recomendación basado en el filtrado de emociones y de similaridad. El sistema incluye un experimento en donde a partir de una pieza musical comercialmente famosa a nivel mundial, se determina los niveles de similaridad a piezas musicales de artistas noveles.

- **Objetivo 7:** Implementar un experimento para el etiquetado de emociones durante un proceso de apreciación musical.
  - El experimento de etiquetado de emociones a través de un proceso de apreciación emocional se presenta en el apartado 5.1. Este experimento hace parte de la generación del *dataset* ENSA que se detalla en el capítulo 5.

- **Objetivo 8:** Diseñar e implementar un nuevo *dataset* de piezas musicales de artistas noveles.
  - El diseño del nuevo *dataset* se presenta en el capítulo 5 Diseño del *dataset* musical: *Emotional Non-Superstar Artist-Dataset (ENSA)*.

- **Objetivo 9:** Desarrollar un prototipo de sistema recomendador basado en la relación entre las propiedades intrínsecas de la música y las emociones percibidas por el oyente.

  · En el capítulo 6 se presenta el desarrollo del sistema recomendador musical, el cual tiene en cuenta la estructura de la canción, en particular, la percepción emocional sobre los versos.

7.1.2 Conclusiones que responden a las preguntas

A continuación se retoman y responden las preguntas de investigación de esta Tesis:

- **P1: ¿Cómo se puede representar computacionalmente la relación entre la música y las emociones?**
  La relación entre la música y las emociones se puede representar a partir de reglas lógicas de inferencia, en donde para ciertos valores de las características musicales se asocian determinadas emociones, en caso de utilizar un modelo categórico, o una coordenada de *valence* y *arousal*, en caso de utilizar un modelo dimensional. Es importante resaltar que la identificación de todas las reglas lógicas es un trabajo realmente complejo, teniendo en cuenta que no necesariamente existen unas reglas universales, pues éstas podrían variar de un usuario a otro, o entre grupos de usuarios con perfiles demográficos diferentes. Por esta razón, la mejor manera de representar computacionalmente la relación entre la música y las emociones, es a través de modelos de *machine learning*, en donde las reglas se descubren a través de un proceso de aprendizaje. Es así como esta Tesis aporta con el diseño de dos sistemas basados en *machine learning*, el primero para diseñar un sistema de predicción de niveles de *valence* y *arousal* en el apartado 4.1, y el segundo para diseñar un sistema de clasificación emocional en el apartado 4.3.

- **P2: ¿Puede la música mejorar la experiencia de un usuario o provocar emociones específicas?**
  De acuerdo al apartado 2.1 (emociones en la música), la música tiene la capacidad de afectar emocionalmente a una persona, generando una transformación en su estado de ánimo. La relación entre las emociones percibidas por el oyente y la música, se da a través de las diferentes características de la música, como lo son el tono, el modo y el tempo. En general, los artistas experimentan con diferentes valores sobre las características musicales, así como también con la estructura de la canción, para con ello, generar una experiencia que en muchos de los casos, tanto artistas como oyentes, la consideran una experiencia emocional.

- **P3: ¿Cuáles son las técnicas de computación afectiva más apropiadas para determinar qué es lo que realmente siente un oyente?**
  Partiendo de la fundamentación teórica del capítulo 2, y de la revisión del estado del arte del capítulo 3, se determina que en general las técnicas más utilizadas son las relacionadas con procesos de etiquetado; por una parte, estos procesos de etiquetado pueden ser llevados a cabo a través de un modelo afectivo dimensional,

en donde se tiene una coordenada de *valence* y *arousal*. Por otra parte, también es posible la utilización de modelos categóricos, los cuales normalmente comprenden una serie de emociones elegibles por el oyente. Generalmente, los procesos de etiquetado suelen ser estáticos, lo que implica un etiquetado de percepción general sobre la canción. Además, resulta interesante considerar un etiquetado dinámico, debido a su capacidad de describir la percepción emocional a lo largo del tiempo, tal como ocurre con el *dataset* de MediaEval revisado en el apartado 3.3.

- **P4: ¿Qué *frameworks* existen actualmente para el reconocimiento de emociones en la música?**
  Durante la revisión del estado del arte, específicamente en el apartado 3.1, se identificaron algunas librerías de alto nivel para el reconocimiento de emociones en la música, que pueden considerarse como un punto de partida en la exploración de *frameworks* disponibles para utilizar en el campo de MER. Esta revisión permitió estudiar y comparar las diferentes funcionalidades ofrecidas por *Spotify API, jAudio, AcousticBrainz* y *OpenSMILE*.

- **P5: ¿Puede diseñarse computacionalmente un recomendador de piezas musicales basado en la percepción emocional de los usuarios?**
  De acuerdo al apartado 2.9 y apartado 3.7 que soportan respectivamente la fundamentación teórica y el estado del arte de los MRS, el diseño de este tipo de sistemas es posible, y generalmente tiene en cuenta la implementación de diversas estrategias para realizar las recomendaciones. Entre las estrategias más utilizadas se destaca el filtrado basado en emociones, el filtrado basado en contenido, y el filtrado basado en el contexto del usuario. Y aunque existe un notable desarrollo en el campo de MRS, todavía hay muchas oportunidades para realizar aportes, en especial si parte de la investigación se concentra en el tratamiento de los sesgos, la efectividad de las recomendaciones, el diseño de estrategias centradas en el usuario, la evaluación y comparación entre diferentes sistemas de MRS, la generación de nuevos *datasets*, y el diseño de nuevos y mejores descriptores tanto musicales como emocionales. A partir de estas oportunidades, esta Tesis contribuye con dos resultados: el diseño un nuevo *dataset* a través del capítulo 5, y el diseño de un sistema recomendador en el capítulo 6.

7.2 Implicaciones prácticas

A continuación se listan las implicaciones prácticas de los diferentes productos obtenidos en esta Tesis:

- Los archivos de ambos modelos de reconocimiento de emociones en las dimensiones de *valence* y *arousal* se encuentran disponibles en la cuenta de GitHub del tesista[1]. Estos modelos pueden ser utilizados por la comunidad científica para dar continuidad a diferentes experimentos en el campo de MER a través del principio

[1]Disponible en: https://github.com/yesidospitiamedina/ENSA/tree/main/IA_models

de *transfer learning*, el cual permite transferir el conocimiento de un modelo de *machine learning* a otro [149].

- El formulario WEB diseñado y utilizado para el proceso de etiquetado emocional de las canciones del *ENSA dataset*, se encuentra disponible en un servidor con dirección IP pública que fue destinado para los despliegues de diferentes elementos experimentales de la Tesis[2]. Este formulario permite extender a futuro el proceso de etiquetado sobre el *dataset* con nuevos oyentes. Adicionalmente, también puede servir de referencia para extender las características de etiquetado que son consideradas actualmente dentro del alcance de la Tesis.

- La totalidad de las canciones del *ENSA dataset* en formato MP3, como también las evaluaciones emocionales realizadas por oyentes y artistas, se encuentran disponibles en la cuenta de GitHub del tesista[3]. El *ENSA dataset* abre nuevas posibilidades de experimentación para la comunidad científica interesada especialmente en la recuperación de información musical, reconocimiento de emociones, estrategias de recomendación, y el estudio de sesgos.

- Las estrategia de recomendación y sus respectivos algoritmos desarrollados en el capítulo 6 resultan útiles para fines académicos, como lo es la participación en congresos, material de clase para enseñar y divulgar conocimiento, entre otros más. También puede ser utilizada como punto de partida para el desarrollo un nuevo prototipo de sistema de reproducción musical que consideré un *frontend* para el usuario final.

7.3 Producción científica

Las principales contribuciones relacionadas con los resultados de esta Tesis son:

- La revisión y discusión del estado actual de las diferentes áreas involucradas en el reconocimiento de emociones en la música.

- La propuesta de algunas recomendaciones para el tratamiento de sesgos en sistemas recomendadores.

- El diseño de tres prototipos para el reconocimiento de emociones en la música.

- El diseño de un *dataset* musical.

- El diseño de un prototipo de sistema recomendador basado en emociones y en la estructura musical de las canciones.

La producción científica que respalda estas contribuciones es la siguiente:

[2]Disponible en: http://104.237.5.250/evaluacionensa/form.php
[3]Disponible en: https://github.com/yesidospitiamedina/ENSA/

- Ospitia-Medina Y., Baldassarri S., Beltrán J.R. (2019) High-Level Libraries for Emotion Recognition in Music: A Review. In: Agredo-Delgado V., Ruiz P. (eds) Human-Computer Interaction. HCI-COLLAB 2018. Communications in Computer and Information Science, vol 847. Springer, Cham. [150]

  Capítulos relacionados:

  · Apartado 3.1 Librerías de alto nivel para la extracción de características musicales

- Yesid Ospitia Medina, José Ramón Beltrán, Cecilia Sanz, and Sandra Baldassarri. 2019. Dimensional Emotion Prediction through Low-Level Musical Features. In Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound on ZZZ (AM'19). ACM, New York, NY, USA, 231-234. DOI: https://doi.org/10.1145/3356590.3356626. [131]

  Capítulos relacionados:

  · Apartado 3.3 MediaEval Dataset

  · Apartado 3.4 Sistemas de predicción

  · Apartado 4.1 Sistema de predicción de emociones

- Ospitia Medina Y., Baldassarri S., Beltrán J.R. (2020) Librerías de alto nivel para el reconocimiento de emociones en la música. En: Ingeniería colaborativa desde la interacción humano-computador. [151]

  Capítulos relacionados:

  · Apartado 3.1 Librerías de alto nivel para la extracción de características musicales

- Y. Ospitia-Medina, S. Baldassarri, C. Sanz, J. R. Beltrán and J. A. Olivas. Fuzzy Approach for Emotion Recognition in Music. IEEE Congreso Bienal de Argentina (ARGENCON), 2020, pp. 1-7. DOI: 10.1109/ARGENCON49523.2020.9505382. [129]

  Capítulos relacionados:

  · Apartado 3.5 Sistemas no determinísticos (*Fuzzy)*

  · Apartado 4.2 Sistema de clasificación emocional no determinística

- Medina, Y.O., Beltrán, J.R. & Baldassarri, S. Emotional classification of music using neural networks with the MediaEval dataset. Pers Ubiquit Comput (2020). https://doi.org/10.1007/s00779-020-01393-4. [83]

  Capítulos relacionados:

- Apartado 3.3 MediaEval Dataset
- Apartado 3.6 Sistemas de clasificación determinísticos
- Apartado 4.3 Sistema de clasificación emocional determinística

- Ospitia-Medina, Y., Baldassarri, S., Sanz, C., Beltrán, J.R. (2023). Music Recommender Systems: A Review Centered on Biases. In: Biswas, A., Wennekes, E., Wieczorkowska, A., Laskar, R.H. (eds) Advances in Speech and Music Technology. Signals and Communication Technology. Springer, Cham. DOI: https://doi.org/10.1007/978-3-031-18444-4. [152]

  Capítulos relacionados:

  - Apartado 3.2 Datasets musicales
  - Apartado 3.3 MediaEval Dataset
  - Apartado 3.7 Sistemas recomendadores musicales
  - Apartado 3.8 Tratamiento de sesgos en MRS

- Ospitia-Medina, Y., Beltrán, J.R. & Baldassarri, S. ENSA dataset: a dataset of songs by non-superstar artists tested with an emotional analysis based on time-series. Pers Ubiquit Comput (2023). https://doi.org/10.1007/s00779-023-01721-4. [153]

  Capítulos relacionados:

  - Capítulo 5 Diseño del *dataset* musical: *Emotional Non-Superstar Artist-Dataset (ENSA)*
  - Capítulo 6 Desarrollo de un sistema de recomendación musical

7.4 Limitaciones del estudio

A continuación se describen las limitaciones identificadas a través de los diferentes estudios y experimentos desarrollados en esta Tesis:

- En esta Tesis, se abordan los conceptos más importantes de la teoría musical, pero solo de forma superficial, sin entrar en un alto grado de profundización, dado que el foco del trabajo se centra en el plano técnico requerido en el Doctorado en Ciencias Informáticas.
- El *ENSA dataset* presenta algunas limitaciones relacionadas con el tamaño del *dataset* (integrado por 60 canciones), la variedad y representación de géneros musicales, la distribución de canciones entre voces masculinas y femeninas, la distribución de canciones entre instrumentales y no instrumentales, y la distribución de canciones entre las que siguen una estructura musical y las que no siguen una estructura musical.

- Los procesos de etiquetado por parte de los oyentes en el *ENSA dataset*, sólo consideraron un modelo de etiqueta única, lo que simplifica la estrategia de recomendación, pero al mismo tiempo limita la posibilidad de que los oyentes consideren etiquetas alternativas.

- Algunas canciones en el *ENSA dataset* se etiquetaron emocionalmente como una unidad completa porque no seguían una estructura musical que se pudiera describir por versos y coros. Lo que no permite garantizar por completo una estrategia de agrupación basada en el mismo tipo de estructura musical.

- El análisis preliminar del *ENSA dataset* sugiere la existencia de algunas relaciones entre diferentes características de los datos, pero estas relaciones aún no están muy claras y sería interesante estudiarlas en profundidad, ya que podrían ser útiles para mejorar la estrategia de recomendación.

- Todos los experimentos de agrupación se basaron en una única métrica de similaridad, DTW. Sería interesante repetir los experimentos con la versión *soft* de DTW, como también explorar otras métricas aplicables, y así evaluar los resultados entre diferentes experimentos de *clustering* con diferentes métricas de similaridad.

- Aunque en líneas generales este trabajo estudia la inclusión de sesgos, y se ha trabajado especialmente en evitar el sesgo por popularidad, es posible que el *ENSA dataset* incluya un sesgo cultural, teniendo en cuenta que todos los artistas son procedentes de un mismo país.

7.5 Futuras líneas de trabajo

A continuación se presentan algunas líneas de investigación para profundizar a futuro, las cuales han sido identificadas a lo largo del recorrido de esta Tesis.

- Aunque existen en la actualidad una variedad de librerías para la extracción de características de sonido, esta línea de investigación sigue siendo de amplio interés para la comunidad de MIR, en donde parte de su trabajo consiste en estudiar la extracción de diversas características musicales. Este proceso de extracción y representación de estas características es un factor clave en el campo de MER y MRS, debido a que permite diseñar nuevos y mejores descriptores musicales.

- Las estrategias de etiquetado emocional también presentan muchas oportunidades a futuro. Resulta interesante analizar la posibilidad de incluir nuevas etiquetas con información que se considere relevante para los campos de MER y de MRS. También es importante trabajar en nuevas propuestas de etiquetado que permitan mejorar la calidad de los datos recolectados.

- Desde la computación afectiva existen algunos avances interesantes en la captura, procesamiento, e interpretación de señales fisiológicas que permiten inferir

emociones relacionadas con experiencia de usuario, en el caso particular de los intereses de esta Tesis, una experiencia de apreciación musical. La inclusión de este tipo de información a futuro permitiría llevar los sistemas recomendadores musicales a otro nivel.

- La inclusión del artista, para comprender sus necesidades y así diseñar sistemas recomendadores que consideren sus intereses, seguirá marcando a futuro un factor relevante en el éxito del campo de MRS. En la actualidad muchos de los trabajos se concentran en el oyente, olvidando que el artista es otro usuario importante de los MRS, con sus propio intereses y necesidades.

- El sesgo de la popularidad tiene un impacto sobre la percepción de la eficiencia de los MRS; artistas y oyentes expresan su inconformidad. Como trabajo futuro se propone profundizar en el estudio de los sesgos, y proponer nuevas estrategias de recomendación que incluyan su mitigación.

- Aunque existen algunos *datasets* disponibles para experimentar en el campo de MRS, muchos de éstos no han sido diseñados específicamente para realizar experimentos dentro del campo de MRS, en lugar de ello, se han diseñado para cubrir aspectos más generales del campo de MER y MIR. Por lo tanto, se propone a futuro, hacer nuevos experimentos con el *dataset* ENSA, resultado de esta Tesis, de tal manera que se logre generar más datos de etiquetado, y a su vez, incluir nuevas canciones.

- La estrategia de recomendación diseñada en esta Tesis se ha centrado en los versos de las canciones, por lo que se propone a futuro, la inclusión de otras partes de la estructura musical, como lo son la introducción, el coro, y los solos. Considerando también la posibilidad de que el reconocimiento de la estructura musical sea completamente automático.

Se espera que esta Tesis permita a otros investigadores profundizar y generar nuevos aportes en la línea de los sistemas recomendadores musicales y otras áreas relacionadas.

A. Fundamentos de teoría musical

A continuación en la Tabla 7.1 se presenta una selección de conceptos fundamentales de teoría musical que son relevantes para los objetivos de esta Tesis. Los conceptos han sido seleccionados de los trabajos de J. Powell [9] y A. Candelaria [154], teniendo en cuenta aquellos que tienen mayor presencia y estudio en las áreas de recuperación de información musical, reconocimiento de emociones y sistemas recomendadores. Este anexo no pretende explicar en profundidad los diferentes aspectos de la teoría musical, pues se reconoce su alto nivel de complejidad, el cual va mas allá del alcance y de los intereses de esta Tesis.

Tabla 7.1: Conceptos fundamentales sobre teoría musical

Concepto	Definición
Nota musical	Sonido que se caracteriza por su frecuencia de vibración o altura musical y por su intensidad sonora.
Escala	Serie secuencial de siete notas organizadas como si fueran escalones que permiten subir o bajar de unos a otros. La escala temperada occidental consiste en 12 notas separadas cada una de ellas por una distancia que se denomina semitono. Dos semitonos constituyen un tono.
Tonalidad	Nombre que recibe cada una de las notas de cualquier instrumento musical: C, D, E, F, G, A, B. Para algunas de estas notas existe la posibilidad de notas sostenidas o bemoles, que corresponden a una modificación de un semitono hacía arriba o hacía abajo respectivamente.
Melodía	Sucesión de notas con distintas tonalidades.
Acorde	Sonido generado por tres o más notas tocadas al mismo tiempo.
Armonía	Sucesión de acordes.
Contrapunto	Considerado uno de los tipos más complejos de armonía. Consiste en acompañar una melodía con otra melodía.
Modo	Reglas compositivas usadas en los sistemas musicales. Entre las más conocidas se encuentra el modo mayor y el modo menor. Los modos se construyen a partir de escalas de 7 notas formadas por sucesiones de tonos y semitonos.
Modulación	Alternar de un modo a otro. Típicamente de mayor a modo menor, y viceversa; pero no es el único caso.
Ritmo	Describe la distribución temporal de los sonidos, y el énfasis sobre estos mismos.
Tempo	Asociado a la definición del término BPM (*beats per minute* / pulsos por minuto). Se suele entender a través de la cantidad de veces que se dan golpes con el pie al seguir una pieza musical.
Timbre	Las características del sonido que permiten distinguir cada instrumento musical.

B. Experimento con *AcousticBrainz*

Como un ejemplo práctico de *AcousticBrainz* se muestra a continuación los resultados obtenidos después de utilizar la librería para extraer características sobre la canción *Rocking in the free world* de *Bon Jovi*. En la respuesta JSON de la Figura 7.1 se identifica una tonalidad de Re sostenido (D#) en modo menor (*minor*), mientras que en la respuesta JSON de la Figura 7.2 se reconoce un tempo aproximado de 142 bpm. En cuanto a la clasificación emocional en la respuesta JSON de la Figura 7.3 se muestran las emociones predominantes.


```
"chords_key": "D#",
"chords_number_rate": 0.00177329150029,
"chords_scale": "minor",
"chords_strength": {
  "dmean": 0.0106518547982,
  "dmean2": 0.0108457338065,
  "dvar": 7.51757543185e-05,
  "dvar2": 8.68747229106e-05,
  "max": 0.810177087784,
  "mean": 0.522222876549,
  "median": 0.52260440588,
  "min": 0.254474312067,
  "var": 0.0089641045779
```


Figura 7.1: Extracción de características de bajo nivel (tonalidad y modo). Captura del código de respuesta JSON.


```
"bpm": 142.085388184,
"bpm_histogram_first_peak_bpm": {
  "dmean": 0,
  "dmean2": 0,
  "dvar": 0,
  "dvar2": 0,
  "max": 144,
  "mean": 144,
  "median": 144,
  "min": 144,
  "var": 0
},
```


Figura 7.2: Extracción de características de bajo nivel (bpm). Captura del código de respuesta JSON.

```
"mood_aggressive": {
      "all": {
        "aggressive": 0.98418200016,
        "not_aggressive": 0.0158179998398
      },
"mood_happy": {
      "all": {
        "happy": 0.26508384943,
        "not_happy": 0.73491615057
      },

"mood_relaxed": {
      "all": {
        "not_relaxed": 0.834616065025,
        "relaxed": 0.165383920074
      },

"mood_sad": {
      "all": {
        "not_sad": 0.819041311741,
        "sad": 0.180958673358
      },
"moods_mirex": {
      "all": {
        "Cluster1": 0.0857282057405,
        "Cluster2": 0.0417659841478,
        "Cluster3": 0.18078148365,
        "Cluster4": 0.0692446380854,
        "Cluster5": 0.6224796772
      },
```


Figura 7.3: Extracción de características de alto nivel (emociones). Captura del código de respuesta JSON.

El modelo aplicado para la clasificación emocional por parte de *AcousticBrainz*, reconoce en la canción lo siguiente:

- En un 98 % de clasificación la canción transmite una emoción de agresividad.
- En un 73 % de clasificación no transmite una emoción de felicidad.
- En un 83 % de clasificación no transmite relajación. Lo que es bastante coherente con el valor del atributo *mood_aggresive*.
- En un 82 % no transmite una emoción de tristeza.
- De acuerdo al atributo *moods_mirex*, la canción predomina en el *cluster* 5, que según el modelo definido en [76], en este *cluster* se contempla las emociones del tipo: agresivo, ardiente, tenso, ansioso, intenso, volátil, visceral.

C. Cuadro comparativo de librerías de alto nivel

La Tabla 7.2 presenta una comparación entre las librerías *Spotify API*, *jAudio*, *Acoustic-Brainz* y *OpenSMILE*. Para esta comparación, además de las características generales presentadas en el apartado 3.1, se han incluido los atributos más representativos y comunes de las librerías, así como también algunas características adicionales que se consideran pertinentes en los sistemas de MER y cuya importancia se resalta en [82].

Tabla 7.2: Cuadro comparativo de las librerías de alto nivel.

Atributo	**Spotify**	**jAudio**	**AcousticBrainz**	**OpenSMILE**
Características reconocidas	Alto nivel (18).	Bajo nivel (138).	Bajo (80), y alto nivel (20).	–
Licenciamiento	Comercial	Libre	Libre	Libre
Formato de salida	Archivo JSON	Archivo XML	Archivo JSON	Archivos CSV, HTK, ARFF, LIBSVM
Tipo de Arquitectura	Servicio en la Nube	Local	Servicio en la Nube	Local
Emociones reconocidas	Clasificación de emociones con *valence* positivo y *valence* negativo.	No	Directamente identifica las emociones de: Feliz, relajado, triste, agresivo, divertido/fiesta. Adicionalmente incluye 5 *clusters*, que asocian la pieza musical a una categoría emocional.	No
Reconoce género musical	No	No	Sí	No
Reconoce canciones en vivo	Sí	No	No	No
Reconoce canciones instrumentales	Sí	No	Sí	No
Identifica la voz por género	No	No	Sí	No
Identifica la velocidad Bpm	Sí	No	Sí	No
Identifica la tonalidad (*Chord recognition*)	Sí	No	Sí	No
Identifica el modo presente en la tonalidad de la canción	Sí	No	Sí	No

Continúa en la siguiente página

Tabla 7.2 – Continuación desde la página anterior

Atributo	Spotify	jAudio	AcousticBrainz	OpenSMILE
Permite implementar modelos propios para características de alto nivel dentro de la misma librería	No	Sí	No	No
Permite extender su uso y parametrizar la librería.	No	Sí	No	Sí
Facilidad de uso (Normal, Intermedio, Avanzado)	Intermedio	Avanzado	Intermedio	Avanzado
Clasificación *Fuzzy* en la detección de emociones en la música, por clústeres y emociones predominantes	No	No	Sí	No
Detección de la variación de la emoción en la música a lo largo del tiempo	No	No	No	No
Modelo afectivo de clasificación	Dimensional (1D)	-	Dimensional (2D)	-
***Personalized* MER (PGER): anotación personalizada por el usuario para el entrenamiento del reconocimiento de emociones.**	No	No	No	No
General MER: Utilización de algoritmo de regresión general	Sí	Sí	Sí	No
***Groupwise* MER: Utilización de algoritmo de regresión por grupos de interés.**	No	No	No	No
***Residual Modeling:* Capacidad para calcular la distancia entre PGER y General MER**	No	No	No	No

Continúa en la siguiente página

Tabla 7.2 – Continuación desde la página anterior

Atributo	Spotify	jAudio	AcousticBrainz	OpenSMILE
Lyrics feature extraction: **extracción de la letra contenida en la canción**	No	No	No	No
Consideración de los factores situacionales del oyente, frente a la percepción emocional. Señales psicológicas o fisiológicas, situación de contexto	No	No	No	No

D. Análisis del sistema de etiquetado en MediaEval

Aunque los usuarios tenían que anotar sus emociones, en muchos casos, parece que indicaron su estado de ánimo, pues se detectaron muchas diferencias entre anotaciones de diferentes anotadores para una misma canción (ver Figuras 7.4 y 7.5). Es importante destacar que cada canción es anotada por múltiples usuarios, por lo que se dispone de la anotación dinámica de cada usuario, de tal manera es posible calcular la anotación promedio de todos los usuarios para cada ventana de tiempo de duración de una determinada canción. Además, para cada canción también es posible calcular un valor V/A promediado. El proceso de cálculo de este V/A promediado implica, en primer lugar, promediar las anotaciones de todos los anotadores para cada ventana de tiempo y, a continuación, calcular la media de todos los valores V/A resultantes para todas las longitudes de ventana de tiempo. De este modo, cada anotación final puede relacionarse con el valor medio de cada característica de sonido, que se calcula a partir de todas las ventanas temporales. La generación de descriptores con valores medios para el contexto de las emociones, como también de las características de sonido, es algo que se suele encontrar con frecuencia en la literatura, sin embargo, es una situación que vale la pena analizar en detalle desde el punto de vista artístico de la música para discutir si es realmente representativo y válido.

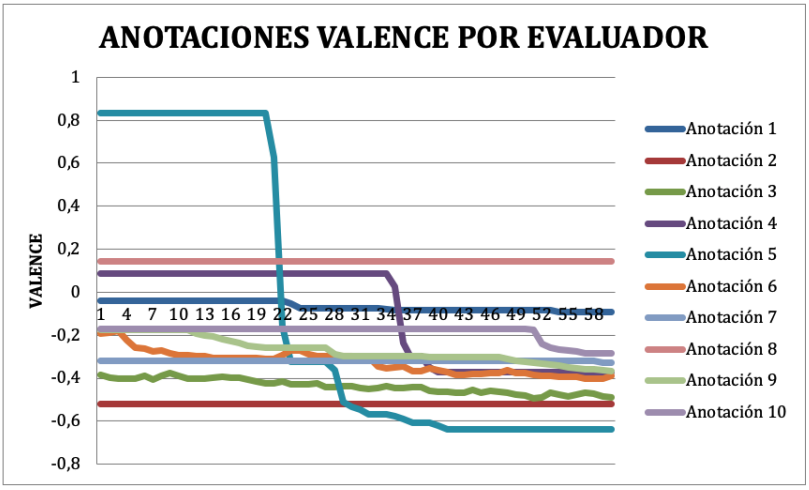


Figura 7.4: Ejemplo de anotaciones de valencia (*valence*): 10 anotadores etiquetando sobre el tiempo el *valence* de la canción 2.mp3. Elaboración propia.

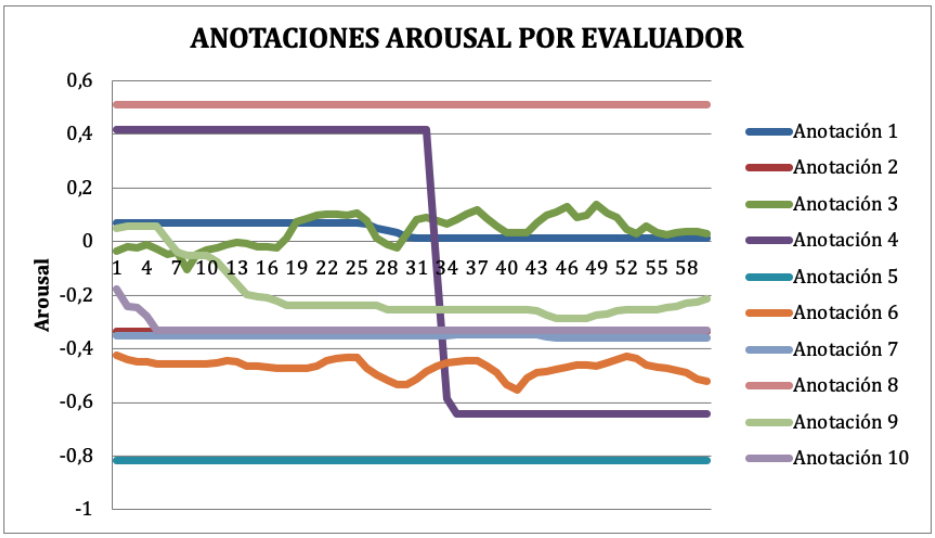


Figura 7.5: Ejemplo de anotaciones de excitación (*arousal*): 10 anotadores etiquetando sobre el tiempo el *arousal* de la canción 2.mp3. Elaboración propia.

Bibliografía

[1] IFPI. Global Music Report 2023. Technical report, IFPI, London, 2023. x, 1, 2

[2] Rafael Santandreu. *El arte de no amargarse la vida*. ONIRO, Barcelona, oniro edic edition, 2011. x, 2, 10

[3] Pieter M A Desmet, C J Overbeeke, and Tax Stefan. Designing products with added emotional value. *The Design Journal*, 4(1):32–47, 2001. x, 13, 14

[4] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980. x, 14, 15, 31, 34, 43, 46, 84, 93

[5] JMIR. JMIR Audio Utilities. Recuperado de: http://jmir.sourceforge.net/index_jAudio.html. [Última consulta 06/06/2018]. x, 32

[6] Yesid Ospitia Medina, José Ramón Beltrán, Cecilia Sanz, and Sandra Baldassarri. Dimensional emotion prediction through low-level musical features. In *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound*, AM'19, page 231–234, New York, NY, USA, 2019. Association for Computing Machinery. x, 42

[7] Bruno G. Costa, Jean Carlos Arouche Freire, Hamilton S. Cavalcante, Marcia Homci, Adriana Rosa Garcez Castro, Raimundo Viegas, Bianchi Serique Meiguins, and Jefferson M. Morais. Fault classification on transmission lines using knn-dtw. In *Computational Science and Its Applications – ICCSA 2017*, pages 174–187, Cham, 2017. Springer International Publishing. xi, 97

[8] John A. Sloboda. *The Musical Mind*. Oxford University Press, New York, oxford psy edition, apr 1986. 2, 10, 11, 43, 67, 68, 93

[9] John Powell. *Así es la música*. Titivilus, Barcelona, titivilus edition, 2012. 2, 10, 11, 67, 68, 115

[10] Barbara Kitchenham, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. Systematic literature reviews in software engineering - A systematic literature review. *Information and Software Technology*, 51(1):7–15, 2009. 5

[11] Ken Schwaber and Mike Beedle. *Agile Software Development with Scrum*. Prentice Hall, Upper Saddle River, New Jersey, 2002. 7

[12] Jenefer Robinson and Robert Hatten. Emotions in music. *Music Theory Spectrum*, 34:71–106, 10 2012. 10

[13] Ian Cross & Tolbert and Elizabeth. Music and meaning. In Susan Hallam, Ian Cross, and Michael Thaut, editors, *Oxford Handbook of Music Psychology*. Oxford University Press, 2008. 11

[14] Jianhua Tao and Tieniu Tan. Affective computing: A review. In Jianhua Tao, Tieniu Tan, and Rosalind W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 981–995, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. 11

[15] Renato Eduardo Silva Panda. *Automatic mood tracking in audio music*. Master in informatics engineering, Universidade de Coimbra, 2010. 12

[16] Renato Panda and Rui Pedro Paiva. Music Emotion Classification: Dataset Acquisition and Comparative Analysis. In *5th Int. Conference on Digital Audio Effects (DAFx-12), York, September 17-21, 2012, pp. 1-7.*, pages 1–7, York, 2012. 12, 13, 46, 47, 48

[17] João Fernandes. *Automatic Playlist Generation via Music Mood Analysis*. Msc. thesis, University of Coimbra, 2010. 12, 43, 46, 47

[18] Yi-hsuan Yang and Homer H Chen. Machine Recognition of Music Emotion: A Review. *ACM Transactions on Intelligent Systems and Technology*, 3(3):30, 2012. 12

[19] George Tzanetakis and Perry Cook. MARSYAS: a framework for audio analysis. *Organised Sound*, 4(3):S1355771800003071, dec 2000. 12

[20] Olivier Lartillot, Petri Toiviainen, and Tuomas Eerola. A Matlab Toolbox for Music Information Retrieval. In C. Preisach, H. Burkhardt, Schmidt-Thieme, and R. L., Decker, editors, *Data Analysis, Machine Learning and Applications*, pages 261–268. Springer-Verlag Berlin Heidelberg, Berlin, 2008. 12

[21] Densil Cabrera, Sam Ferguson, and Emery Schubert. PsySound3: Software for Acoustical and Psychoacoustical Analysis of Sound Recordings. In Proceedings of the 13th International Conference on Auditory Display, editor, *Proc. International Conference on Auditory Display*, number July, pages 356–363, Canada, 2007. Proceedings of the 13th International Conference on Auditory Display. 12

[22] Bhagyashree Shirke, Jonathan Wong, Jonathan Libut, Kiran George, and Sang Oh. Brain-iot based emotion recognition system. pages 0991–0995, 01 2020. 13

[23] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. A systematic survey on multimodal emotion recognition using learning algorithms. *International Journal of Intelligent Systems and Applications*, 17, 01 2023. 13

[24] Mordor Intelligence. Emotion detection and recognition market, 2023. 13

[25] Subhasmita Sahoo and Aurobinda Routray. Emotion recognition from audio-visual data using rule based decision level fusion. pages 7–12, 09 2016. 13

[26] Alberto Betella and Paul Verschure. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PLoS ONE*, 11:e0148037, 02 2016. 13

[27] Teah-Marie Bynion and Matthew Feldner. *Self-Assessment Manikin*, pages 1–3. 01 2017. 13

[28] Pieter Desmet. Measuring emotion: Development and application of an instrument to measure emotional responses to products. *Human-Computer Interaction Series*, 3:111–123, 01 2004. 13

[29] Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994. 13

[30] Pieter MA Desmet, Martijn H Vastenburg, and Natalia Romero. Mood measurement with pick-a-mood: review of current methods and design of a pictorial self-report scale. *Journal of Design Research*, 14(3):241–279, 2016. 13

[31] Konstantinos Trohidis, George Kalliris, Griogorios Tsoumakas, and Ioannis Vlahavas. Multi-Label Classification of Music Into Emotions. In ISMIR 2008, editor, *ISMIR 2008 – Session 3a – Content-Based Retrieval, Categorization and Similarity 1*, volume 3a, pages 325–330, Philadelphia, 2008. ISMIR 2008. 14

[32] Yading Song, Simon Dixon, and Marcus Pearce. Evaluation of Musical Features for Emotion Classification. In Fabien Gouyon, Perfecto Herrera, Luis Gustavo Martins, and Meinard M, editors, *13 th International Society for Music Information Retrieval Conference (ISMIR)*, pages 523–528, Porto, 2012. ISMIR 2012. 14

[33] Paul Ekman. *Basic Emotions*, chapter 3, pages 45–60. John Wiley and Sons, Ltd, 1999. 14

[34] Soren Nielzen and Zvonimir Cesarec. Emotional experience of music as a function of musical structure. *Psychology of Music*, 10(2):7–17, 1982. 14, 54, 83, 94

[35] CYNTHIA M. WHISSELL. Chapter 5 - the dictionary of affect in language. In Robert Plutchik and Henry Kellerman, editors, *The Measurement of Emotions*, pages 113–131. Academic Press, 1989. 14

[36] Robert Plutchik. *Emotion, a Psychoevolutionary Synthesis*. Harper & Row, 1980. 14

[37] Tuomas Eerola and Jonna K Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011. 15, 46

[38] Hatice Gunes and Björn Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136, 2013. 15

[39] Cyril Laurier, Mohamed Sordo, Joan Serra, and Perfecto Herrera. Music mood representations from social tags. In *Information Retrieval*, volume 31, pages 381–386, Kobe, 2009. 16

[40] Pasi Saari and Tuomas Eerola. Semantic Computing of Moods Based on Tags in Social Media of Music. *IEEE Transactions on Knowledge and Data Engineering*, 26(10):2548–2560, 2014. 16

[41] Samira Pouyanfar and Hossein Sameti. Music emotion recognition using two level classification. *Proc. Intelligent Systems*, pages 1–6, 2014. 16

[42] You Shyang Chen, Ching Hsue Cheng, Da Ren Chen, and Cheng Huan Lai. A mood- and situation-based model for developing intuitive Pop music recommendation systems. *Expert Systems*, 33(1):77–91, 2016. 16

[43] Cory Mckay. *Automatic Music Classification with jMIR*. PhD thesis, McGill University, 2010. 17

[44] Jacek Grekow. Audio features dedicated to the detection of arousal and valence in music recordings. In IEEE, editor, *2017 IEEE International Conference on Innovations in Intelligent Systems and Applications (INISTA)*, pages 40–44, Gdynia, jul 2017. IEEE. 18, 36, 46, 47

[45] M. R.H. Mohd Adnan, Arezoo Sarkheyli, Azlan Mohd Zain, and Habibollah Haron. Fuzzy logic for modeling machining process: a review. *Artificial Intelligence Review*, 43(3):345–379, 2013. 20

[46] Zadeh LA. Fuzzy sets. *Information and Control*, 8:228–353, 1965. 21, 43

[47] Foram Shah, Madhavi Desai, Supriya Pati, and Vipul Mistry. Hybrid Music Recommendation System Based on Temporal Effects. In *Advances in Intelligent Systems and Computing*, volume 1034, pages 569–577. 2020. 23, 24, 25, 49

[48] Xinxi Wang, David Rosenblum, and Ye Wang. Context-aware mobile music recommendation for daily activities. In *Proceedings of the 20th ACM international conference on Multimedia - MM '12*, number October, pages 99–108, Nara, 2012. ACM. 23

[49] Xinxi Wang, Yi Wang, David Hsu, and Ye Wang. Exploration in Interactive Personalized Music Recommendation: A Reinforcement Learning Approach. *ACM Trans. Multimedia Comput. Commun. Appl.*, 11(1):7:1 – 7:22, 2014. 23

[50] Seungmin Rho, Byeong-jun J Han, and Eenjun Hwang. SVR-based music mood classification and context-based music recommendation. In *Proceedings of the seventeen ACM international conference on Multimedia MM 09*, number January, pages 713–716, Beijing, 2009. ACM. 24

[51] Puja Deshmukh and Geetanjali Kale. A Survey of Music Recommendation System. In *International Journal of Scientific Research in Computer Science,*, volume 3, page 27. 2018. 24, 25, 49

[52] Dip Paul and Subhradeep Kundu. A Survey of Music Recommendation Systems with a Proposed Music Recommendation System. volume 937 of *Advances in Intelligent Systems and Computing*, pages 279–285. Springer Singapore, Singapore, 2020. 24, 49

[53] Yucheng Jin, Nyi Nyi Htun, Nava Tintarev, and Katrien Verbert. ContextPlay: Evaluating user control for context-aware music recommendation. *ACM UMAP 2019 - Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pages 294–302, 2019. 24, 25, 49

[54] Christine Bauer, Marta Kholodylo, and Christine Strauss. Music Recommender Systems Challenges and Opportunities for Non-Superstar Artists. In *Digital Transformation – From Connecting Things to Transforming Our Lives*, pages 21–32. University of Maribor Press, jun 2017. 24, 26, 41, 48, 49

[55] Gabriel Vigliensoni and Ichiro Fujinaga. Automatic music recommendation systems: Do demographic, profiling, and contextual features improve their performance? In *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016*, pages 94–100, 2016. 24, 49

[56] Ivana Andjelkovic, Denis Parra, and John O'Donovan. Moodplay: Interactive music recommendation based on Artists' mood similarity. *International Journal of Human Computer Studies*, 121:142–159, 2019. 25, 49

[57] Jean Garcia-Gathright, Brian St. Thomas, Christine Hosey, Zahra Nazari, and Fernando Diaz. Understanding and Evaluating User Satisfaction with Music Discovery. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '18*, pages 55–64, New York, New York, USA, jun 2018. ACM Press. 25, 49

[58] Andres Ferraro, Dmitry Bogdanov, Kyumin Choi, and Xavier Serra. Using offline metrics and user behavior analysis to combine multiple systems for music recommendation. jan 2019. 25, 49

[59] Christine Bauer and Markus Schedl. *Global and country-specific mainstreaminess measures: Definitions, analysis, and usage for improving personalized music recommendation systems*, volume 14. 2019. 25, 49

[60] Rahul Katarya and Om Prakash Verma. Efficient music recommender system using context graph and particle swarm. *Multimedia Tools and Applications*, 77(2):2673–2687, 2018. 25, 49

[61] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3):330–347, jul 1996. 26, 50

[62] Òscar Celma and Pedro Cano. From hits to niches? or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, NETFLIX '08, New York, NY, USA, 2008. Association for Computing Machinery. 26

[63] Spotify Developer API. Recuperado de: https://developer.spotify.com. [Última consulta 07/06/2023]. 30

[64] Jehan Tristan and Whitman Brian. Echonest. Recuperado de: http://the.echonest.com/. [Última consulta 06/06/2017]. 30

[65] Luis Solarte, Mauricio Sánches, Gabriel Elías Chanchí, Diego Duran, and José Luis Arciniegas. Dataset de contenidos musicales de video, basado en emociones. 7(1):37–46, 2016. 31

[66] Gabriel Elías Chanchí. *Arquitectura basada en contexto para el soporte del servicio de VOD de IPTV móvil, apoyada en sistemas de recomendaciones y streaming adaptativo*. PhD thesis, Universidad del Cauca, 2016. 31

[67] Ivana Andjelkovic, Denis Parra, and John O'Donovan. Moodplay. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization - UMAP '16*, pages 275–279, Canada, 2016. ACM Press. 31

[68] Mohammad Soleymani, Micheal N. Caro, Erik M. Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In ACM New York, editor, *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia - CrowdMM '13*, pages 1–6, Barcelona, 2013. ACM Press. 31

[69] Daniel McEnnis, Cory McKay, Ichiro Fujinaga, and Philippe Depalle. jaudio: An feature extraction library. pages 600–603, 01 2005. 33

[70] Daniel McEnnis, Cory McKay, Ichiro Fujinaga, and Philippe Depalle. jAudio: A feature extraction library. In *Proceedings of the International Conference on Music Information Retrieval*, pages 600–603, 2005. 34

[71] Yong-Hun Cho, Hyunki Lim, Dae-Won Kim, and In-Kwon Lee. Music emotion recognition using chord progressions. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 002588–002593, Hungary, oct 2016. IEEE. 34

[72] Universitat Pompeu Fabra Music Technology Group. AcousticBrainz. Recuperado de: https://acousticbrainz.org. [Última consulta 06/06/2018]. 34

[73] Universitat Pompeu Fabra Music Technology Group. Essentia. Recuperado de: http://essentia.upf.edu/documentation/. [Última consulta 06/06/2018]. 34

[74] Robert Kaye. Musicbrainz. Recuperado de: https://musicbrainz.org/. [Última consulta 06/06/2018]. 34

[75] Cyril Laurier, Owen Meyers, Joan Serra, Martin Blech, and Perfecto Herrera. Music Mood Annotator Design and Integration. In *2009 Seventh International Workshop on Content-Based Multimedia Indexing*, pages 156–161. IEEE, jun 2009. 34

[76] Xiao Hu and J Stephen Downie. Exploring Mood Metadata: Relationships with Genre, Artist and Usage Metadata. *Proceedings of the 8th International Conference on Music Information Retrieval ISMIR'07*, pages 67–72, 2007. 35, 117

[77] Jefferson Martins de Sousa, Eanes Torres Pereira, and Luciana Ribeiro Veloso. A robust music genre classification approach for global and regional music datasets evaluation. In IEEE, editor, *2016 IEEE International Conference on Digital Signal Processing (DSP)*, pages 109–113, Beijing, oct 2016. IEEE. 36

[78] Jacek Grekow. Audio Features Dedicated to the Detection of Four Basic Emotions. volume 9339 of *Lecture Notes in Computer Science*, pages 583–591. Springer International Publishing, Cham, 2015. 36, 46, 47, 48

[79] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA, 2010. Association for Computing Machinery. 36

[80] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, page 835–838, New York, NY, USA, 2013. Association for Computing Machinery. 36

[81] Mohammad Soleymani, Anna Aljanaki, and Yi-Hsuan Yang. DEAM: MediaEval Database for Emotional Analysis in Music. pages 3–5, 2016. 37, 40, 41, 83

[82] Yi-Hsuan Yang and Homer H. Chen. *Music Emotion Recognition*. Taylor & Francis Group, crc press edition, 2011. 38, 44, 45, 118

[83] Yesid Ospitia-Medina, José Ramón Beltrán, and Sandra Baldassarri. Emotional classification of music using neural networks with the MediaEval dataset. *Personal and Ubiquitous Computing*, apr 2020. 38, 54, 70, 111

[84] Rushi Longadge and Snehalata Dongre. Class Imbalance Problem in Data Mining Review. *European Journal of Internal Medicine*, 24(1):e256, may 2013. 38, 48, 72

[85] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002. 39, 40, 83

[86] Fabien Gouyon, Anssi Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle, and Pedro Cano. An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1832–1844, 2006. 39, 40, 83

[87] Edith Law, Kris West, Michael Mandel, Mert Bay, and J. Stephen Downie. Evaluation of algorithms using games : The case of music tagging. In *In Proc. wISMIR 2009*, 2009. 39, 40, 83

[88] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011. 40, 83

[89] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 1041–1044, New York, NY, USA, 2014. Association for Computing Machinery. 39, 40, 83

[90] Karol J. Piczak. Esc: Dataset for environmental sound classification. MM '15, page 1015–1018, New York, NY, USA, 2015. Association for Computing Machinery. 39, 40, 83

[91] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1128–1132, 2016. 39, 40, 83

[92] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 39, 40, 83

[93] Eduardo Coutinho, Felix Weninger, Björn Schuller, and Klaus R. Scherer. The Munich LSTM-RNN approach to the MediaEval 2014 Emotion in Music Task. In *MediaEval 2014 Workshop*, Barcelona, 2014. 43

[94] Mingxing Xu, Xinxing Li, Haishu Xianyu, Jiashen Tian, Fanhang Meng, and Wenxiao Chen. Multi-scale approaches to the MediaEval 2015 (emotion in music) task. In *CEUR Workshop Proceedings*, volume 1436, Wurzen, 2015. 43

[95] Satoru Fukayama and Masataka Goto. Music emotion recognition with adaptive aggregation of Gaussian process regressors. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2016-May, pages 71–75. IEEE, mar 2016. 43

[96] Xinyu Yang, Yizhuo Dong, and Juan Li. Review of data features-based music emotion recognition methods. *Multimedia Systems*, 24(4):365–389, jul 2018. 43

[97] D Meister, H Liao, Y Guo, A Savoy, G Salvendy, P Rau, T Plocher, Y Choong, V G Duffy, P Carayon, K Vu, R Proctor, D Koradecka, and W Karwowski. *Data Mining, Theories, Algorithms, and Examples*. 43

[98] Yi-Hsuan Yang, Chia-Chu Liu, and Homer H Chen. Music Emotion Classification: A Fuzzy Approach. In *Proceedings of the 14th annual ACM international conference on Multimedia - MULTIMEDIA '06*, page 81, New York, New York, USA, 2006. ACM Press. 44, 45

[99] Sanghoon Jun, Seungmin Rho, Byeong-jun Han, and Eenjun Hwang. A fuzzy inference-based music emotion recognition system. In *5th International Conference on Visual Information Engineering (VIE 2008)*, number 543 CP, pages 673–677. IEE, oct 2008. 44, 45

[100] Bin Zhu and Kejun Zhang. Music emotion recognition system based on improved GA-BP. In *2010 International Conference On Computer Design and Applications*, volume 2, pages V2–409–V2–412. IEEE, jun 2010. 44, 45

[101] Mohsen Naji, Mohammd Firoozabadi, and Parviz Azadfallah. Emotion classification during music listening from forehead biosignals. *Signal, Image and Video Processing*, 9(6):1365–1375, sep 2015. 44, 45

[102] Siqi Huang, Li Zhou, Zhentao Liu, Shan Ni, and Jingxian He. Empirical Research on a Fuzzy Model of Music Emotion Classification Based on Pleasure-Arousal Model. In *2018 37th Chinese Control Conference (CCC)*, volume 2018-July, pages 3239–3244. IEEE, jul 2018. 44, 45, 47

[103] Renato Panda, Bruno Rocha, and Rui Pedro Paiva. Dimensional music emotion recognition: Combining standard and melodic audio features. *Proc. 10th International Symposium on Computer Music Multidisciplinary Research*, pages 1–11, 2013. 46, 47

[104] Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. Music Mood Detection Based On Audio And Lyrics With Deep Neural Net. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, pages 370–375, Paris, 2018. 46, 47

[105] Fan Zhang, Hongying Meng, and Maozhen Li. Emotion extraction and recognition from music. In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, ICNC-FSKD 2016*, pages 1728–1733, 2016. 46, 47, 48

[106] Erik M. Schmidt, Douglas Turnbull, and Youngmoo E. Kim. Feature selection for content-based, time-varying musical emotion regression. In *Proceedings of the international conference on Multimedia information retrieval - MIR '10*, page 267, New York, New York, USA, 2010. ACM Press. 46, 47, 48

[107] Junjie Bai, Jun Peng, Jinliang Shi, Dedong Tang, Ying Wu, Jianqing Li, and Kan Luo. Dimensional music emotion recognition by valence-arousal regression. In IEEE, editor, *2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, pages 42–49, Palo Alto, CA, USA, aug 2016. IEEE. 46, 47, 48

[108] Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. Music Mood Detection Based On Audio And Lyrics With Deep Neural Net. *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*, pages 383–391, sep 2018. 47, 48, 57

[109] Hai Tao Zheng, Jin Yuan Chen, Nan Liang, Arun Kumar Sangaiah, Yong Jiang, and Cong Zhi Zhao. A Deep Temporal Neural Music recommendation model utilizing music and user metadata. *Applied Sciences (Switzerland)*, 9(4), 2019. 49

[110] Ferdos Fessahaye, Luis Perez, Tiffany Zhan, Raymond Zhang, Calais Fossier, Robyn Markarian, Carter Chiu, Justin Zhan, Laxmi Gewali, and Paul Oh. T-RECSYS: A Novel Music Recommendation System Using Deep Learning. In *2019 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–6. IEEE, jan 2019. 49

[111] Jinpeng Chen, Pinguang Ying, and Ming Zou. Improving music recommendation by incorporating social influence. *Multimedia Tools and Applications*, 78(3):2667–2687, 2019. 49

[112] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2):95–116, jun 2018. 49, 50, 70

[113] Rui Cheng and Boyang Tang. A Music Recommendation System Based on Acoustic Features and User Personalities. volume 9794 of *Lecture Notes in Computer Science*, pages 203–213. Springer International Publishing, Cham, 2016. 49

[114] Rahul Katarya and Om Prakash Verma. Recent developments in affective recommender systems. *Physica A: Statistical Mechanics and its Applications*, 461:182–190, nov 2016. 49

[115] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, jul 2013. 49

[116] Himan Abdollahpouri, Robin Burke, and Masoud Mansoury. Unfair Exposure of Artists in Music Recommendation. 2020. 50, 51, 52

[117] Marius Kaminskas and Derek Bridge. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. 7(1), December 2016. 50

[118] Dushani Perera, Maneesha Rajaratne, Shiromi Arunathilake, Kasun Karunanayaka, and Buddy Liyanage. A Critical Analysis of Music Recommendation Systems and New Perspectives. In *Advances in Intelligent Systems and Computing*, volume 1152 AISC, pages 82–87. 2020. 51

[119] Alessandro B. Melchiorre, Eva Zangerle, and Markus Schedl. Personality Bias of Music Recommendation Algorithms. In *Fourteenth ACM Conference on Recommender Systems*, pages 533–538. ACM, sep 2020. 51

[120] Exploiting the User Social Context to Address Neighborhood Bias in Collaborative Filtering Music Recommender Systems. *Information*, 11(9):439, sep 2020. 51, 52

[121] Meghan Patil, Sainaya Brid, and Stuti Dhebar. Comparison of different music recommendation system algorithms. *International Journal of Engineering Applied Sciences and Technology*, 5(6):242–248, oct 2020. 51, 52

[122] Himan Abdollahpouri and Masoud Mansoury. Multi-sided Exposure Bias in Recommendation. jun 2020. 51, 52

[123] Dougal Shakespeare, Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. Exploring artist gender bias in music recommendation. *CEUR Workshop Proceedings*, 2697, 2020. 51, 52

[124] Andres Ferraro. Music cold-start and long-tail recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 586–590, New York, NY, USA, sep 2019. ACM. 51, 52

[125] Arthur Flexer, Monika Dorfler, Jan Schluter, and Thomas Grill. Hubness as a Case of Technical Algorithmic Bias in Music Recommendation. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, volume 2018-Novem, pages 1062–1069. IEEE, nov 2018. 51, 53

[126] Markus Schedl. The lfm-1b dataset for music retrieval and recommendation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, page 103–110, New York, NY, USA, 2016. Association for Computing Machinery. 52

[127] Òscar Celma. *Music Recommendation and Discovery*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. 52

[128] PMI, editor. *A Guide to the Project Management Body of Knowledge (PMBOK Guide)*. Project Management Institute, Newtown Square, PA, 5 edition, 2013. 53

[129] Yesid Ospitia-Medina, Sandra Baldassarri, Cecilia Sanz, José Ramón Beltrán, and José A. Olivas. Fuzzy approach for emotion recognition in music. In *2020 IEEE Congreso Bienal de Argentina (ARGENCON)*, pages 1–7, 2020. 55, 111

[130] Anna Aljanaki, Yi Hsuan Yang, and Mohammad Soleymani. Developing a benchmark for emotional analysis of music. *PLoS ONE*, 12(3):1–22, 2017. 57

[131] Yesid Ospitia-Medina, José Ramón Beltrán, Cecilia Sanz, and Sandra Baldassarri. Dimensional Emotion Prediction through Low-Level Musical Features. In ACM, editor, *Audio Mostly (AM'19)*, page 4, Nottingham, 2019. 59, 95, 111

[132] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, apr 2016. 62

[133] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. 4:40–79, 2009. 63

[134] K. Gnana Sheela and S. N. Deepa. Review on Methods to Fix Number of Hidden Neurons in Neural Networks. *Mathematical Problems in Engineering*, 2013:1–11, 2013. 63

[135] Carlos González Morcillo. *Lógica Difusa: Una introducción práctica*. 68

[136] Liping Jing, Kuang Tian, and Joshua Z. Huang. Stratified feature sampling method for ensemble clustering of high dimensional data. *Pattern Recognition*, 48(11):3688–3702, nov 2015. 72

[137] Jingjun Bi and Chongsheng Zhang. An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme. *Knowledge-Based Systems*, 158(June):81–93, 2018. 72

[138] Sotiris Kotsiantis, Dimitris Kanellopoulos, and Panayiotis Pintelas. Handling imbalanced datasets : A review. *Science*, 30(1):25–36, 2006. 72

[139] Chih-Chung Chang and Chih-Jen Lin. LIBSVM. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, apr 2011. 72

[140] Annalyn Ng and Kenneth Soo. Random Forests. In *Data Science – was ist das eigentlich?!*, pages 117–127. Springer Berlin Heidelberg, Berlin, Heidelberg, 2018. 72

[141] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15, dec 2014. 72

[142] Saurabh Karsoliya. Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture. *International Journal of Engineering Trends and Technology*, 3(6):714–717, 2012. 73

[143] Tor Jan Derek Berstad, Michael Alexander Riegler, Håvard Espeland, Thomas De Lange, Pia Helén Smedsrud, Konstantin Pogorelov, Håkon Kvale Stensland, and Pål Halvorsen. Tradeoffs using binary and multiclass neural network classification for medical multidisease detection. *Proceedings - 2018 IEEE International Symposium on Multimedia, ISM 2018*, pages 1–8, 2019. 73

[144] David Diaz-Vico, Anibal R. Figueiras-Vidal, and Jose R. Dorronsoro. Deep MLPs for Imbalanced Classification. *Proceedings of the International Joint Conference on Neural Networks*, 2018-July, 2018. 73

[145] Sarah Vluymans, Alberto Fernández, Yvan Saeys, Chris Cornelis, and Francisco Herrera. Dynamic affinity-based classification of multi-class imbalanced data with one-versus-one decomposition: a fuzzy rough set approach. *Knowledge and Information Systems*, 56(1):55–84, 2018. 73

[146] Gustavo E. A. P. A. Batista, Ana L. C. Bazzan, and Maria C. Monard. Balancing Training Data for Automated Annotation of Keywords: a Case Study. *In Proceedings of the Second Brazilian Workshop on Bioinformatics*, pages 35–43, 2003. 75

[147] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, and Eli Woods. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21:1–6, 2020. 96

[148] Marco Cuturi and Mathieu Blondel. Soft-DTW: A differentiable loss function for time-series. *34th International Conference on Machine Learning, ICML 2017*, 2:1483–1505, 2017. 96

[149] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining concepts and techniques, third edition*. Morgan Kaufmann Publishers, Waltham, Mass., 2012. 110

[150] Yesid Ospitia Medina, Sandra Baldassarri, and José Ramón Beltrán. High-Level Libraries for Emotion Recognition in Music: A Review. In Vanessa Agredo and Pablo Ruiz, editors, *Human-Computer Interaction. HCI-COLLAB 2018.*, pages 158–168, Popayán, 2019. Springer. 111

[151] Yesid Ospitia-Medina, Sandra Baldassarri, and José Ramón Beltrán. *Librerías de alto nivel para el reconocimiento de emociones en la música*, pages 173–186. Editorial Bonaventuriana, 2020. 111

[152] Yesid Ospitia-Medina, Sandra Baldassarri, Cecilia Sanz, and José Ramón Beltrán. *Music Recommender Systems: A Review Centered on Biases*, pages 71–90. Springer International Publishing, Cham, 2023. 112

[153] Yesid Ospitia-Medina, José Ramón Beltrán, and Sandra Baldassarri. Ensa dataset: a dataset of songs by non-superstar artists tested with an emotional analysis based on time-series. *Personal and Ubiquitous Computing*, 6 2023. 112

[154] Angel Candelaria. *Teoría de la Música: Niveles 1 - 2*. CreateSpace Independent Publishing Platform, 2014. 115