DESARROLLO DE UN SISTEMA DE RECONOCIMIENTO DE EMOCIONES EN LA MÚSICA, BASADO EN TÉCNICAS DE APRENDIZAJE PROFUNDO

SANTIAGO ALBERTO TRUJILLO SANCHEZ

INSTITUCIÓN UNIVERSITARIA COLEGIO MAYOR DEL CAUCA
FACULTAD DE INGENIERIA
INGENIERÍA INFORMÁTICA
POPAYÁN-CAUCA
2024

DESARROLLO DE UN SISTEMA DE RECONOCIMIENTO DE EMOCIONES EN LA MÚSICA, BASADO EN TÉCNICAS DE APRENDIZAJE PROFUNDO

MONOGRAFÍA PRESENTADA COMO REQUISITO PARA OPTAR AL TITULO DE INGENIERÍA INFORMÁTICA

AUTOR
SANTIAGO ALBERTO TRUJILLO SANCHEZ

DIRECTOR

MG. FRANCISCO JAVIER OBANDO

INSTITUCIÓN UNIVERSITARIA COLEGIO MAYOR DEL CAUCA

FACULTAD DE INGENIERIA

INGENIERÍA INFORMÁTICA

POPAYÁN-CAUCA

2024

I. Tabla de contenido

1. CAPITULO 1 - INTRODUCCIÓN 1

1.1. RESUMEN 1

1.2. INTRODUCCION 2

1.3. PLANTEAMIENTO DEL PROBLEMA 3

1.4. OBJETIVOS 4

1.4.1. OBJETIVO GENERAL 4

1.4.2. OBJETIVOS ESPECÍFICOS 5

1.5. JUSTIFICACION 5

1.6. MARCO TEORICO 6

1.6.1. CONCEPTOS GENERALES: 6

1.6.2. ANTECEDENTES 10

2. CAPÍTULO 2 - REVISIÓN DE LA LITERATURA 12

2.1. ESTUDIOS PRIMARIOS 13

2.2. METODOLOGÍA 17

2.3. RESULTADOS 19

2.3.1. Pregunta N°1 20

2.3.2. Preguntas N°2 26

2.3.3. Pregunta N°3 33

2.3.4. Pregunta N°4 39

2.4. CONCLUSIONES DE LA REVISIÓN 44

3. CAPÍTULO 3 - APLICACIÓN DE LA METODOLOGIA CRISP-DM, PARA IDENTIFICACIÓN DE EMOCIONES EN LA MÚSICA 45

3.1. FASE 1 COMPRENSIÓN DEL NEGOCIO 45

3.1.1. EVALUACIÓN DEL CONTEXTO 45

3.1.2. OBJETIVOS DEL NEGOCIO 46

3.1.3. RECURSOS 46

3.1.4. COSTOS Y BENEFICIOS 47

3.1.5. RIESGOS Y CONTINGENCIAS 48

3.1.6. TERMINOLOGÍA 49

3.1.7. OBJETIVOS DE LA MINERÍA DE DATOS 51

3.1.8. CRITERIOS DE ÉXITO DE LA MINERÍA DE DATOS 51

3.1.9. PLAN DEL PROYECTO 51

3.2. FASE 2 ENTENDIMIENTO DE LOS DATOS: 53

3.2.1. REPORTE DE BÚSQUEDA INICIAL Y SELECCIÓN DE DATOS 53

3.2.2. ESTRUCTURA Y CONTENIDO DE LOS DATOS 54

3.3. FASE 3 PREPARACIÓN DE LOS DATOS 56

3.3.1. CONSTRUCCIÓN, LIMPIEZA Y TRANSFORMACIÓN DE DATOS 56

3.4. FASE 4 MODELADO 63

3.4.1. SELECCIÓN DE TÉCNICA DE MODELADO 63

3.4.2. UNIÓN DE CONJUNTO DATOS 64

3.4.3. PREPROCESAMIENTO DE DATOS 65

3.4.4. ETIQUETAS ONE-HOT O ETIQUETAS DUMMY 65

3.4.5. DIVISIÓN EN CONJUNTOS DE ENTRENAMIENTO Y VALIDACIÓN 66

3.4.6. OPTIMIZADOR 67

3.4.7. FUNCIÓN DE COSTE 68

3.4.8. MÉTRICAS 68

3.4.9. CONSTRUCCIÓN DEL MODELO 69

3.5. CONSTRUCCIÓN DEL MODELO EXPERIMENTACIÓN 72

FASE 5 PRUEBAS Y EVALUACION: 118

FASE 6: DESPLIEGUE 121

4. ALCANCES Y LIMITACIONES 122

5. CONCLUSIONES Y TRABAJOS FUTUROS 123

5.1 CONCLUSIONES 123

5.2 TRABAJOS FUTUROS 125

6. BIBLIOGRAFIA 126

7. ANEXOS 132

II. Lista de Figuras

Figura 1. Porcentaje de Uso Modelos 26

Figura 2. Características por estudios 33

Figura 3. Estudios enfocados en Clasificación, Fuente: Propia 42

Figura 4. Estudios enfocados en Regresión, Fuente: Propia 43

Figura 5. Ejemplo de las etiquetas en el dataset original, Fuente: Propia 55

Figura 6. Distribución de etiquetas del dataset Soundtracks, Fuente: Propia 56

Figura 7. Construcción de fragmento, Fuente: Propia 58

Figura 8. Desplazamiento del tiempo, Fuente: Propia 59

Figura 9. Incorporación del ruido, Fuente: Propia 60

Figura 10. Estiramiento del fragmento, Fuente: Propia 61

Figura 11. Ganancia aleatoria a un fragmento, Fuente: Propia 62

Figura 12. Formula de la precisión correctos vs toral de clases 68

Figura 13. Formula de la precisión, Matrix de confusión 69

Figura 14. Formula de la precisión 72

Figura 15. Rendimiento del modelo 1 en el conjunto de entrenamiento, Fuente: Propia 75

Figura 16. Rendimiento del modelo 1 en el conjunto de validacion, Fuente: Propia 76

Figura 17. Rendimiento del modelo 2 en el conjunto de validación, Fuente: Propia 84

Figura 18. Rendimiento del modelo 2 en el conjunto de validación, Fuente: Propia 85

Figura 19. Rendimiento del modelo 3 en el conjunto de entrenamiento, Fuente: Propia 92

Figura 20. Rendimiento del modelo 3 en el conjunto de validación, Fuente: Propia 93

Figura 21. Rendimiento del modelo 4 en el conjunto de entrenamiento, Fuente: Propia 100

Figura 22. Rendimiento del modelo 4 en el conjunto de validación, Fuente: Propia 101

Figura 23. Ejemplo de partición de conjunto de datos durante la validación cruzada, Fuente: Propia 107

Figura 24. Rendimiento y perdida de los conjuntos de entrenamiento y validación durante la interacción 1, Fuente: Propia 109

Figura 25. Rendimiento y perdida de los conjuntos de entrenamiento y validación durante la interacción 2, Fuente: Propia 111

Figura 26. Rendimiento y perdida de los conjuntos de entrenamiento y validación durante la interacción 3, Fuente: Propia 114

Figura 27. Rendimiento y perdida de los conjuntos de entrenamiento y validación durante la interacción 4, Fuente: Propia 117

Figura 28. Rendimiento y perdida de los conjuntos de entrenamiento y validación durante la interacción 5, Fuente: Propia 117

III. Lista de Tablas

Tabla 1. Preguntas de Investigación 17

Tabla 2. Puntaje de inclusión a la bitácora 18

Tabla 3. Resultados por base de datos Científica 19

Tabla 4. Estudios primarios Pregunta 1 20

Tabla 5. Estudios primarios Pregunta 2 27

Tabla 6. Estudios primarios Pregunta 3 34

Tabla 7. Resultados Pregunta 3 40

Tabla 8. Recursos Proyecto 47

Tabla 9. Riesgos y contingencias 48

Tabla 10. Etiqueta emocional 65

Tabla 11. Etiqueta target procesada 66

Tabla 12. Resultados de entrenamiento modelo 1 74

Tabla 13. Precisión obtenida en las 3 CFG del modelo 1 77

Tabla 14. Recall obtenida en las 3 CFG del modelo 1..... 78

Tabla 15. F1-Score obtenido en las 3 CFG del modelo 1 79

Tabla 16. Resumen de los datos obtenidos en la matriz de confusión del modelo 1 80

Tabla 17.Resultados del entrenamiento del modelo 2..... 82

Tabla 18. Precisión por emoción obtenidas durante el entrenamiento del modelo 2 86

Tabla 19. Recall por emocion obtenidas durante el entrenamiento del modelo 2..... 87

Tabla 20. F1-Score por emoción obtenidas durante el entrenamiento del modelo 2 88

Tabla 21. Resumen de los datos obtenidos en la matriz de confusión del modelo 2 89

Tabla 22. Resultados generales del entrenamiento del modelo 3..... 91

Tabla 23. Precisión por emoción obtenidas durante el entrenamiento del modelo 3 94

Tabla 24.Recall por emoción obtenidas durante el entrenamiento del modelo 3..... 94

Tabla 25. F1-Score por emoción obtenidas durante el entrenamiento del modelo 3 95

Tabla 26. Resumen de los datos obtenidos en la matriz de confusión del modelo 3 96

Tabla 27. Resultados generales del entrenamiento del modelo 4..... 98

Tabla 28 Precisión por emoción obtenidas durante el entrenamiento del modelo 4 102

Tabla 29. Recall por emocion obtenidas durante el entrenamiento del modelo 4..... 102

Tabla 30. F1-Score por emocion obtenidas durante el entrenamiento del modelo 4 103

Tabla 31. Resumen de los datos obtenidos en la matriz de confusión del modelo 4 104

Tabla 32. Resumen métricas obtenidas en el entrenamiento de la interacción 1 109

Tabla 33.Resumen métricas obtenidas en el entrenamiento de la interaccion 2 112

Tabla 34. Resumen metricas obtenidas en el entrenamiento de la interaccion 3 115

Tabla 35. Resultados de las pruebas de los datasets con entrenamiento normal 119

Tabla 36. Resultado de las pruebas con validación cruzada 119

1. CAPITULO 1 - INTRODUCCIÓN

1.1. RESUMEN

La detección de emociones constituye un área de investigación clave que permite la identificación de patrones emociones, con diversas aplicaciones como la psicología, medicina e industria musical por lo que en este proyecto de grado se centró en el desarrollo de una herramienta computacional destinada para la detección de emociones en la música empleando técnicas de machine learning, técnicas de aprendizaje profundo y de procesamiento de audio.

Para llevar a cabo este proyecto, se seleccionó un dataset con fragmentos de música etiquetados en un comienzo de duración variada de entre 15 segundos a 1 minuto de duración. Estos fragmentos fueron sometidos a diversas técnicas de aumento de datos con el objetivo de ampliar la cantidad de información del dataset original. Además, se aplicaron diferentes técnicas para la extracción de características específicas de la música como tono, amplitud, ritmo y armonía, entre otras. Por ultimo los datos fueron sometidos a un preprocesamiento en el cual las etiquetas emociones predeterminadas del dataset original se convirtieron en etiquetas tipo one-hot las cuales son ideales para problemas de clasificación multiclase o donde haya más de una etiqueta. Asimismo, después de este proceso se experimentó con 2 tipos de entrenamiento para determinar daba mejores resultados con el modelo el primero fue un entrenamiento normal en el cual se dividió el conjunto de datos en dos subconjuntos el cuales se les denomino como entrenamiento y validación y en el cual se dividió en dos partes o ejes conocidos como X y Y donde en X albergaba todas las características extraídas con anterioridad y en Y las etiquetas emocionales en formato one-hot. El segundo fue la validación cruzada con 5 fold o pliegos el cual dividía el dataset en 5 partes iguales en las cuales todas las partes formarían parte entrenamiento

del modelo a diferencia del entrenamiento con división de conjuntos en la cual solo toma una parte del dataset como entrenamiento.

Todo este proceso se hace con el objetivo de garantizar que las características fueran idóneas para el entrenamiento del modelo para la detección de emociones en la música. Una vez finalizado el proceso de entrenamiento, se generaron diversas estadísticas y metricas para evaluar el rendimiento del modelo algunas como la matriz de confusión, accuracy, precisión, recall y F1-Score.

1.2. **INTRODUCCION**

La música es un arte que ha acompañado a la humanidad a lo largo de la historia y que tiene gran influencia en nuestras vidas cotidianas ya que tiene la capacidad de transmitir emociones y sentimientos de esta forma evocando una amplia gama de emociones a quienes la escuchan, con el avance de la IA se ha desarrollado un gran interés en la detección automática de emociones en la música ya que las personas tienden a escuchar diferentes canciones cuando se encuentran en diferentes estados emocionales, por lo tanto, categorizar música de acuerdo al tipo de emociones se vuelve más y más importante [3]. Esto es un tema de investigación interesante e importante en los últimos años, ya que puede proporcionar una visión más profunda de como la música influye en la mente de una persona y por lo tanto lo vuelve un problema complejo ya que requiere un análisis cuidadoso de diferentes características que componen a la música [5]. Las técnicas de procesamiento de señales de audio han sido de gran ayuda en este campo ya que puede extraer características relevantes de la música para la detección de emociones [3] de esta forma y con la utilización de diversas técnicas de ML y aprendizaje profundo se busca abordar esta problemática tales como SVG [14], C4.5 [2], Naive Bayes [2]. O se hace uso de técnicas más sofisticadas como el uso de encefalogramas para descubrir lazos

entre estados emociones provocados por la música y la actividad cerebral el cual se lleva a cabo usando técnicas conocidas como EGG y clasificadores como el SVM [13] incluso se usan se usa el rostro de las personas como punto de partida de tal manera que se detecten cambios faciales que ayuden a detectar que emociones produce una determinada canción en una persona y como esta afecta su estado de ánimo [6]. Como se mencionó anteriormente hay muchas técnicas, pero todas buscan el solucionar el mismo objetivo. Para el este proyecto desarrollo una herramienta que sea capaz de detectar emociones en un fragmento de música utilizando un enfoque basado en aprendizaje profundo o Deep learning. Con el fin de cumplir con estos objetivos, se realizó una evaluación y revisión de diferentes algoritmos de aprendizaje profundo para determinar cuál es el más adecuado para esta tarea. Además, se buscó y utilizaron bases de datos disponibles para entrenar y validar el modelo que, con el uso de diversas métricas de evaluación como precisión, recall, F1-Score. Matrices de confusión y curvas ROC obtienen resultados precisos y confiables. Con estas herramientas y estrategias, se espera logra el desarrollo de una herramienta computacional automatizada y precisa que permita la detección de emociones de un amanera eficiente y efectiva.

1.3. **PLANTEAMIENTO DEL PROBLEMA**

La música ha sido siempre un medio popular para relajar y dejar la presión de la vida cotidiana. Sin embargo, la percepción de la misma podría cambiar bajo diferentes emociones, por ejemplo, si la canción favorita de una persona se escucha estando en estado triste o enojado es muy probable que tome otro significado si se escucha estando en un estado feliz [20]. Numerosos métodos han sido usados en el pasado con el fin de identificar y evaluar emociones humanas, las mayores técnicas usadas para el reconocimiento de emociones incluyen el uso de material textual, expresiones faciales, diálogos, movimiento del cuerpo y señales psicológicas [21]

El reconocimiento de emociones en la música es una tarea importante en diversas áreas, como la psicología, la industria musical y la tecnología. El análisis de las emociones en la música puede proporcionar información valiosa para la comprensión del comportamiento humano y para la creación de experiencias musicales personalizadas. Sin embargo, la detección precisa de las emociones en la música es un desafío debido a la complejidad y subjetividad de la interpretación emocional.

Métodos tradicionales de machine learning toman características acústicas extraídas manualmente de señales de audio como entradas, generalmente se toman características acústicas básicas que contienen ritmo, timbre, tono y volumen además que existen muchas características preprocesadas como ZCR, MFCC, etc. Sin embargo, si analizamos estas características por sí solas no nos darían una visión clara de qué característica pueden afectar las emociones cuando se escucha música. [3]

Para abordar estos problemas, se propone el desarrollo de una herramienta automatizada de reconocimiento de emociones basada en técnicas de aprendizaje profundo. El objetivo de este proyecto es diseñar e implementar un modelo de aprendizaje automático capaz de detectar y clasificar las emociones presentes en fragmentos de música basándose en las etiquetas presentes en un conjunto de datos. Para lograr este objetivo, se requiere la revisión de la literatura existente, la exploración y selección del algoritmo de aprendizaje profundo más adecuado y la implementación de un módulo de software que permita el reconocimiento de emociones en la música.

1.4. **OBJETIVOS**

1.4.1. OBJETIVO GENERAL

Desarrollar un sistema de reconocimiento de emociones basado en técnicas de aprendizaje profundo.

1.4.2. OBJETIVOS ESPECÍFICOS

- Identificar los modelos definidos por la computación afectiva aplicables en la música y las técnicas de procesamiento de audio.
- Entrenar un algoritmo de aprendizaje profundo que sea capaz de detectar emociones asociadas con la música a partir de características extraídas de un dataset.
- Evaluar el rendimiento del algoritmo de aprendizaje profundo

1.5. **JUSTIFICACION**

La detección de emociones en la música puede tener una amplia gama de aplicaciones prácticas en el mundo real, como por ejemplo la clasificación automática de canciones según el estado de ánimo o según la lista de reproducción de un usuario para tener recomendaciones personalizadas en aplicaciones como Spotify o YouTube.

La emoción en la música puede ayudar a mejorar la personalización de servicios y contenidos ofrecidos por aplicaciones de música, muchos trabajos de investigación basados en el uso de técnicas de machine learning han apuntado al problema de reconocimiento de emociones en la música durante los últimos años y los resultados obtenidos solo se han aplicado a repositorios de música de tamaño muy pequeños o no se han hecho públicos y no consideran como los usuarios se sienten en realidad cuando escuchan canciones este tipo de lógica impide que las propuestas existentes se no integren a mecanismos de personalización de la red como Spotify o YouTube Music [22]

Lo que se busca con este proyecto es desarrollar una herramienta computacional que permita identificar las emociones presentes en la música de manera automatizada lo que puede tener un gran impacto tanto en campo de la industria musical como por ejemplo en el campo de la psicología, en este último campo específicamente puede ser usado como apoyo a procesos terapéuticos. A nivel de investigación, este proyecto contribuye al campo de la inteligencia artificial y el procesamiento de señales lo que puede tener diversas aplicaciones en campos de la vida real como la medicina, seguridad y tecnología

La propuesta en este proyecto es usar una un algoritmo de aprendizaje profundo para el reconocimiento de emociones en la música permitiendo el procesamiento automático de grandes cantidades de datos de audio y la detección de patrones complejos de los mismo, la utilización de estos tipos de algoritmos puede mejorar significativamente la precisión y eficiencia del proceso de detección de emociones en la música lo que la hace una herramienta valiosa para diversa aplicación prácticas.

1.6. **MARCO TEORICO**

1.6.1. **CONCEPTOS GENERALES:**

- Emoción: Las emociones son reacciones afectivas intensas que suponen un impulso a actual, una condición psicológica y biológica de respuesta a ciertos estímulos internos y externos al organismo con frecuencia traducido a gestos, actitudes u otras formas de expresión, en el caso de la música pueden ser vistas como la armonía, melodía, ritmo, letra, intensidad u otros elementos de una canción [23]

- Etiquetado emocional: Es la tarea de reconocer emociones para clasificarlas en su correspondiente categoría. Dada una entrada clasificarla como emoción, neutral o no emoción como una o varias de las varias emociones que mejor representen el estado mental de una persona

- Características de audio: Son descripciones de sonido o señales de audio que pueden ser alimentadas en modelos estadísticos o ML para construir sistemas de audio inteligentes. Algunas aplicaciones que usen estas características pueden ser:

Clasificación de audio

Reconocimiento de voz

Etiquetado automático de música

Segmentación de audio

Eliminación de sonido del audio

A su vez diferentes características pueden captar diferentes aspectos de sonidos generalmente las categorías de audio son categorizadas por los siguientes aspectos:

Nivel de abstracción: El cual comprende 3 niveles para la clasificación de señales de audio: señales de audio nivel alto los cuales son características abstractas que son entendidas y disfrutadas por los humanos las cuales incluyen instrumentación, acordes, melodía, armonía, ritmo, género, humor, etc. Señales de audio nivel medio son características que están un paso más arriba en la abstracción y representación de datos en comparación a características de un nivel más bajo, pero aún no son completamente abstractas o semánticas tales como el tono, patrones de fluctuación, MFCC, etc. A menudo se extraen para representar aspectos específicos de la señal de audio que son relevantes en diversas aplicaciones. Señales de audio de bajo nivel las cuales son representaciones crudas y básicas de características extraídas de un audio o imagen y la cual una máquina puede entender perfectamente, normalmente la información se extrae directamente de los datos de entrada sin un procesamiento o abstracción significativa algunos ejemplos son la amplitud de onda en un dominio de tiempo, spectral centroid, spectral flex, zero-crossing rate. [28]

- Temporal Scope: En el contexto del MER y procesamiento de audio se refiere a la ventana de tiempo sobre la cual se aplican ciertas operaciones o se extraen ciertas características, cuando se busca hacer extracciones de características se puede usar alguno de estos enfoques: Instantáneo el cual considera un punto específico en el tiempo aquí las características se calculan en un instante particular sin tener un contexto temporal donde se usan pequeños fragmentos de una señal de audio en el rango de milisegundo.
- Segmentación aquí se busca dividir la señal de entrada en segmentos o ventanas temporales más pequeñas y aplicar la extracción de características a cada segmento, Global en este caso en particular se usa la señal en su totalidad sin subdividirla en segmentos lo cual puede ser útil en ciertas tareas que requieran una visión general de la señal completa [28]

- Uso de enfoques de ML: Los enfoques tradicionales de ML consideran todos o la gran mayoría de características tanto de tiempo como de frecuencia como entrada para el modelo. Las características necesitan ser escogidas cuidadosamente ya que estas afectarán el modelo y su rendimiento algunas de las más comunes y usadas son el ZCR, RMS, SC. BER y características denominadas Chroma y SB. [28]

- Deep Learning: Es un subconjunto de las redes neuronales que abarca estadísticas y modelos predictivos tienen una tarea crucial en la tarea de la ciencia de datos ya que se utilizan para analizar grandes conjuntos de datos y extraer patrones complejos, este enfoque considera representaciones de audio sin estructurar como espectrogramas o MFCC los cuales pueden extraen patrones por su cuenta, las características que

normalmente se utilizan para alimentar la red neuronal son los espectrogramas, espectrogramas de mel o MFCC. [28]

- Procesamiento de señales de audio: Conjunto de técnicas de señales de audio que se utilizan para analizar y manipular señales de audio, las técnicas más usadas incluyen el análisis de espectro, la extracción de características como el MFCC, ZCR, RMS, diversas de características chroma, entre otros [28]

- Dataset: Se refiere a archivos que contienen uno más registros los cuales pueden tener información tales como registros médicos o registros de seguros para ser usados por un sistema informático, además los datasets son usados para guardar información necesitada por aplicaciones o sistemas operativos tales como macro librerías, sistemas variables o parámetros

- Machine Learning: Es una técnica de análisis de datos dentro de la inteligencia artificial que pretende dotar a las máquinas o sistemas informáticos de la capacidad para asociar, procesar información y aprender de la experiencia, todo ello directamente de los datos que recibe sin ser programadas de manera explícita [26]. El proceso de machine learning crea modelos que generalizan la información recibida para entregar una salida y realizar predicciones [26]

- Redes Neuronales: Son un modelo matemático que se inspira en el funcionamiento del cerebro humano, consiste en un conjunto de nodos interconectados o mejor visto como

neuronas que procesan y transmiten información, cada neurona recibe entradas, las procesa y produce una salida que puede la entrada a otras neuronas [46]

1.6.2. ANTECEDENTES

En [8] el autor propone el uso de una CLDNN con la creación de un dataset de música turca desde cero, este dataset consta de 124 fragmentos de música turca cada una con una duración máxima de 30 segundos. Para la extracción de características de los fragmentos musicales se lleva a cabo una división previa en la cual cada fragmento se segmenta en tres partes con una duración de 10 segundos cada uno. Se extraen dos tipos de características por fragmento las MFCC y las log-mel filterbank energy estas se introducen a la CNN compuesta de cuatro capas convolucionales y cuatro capas MaxPooling generando por 768 características compuestas por la MFCC y log-mel filterbank además que se extraen características de tipo estándar como tono, espectrales, chroma, etc. Para el entrenamiento del modelo se usa una LSTM+DNN y la evaluación del modelo se realiza usando la técnica de validación cruzada con 10 interacciones los resultados obtenidos indican que el clasificador supera otros algoritmos como los KNN, SVM y bosques aleatorios logrando una precisión de 99.19%

En [16] el autor propone la implementación de una CRNN basado en un enfoque de regresión. Este enfoque implica la “superposición” de capas convolucionales y recurrentes para realizar predicciones continuas de emociones, tomando como referencia el modelo de Thayrel que se centra en las dimensiones de valencia y arousal. Para extraer características se emplean las Log mel-band energy las cuales según el autor capturan información equivalente a características de alto, medio o bajo nivel. Para evaluar el modelo se realiza utilizando la métrica RMSE la cual compara las predicciones hechas por el modelo con valores reales, un valor bajo de RMSE indica que las predicciones del modelo están cercanas a los valores reales de las etiquetas. Se

presentan resultados utilizando diferentes longitudes de secuencia variando en 10 a 60 lo resultados expuestos en las dimensiones de valencia oscilan entre 0.268 y 0.275 mientras para arousal oscilan entre 0.060 a 0.085

En [40] el autor se inspira en la arquitectura de la red VGGnet la cual se utiliza para el procesamiento de imágenes para desarrollar una red personalizada CNN. Se hacen diversos experimentos con diversos datasets como SoundTrack y Bi-modal y el autor propone una técnica novedosa de post-procesamiento para mejorar el rendimiento de los datasets. Se adopta la estrategia de solo tomar el diagrama de Rusell como referencia para las etiquetas emocionales por lo que “happy”, “sad”, “angry” y “tender” como “neutral” son lo que el autor utiliza. El autor utiliza una arquitectura de red diferente a una CNN convencional donde se usan múltiples capas por bloques esto puede ser beneficioso ya que permite al modelo aprender patrones más complejos, pero a su vez puede tener una arquitectura muy compleja lo cual puede afectar el entrenamiento y aumentar el uso de recursos computacionales. Por último, el autor para evaluar su red utiliza otras redes como SVM, RF Y NN con características extraídas de los clips de música de los dataset como características espectrales, temporales y MFCC los resultados obtenidos no son muy alentadores y el autor afirma que encontrar una combinación entre red + características no es sencillo y se concluye que la CNN muestra mejores resultados entre 10% a 15% mejor que las otras arquitecturas.

En [41] el propone usar una CNN para la predicción de emociones musicales utilizando clips de audio como entrada y generando vectores de emociones bidimensionales en valencia y arousal como salida, se emplean diferentes técnicas de extracción y transformación de características para convertir los clips en espectrogramas y se usa una técnica llamada sliding window con diferentes anchos para garantizar la continuidad musical en la predicción de emociones. Se evalúa el rendimiento usando métricas denominadas Top-1, Top-3 y Fuzzy-3 accuracy con el

dataset EmoMusic y demuestra que la ventana (sliding window) de 5 segundos con 50% de superposición logra un rendimiento óptimo logrando resultados de 83% en precisión en valencia y 82% en arousal con características como STFT, CQT, MFCC, Chroma STFT Y CQT

En [43] la autora ha creado diferentes modelos de red neuronal completamente conectada o densa junto con la construcción de un dataset desde cero. En las primeras etapas del proyecto se enfocó en la selección y etiquetado manual de canciones. Durante este proceso se asocian diferentes géneros musicales con emociones como por ejemplo el pop y hip hop con la emoción alegría mientras que otros géneros como el Metal o Hard Rock se asocian con la emoción enfado. Una vez completado el etiquetado, se eligen los 30 segundos más significativos de cada canción y se extraen características con la librería Essentia para su posterior uso en experimentos con modelos ML. La autora ajusta constantemente los modelos durante el desarrollo del proyecto, realizando modificaciones en aspectos como el número de capas ocultas y el optimizador del modelo. Este tipo de enfoque tiene como objetivo alcanzar el mejor equilibrio posible entre la pérdida y la precisión del algoritmo buscando así optimizar los resultados obtenidos en los experimentos

2. CAPÍTULO 2 - REVISIÓN DE LA LITERATURA

Para el desarrollo del sistema de reconocimiento de emociones en la música, basado en técnicas de aprendizaje profundo, se realizó una revisión de la literatura. Para desarrollar esta herramienta se llevaron a cabo diversas actividades, siendo la principal la identificación de los estudios realizados en el campo de detección de emociones en la música, conocido como MER, para ellos se realizó una investigación en distintas bases de datos como Google Scholar, IEEE Xplorer, Springer, entre otras.

Se recopilaron un total de 25600 publicaciones, de las cuales se seleccionaron 35 artículos considerados relevantes para convertirse en estudios primarios. Tras leer y analizar estos

artículos utilizando un sistema de puntuación basado en las preguntas de investigación el cual puede ser visto en la tabla 2 en el cual se daba una puntuación de 1.25 si el articulo contestaba una pregunta de investigación o 0 si el articulo no contestaba la pregunta basada en este filtro se descartaron 17 de los artículos iniciales, quedando finalmente 18 artículos que se utilizaron como base de investigación. Estos artículos proponen diversos algoritmos para la detección de emociones en la música, los cuales fueron analizados y comparados teniendo en cuenta diferentes aspectos, como las herramientas empleadas para la extracción de características musicales, los enfoques utilizados y los resultados obtenidos.

Como Palabras claves se especifican las siguientes: Aprendizaje profundo, emociones, música, redes convolucionales.

El presente capitulo está organizado de la siguiente manera: la sección 2 está dedicada a explicar algunos de los estudios primarios que se usaron en esta investigación. En la sección 3 se explica la metodología que se usó en la revisión sistemática y se generan las preguntas de investigación a ser respondidas con los estudios primarios. La sección 4 está dedicada a analizar cada uno de los estudios primarios escogidos de tal forma que cada uno responda las preguntas de investigación. Finalmente, en la sección 4 se presentan conclusiones y trabajos futuros derivados de la presenten investigación

2.1. **ESTUDIOS PRIMARIOS**

En este apartado se presentan algunas de las técnicas empleadas para el reconocimiento de emociones en la música haciendo uso de diferentes enfoques y redes de aprendizaje profundo en donde se abordan diferentes temas como la extracción de características necesarias para la identificaron de emociones en la música, entrenamiento de las redes, diferentes enfoques como son la regresión y clasificación, procesamiento de información en base a las características extraídas y por ultimo pruebas y rendimiento de la red de aprendizaje profundo

En el estudio titulado “Music emotion recognition using convolutional long short term memory deep neural networks” se utiliza una arquitectura CLDNN para detectar emociones en la música turca. Esta arquitectura utiliza la salida de una CNN como características y emplea una combinación de LSTN + DNN como clasificador. Además, se utiliza una CNN unidimensional para extraer características para el proceso de detección de emociones. El estudio también menciona que la profundidad de la red puede tener un impacto positivo o negativo, si la red es muy profunda existe el riesgo de saturación y esto podría afectar diferentes medidas como precisión, recall, accuracy, etc.

Para extraer las características necesarias para la entrada de la CNN, se utilizan MFCC, log-mel y spectrum. Por último, para evaluar el rendimiento de la red se crean tres conjuntos de datos llamados Fea_set_X donde “x” es un número del 1 al 3. Se realizan pruebas individuales a cada uno de los conjuntos y se obtiene un rendimiento entre el 89% y el 90% con una precisión que varía entre el 80% y el 82%, Al hacer combinación entre conjuntos y especialmente cuando se utilizan juntos se obtiene un rendimiento del 99% y una precisión del 85%. [8]

En el estudio denominado “Automatic Mood Detection and Tracking of Music Audio Signals” se extraen características específicas que se utilizan en base al modelo de Thayer. En concreto se extraen tres características: intensidad, timbre y ritmo las cuales se considera son los suficientemente eficiente para evaluar con este modelo. La intensidad se representa mediante la energía de las sub-bandas obtenidas en cada interacción. El timbre tiene subcaracterísticas que se componen de forma espectrales y contraste espectral. Por otro lado, el ritmo indica tres aspectos cercanamente relacionados a la respuesta individual de las emociones en la música las cuales incluye la fuerza del ritmo, la regularidad del ritmo y el tempo del ritmo. Se menciona que usan un Framework jerárquico para detectar la emoción en un fragmento de música basándose en estas tres características extraídas, después se clasifica en un grupo el clip basándose en el

humor de las características extraídas y por último se aplica el algoritmo de clasificación dentro cada grupo basándose en las características de ritmo y timbre. Para evaluar el desempeño del sistema, se utilizó una base de datos de 800 clips musicales. Los resultados mostraron una precisión promedio de emociones del 86.3% y un recall promedio del 84.1%. [11]

En el estudio denominado “EEG-Based Emotion Recognition in Music Listening” los investigadores se enfocan en el uso de grabaciones de electroencefalogramas para clasificar estados emocionales en la música usando algoritmos de (ML). Se hace uso de varios algoritmos para detectar emociones en la música, siendo el algoritmo principal una combinación de la estimación espectral (EEG) y el Support Vector Machine (SVM) donde se permite identificar cuatro estados emocionales como son la alegría, ira, tristeza y placer.
El SVM se utilizó para realizar la clasificación de características extraídas por el EEG en especialmente el DASM12 y se usó un enfoque de validación cruzada con una configuración de 10 pasos para obtener resultados más precisos, se menciona que se utiliza un enfoque de clasificación para detectar las emociones usando el SVM y se lograron resultados bastante prometedores ya que se proporciona información acerca de la precisión del algoritmo que fue del 82% en la detección de emociones evaluadas y la clasificación de las emociones de alegría, tristeza, ira y placer se obtienen resultados entre el 80% y 90% esto indica una capacidad considerable para distinguir entre diferentes estados emocionales en la música usando este algoritmo. [14]

En el estudio denominado “Music Emotion Recognition Based on a Neural Network with an Inception-GRU Residual Structure” el objetivo principal es mejorar la precisión en la clasificación y reconocimiento de emociones en la música. Para lograrlo se propone el uso de una red neuronal convolucional (CNN) de una dimensión (1D CNN) con una estructura residual llamada inception-GRU la cual permite extraer características y conexión temporales en el procesamiento

de señales en la música, se utilizó el dataset SoundTrack y se aplicaron métodos de aprendizaje superficial como SVM y bosques aleatorios.
Los resultados muestran que el método de aprendizaje propuesto, en conjunto con la estructura residual de un 1D CNN con el Inception-GRU da un porcentaje de precisión del 84% dando un rendimiento mayor en la clasificación y reconocimiento de emociones musicales por último se menciona que se diseñaron y extrajeron características adecuadas para el proyecto se utilizó la transformada de Fourier rápida (FFT) para procesar las muestras y determinar sus características espectrales. El enfoque propuesto se compara con métodos de aprendizaje poco profundos como máquinas de vectores de soporte (SVM) y bosques aleatorios. [13]

En el trabajo “Music Emotion Classification Method Based on Deep Learning and Explicit Sparse Attention Network” se tiene como objetivo mejorar la precisión del reconocimiento y clasificación de las emociones en la música, lo que se hizo para lograr esto es la utilización de un método que combina una red de aprendizaje profundo con una red de atención dispersa explicita para el procesamiento de un dataset, se usan técnicas de segmentación detallada y otros métodos de pre procesamiento para proporcionar un dataset de entrada de alta calidad para el modelo de clasificación. Se menciona que se realizan pruebas utilizando dataset reales, los resultados muestran que la precisión de reconocimiento del método propuesto es del 0,71% con emociones felices en la música y de 0,69 para emociones tristes en la música lo que indica que tienen una buena capacidad de reconocimiento y clasificación de emociones musicales, se describe el método para extraer características de audio que incluye la segmentación final de audio y la separación de voz se extraen características como características de bajo nivel (LLD) y de alto nivel (HSF), frecuencia de mel (MFCC), tasa de cruce por cero, ancho de banda espectral, atenuación espectral y flujo espectral. por último, se describe el modelo de clasificación de audio basado en redes neuronales de múltiples capas que combina una CNN y una LSTM el modelo utiliza las características de espectrograma y características LSTM como entrada las cuales

fusionan las características extraídas de ambas para mejorar el rendimiento de la clasificación. [17]

2.2. **METODOLOGÍA**

El objetivo de este estudio es investigar y evaluar los métodos de reconocimiento de emociones en la música utilizando técnicas de aprendizaje profundo. Para lograr esto, se realizó un análisis exhaustivo de las propuestas existentes en este campo, como el fin de identificar las técnicas y recursos más efectivos. Se examinaron diferentes enfoques, técnicas, algoritmos y características necesarios para la detección de emociones en la música, así como se analizaron los resultados obtenidos en cada uno de los estudios.

Con el objetivo de abordar esta investigación de manera más organizada, se plantearon cuatro preguntas de investigación las cuales permiten un análisis más detallado

Tabla 1. Preguntas de Investigación

N° Pregunta	**Pregunta**
Q1	¿Qué modelos son los más efectivos para la detección de emociones en la música?
Q2	¿Qué tipos de características se extraen de las señales de audio para poder identificar emociones en la música?
Q3	¿Qué investigaciones se han realizado en base a los enfoques de clasificación y regresión y cual ha dado mejores resultados en el campo de la detección de emociones en la música?
Q4	¿Cuál es la precisión y eficiencia de los enfoques de clasificación y regresión en la detección de emociones en la música?

Para responder cada una de las preguntas, se llevó a cabo una recolección de información utilizando diferentes bases de datos científicas como IEEE Xplorer, Google Scholar, Springer, Research Gate. Considerando un rango de tiempo entre los años 2006 y 2023.

Con el fin de identificar los estudios más relevantes para la investigación, se estableció un sistema de búsqueda basado en palabras clave como “Deep learning” “Convolutional networks”, “Music”, “emotions”, entre otros. Además, que se tomaron en cuenta aspectos como título, palabras clave y abstract para determinar la relevancia del mismo

Para evaluar la calidad y relevancia de los estudios seleccionados, se diseñó un sistema de puntuación basado en las cuatro preguntas de investigación anteriormente formuladas.

Tabla 2. Puntaje de inclusión a la bitácora

Puntaje	Criterio de inclusión
1.25	El estudio contesta con la pregunta de investigación propuesta
0.65	El estudio contesta parcialmente la pregunta de investigación propuesta
0	El estudio no esta relacionado o no contesta la pregunta de investigación

Cada pregunta se calificó en una escala de 0 a 1.25, donde 1.25 indicaba una respuesta completa y relacionada con la pregunta, un puntaje de 0.65 se les asignaba a estudios se enfocaba ligeramente en la pregunta y un puntaje de 0 se les asignaba a proyectos que no tenían nada que ver con la o las preguntas de investigación.

Finalmente, se aplicó un filtro para seleccionar los estudios primarios que serían considerados para la investigación. Aquellos estudios que obtenían una puntuación mayor a 3 fueron incluidos y llamados estudios primarios, lo cual resulto en 18 estudios de los 35 estudios seleccionados. Los 17 estudios restantes que no cumplieron con la puntuación mínima establecida, se registraron en una bitácora, pero no se tomaron en cuenta para la investigación principal.

2.3. **RESULTADOS**

En la tabla 2 se muestran el consolidado de los estudios realizados y seleccionados en cada una de las bases de datos científicas. En total se seleccionaron 35 estudios de los cuales 18 fueron seleccionados como estudios primarios después de evaluarlos con un sistema de puntaje

Tabla 3. Resultados por base de datos Científica

Fuente de investigación	**Estudios encontrados**	**Estudios relevantes**	**Estudios seleccionados**
1. MDPI	2	1	1
2. IEEE Xplorer	1108	6	4
3. Springer / SpringerLink	300	2	1
4. Google Scholar	9200	12	7
5. Google Scholar/Research Gate	250	3	3
6. Google	15000	9	2

A continuación, se exponen los resultados obtenidos tras analizar cada uno de los estudios primarios, intentando responder a cada una de las preguntas de investigación

2.3.1. Pregunta N°1

¿Qué modelos son los más efectivos para la detección de emociones en la música?

Todos los estudios seleccionados como estudios primarios muestran que por lo menos se usó un modelo para la detección de emociones en la música. En la tabla 3 se presentan los estudios primarios que abordan a la primera pregunta de investigación y proporcionan información pertinente para la investigación la tabla se divide en 3 columnas la primera columna es el número de investigación, la segunda columna es el nombre del estudio y la tercera columna es el modelo o red utilizado en el estudio.

Tabla 4. Estudios primarios Pregunta 1

Referencia	**Nombre del estudio**	**Modelo**
[2]	Multi-Modal Music Emotion Recognition A new Dataset, Methodology and Comparative Analysis	Se hicieron varios test con diferentes modelos de aprendizaje supervisado como son SMV, K-Nearest neighbors, C4.5 y Naive Bayes
[3]	Audio-based deep music emotion recognition	Se hace una estrategia para reconocer emociones contenida en canciones por medio de espectrogramas haciendo uso de una CNN y SVM para comparar resultados
[4]	A regression approach to music emotion recognition	Se usa una CNN para este proyecto

Referencia	Nombre del estudio	Modelo
[5]	Desarrollo de un sistema de detección del nivel de música sobre voz en señales de audio	Se busca hacer un sistema que busque detectar música sobre voz en señales de audio usando CNN para el procesado de imágenes y audio utilizando 2 bases de datos
[6]	Reconocimiento multi-modal de emociones orientadas al aprendizaje	Se crea un modelo hibrido para la extracción de caracterisrticas con un modelo CNN y las capas de extracción del modelo usando una CNN-LSTM con capas de endebidos
[7]	Audio Features for music emotion recognition a survey	No se enfocan en un solo método para la detección de emociones buscan identificar cual es el mejor. Se exponen diferentes investigaciones en: SVM los cuales son más utilizados en problemas de clasificación Las ANN para redes neuronales profundas Las Dyadic SVM las cuales se tienen en cuenta la relación temporal entre emociones

Referencia	Nombre del estudio	Modelo
[8]	Music emotion recognition using CNN long short term memory	Se hace uso de una arquitectura CLDNN para la detección de emociones en la música turca esta arquitectura usa la CNN como salida y el LSTN + DNN como clasificador
[9]	A survey of music emotion recognition	Este artículo se trata como una encuesta, pero da a conocer algoritmos para la detección de emociones como: -LSMT (Long Short-term Memory) los cuales se utilizan en combinación mecanismos de atención para modelar las emociones en la música a nivel de una canción -Se hacen mención de los BiLSTM se usan con el modelo de arousal y valencia (Russell) - los GRAN el cual es un modelo de aprendizaje profundo utilizando para el reconocimiento de emociones con base en la BD de emotify

Referencia	Nombre del estudio	Modelo
		- Por último, se mencionaron que las CNN y RNN se utilizan en combinación para la detección de emociones usando tanto el audio como las letras (Partitura)
[10]	An Efficient Classification Algorithm For Music Mood Detection In Western and Hindi Music Using Audio Feature Extraction	Se hace uso de una ANN con enfoque de clasificación de emociones en el modelo de Thayer con una red neuronal de múltiples capas la cual se programa para detectar las emociones en una pieza de música
[11]	Automatic Mood Detection and Tracking of Music Audio Signals	No se indica que red neuronal se utiliza Se describe un trabajo de investigación sobre la detección de emociones usando características acústicas
[12]	Multi-Level Speech Emotion Recognition based on HMM and ANN	Se hace enfoque en el uso de las ANN con un enfoque de clasificación en el modelo de Thayer
[13]	EEG-BASED EMOTION RECOGNITION IN MUSIC LISTENING: A COMPARISON OF	Se utilizan varios algoritmos para la detección de emociones en la música el principal algoritmo fue la combinación de la estimación

Referencia	Nombre del estudio	Modelo
	SCHEMES FOR MULTICLASS SUPPORT VECTOR MACHINE	espectral (EEG) y el support vector machine (SVM)
[14]	Music Emotion Recognition Based on a neural Network with an Inception-GRU Residual Structure	Se propone el uso de una red neuronal convolucional (CNN) de una dimensión en conjunto con una estructura residual llamada inception-GRU Se utilizan métodos de aprendizaje superficial como SVG y bosques aleatorios
[15]	Music emotion recognition using recurrent neural networks and pretrained models	Se presentan experimentos realizados utilizando redes neuronales recurrentes (RNN) para detectar emociones en la música a través segmentos musicales
[16]	STACKED CONVULTIONAL AND RECURRENT NEURAL NETWORKS FOR MUSIC EMOTION RECOGNITION	Se propone un método basado en redes CNN y RNN. Se utilizó una capa de la CNN seguido con dos ramas de la RNN entrenadas separadamente para la valencia y arousal
[17]	Music Emotion Classification Method Based on Deep Learning	Se combina una CNN y una LSTM el modelo utiliza las características de

Referencia	Nombre del estudio	Modelo
	and Explicit Sparse Attention Network	espectrograma y características LLDS como entrada y se fusionan las características extraídas de ambas para mejorar el rendimiento de la clasificación
[18]	Music Emotion Classification Method Based on Deep Learning And improved Attention Mechanism	En este estudio se usan una CNN y una LSTM usando técnicas de integración de mecanismos de atención se construyen un modelo de análisis de emociones basados en CNN-LSTM
[19]	DNN BASED Music Emotion Recognition from Raw Audio Signal	Se analiza un enfoque novedoso para el reconocimiento de emociones en la música basado en CNN y RNN el objetivo es extraer características de las señales de audio sin necesidad de pre procesamiento y utilizarlas para reconocer las emociones que pueden surgir en la música

A partir de la información recopilada se puede destacar que:

La red neuronal más utilizada para la detección de emociones en la música es la red neuronal convolucional (CNN) que se usa en un 31% de los proyectos (E2, E3, E4, E5, E8, E13, E14, E15, E16, E17, E18), la red que le sigue es la SMV (E1, E2, E6, E12) y la RNN (E8, E14, E15, E18)

con 11% de los proyectos, después se encuentran las ANN (E6, E9, E11) y LSTM (E8, E16, E17) las cuales son usadas en 9% de los proyectos, por último se encuentran redes como las GRAN, C4.5, SVG, etc. que cumplen la tarea de reconocer emociones pero se usan en un proyecto en específico.

Cabe destacar que muchos proyectos usan una o más de una red y cada red tiene una tarea en específico lo cual hace que calificar esto sea un poco complicado como por ejemplo E1, E2, E8 así que se le coloca 1 punto a cada red que se utilizó en el proyecto.

La figura 1, muestra porcentaje de uso de las redes.

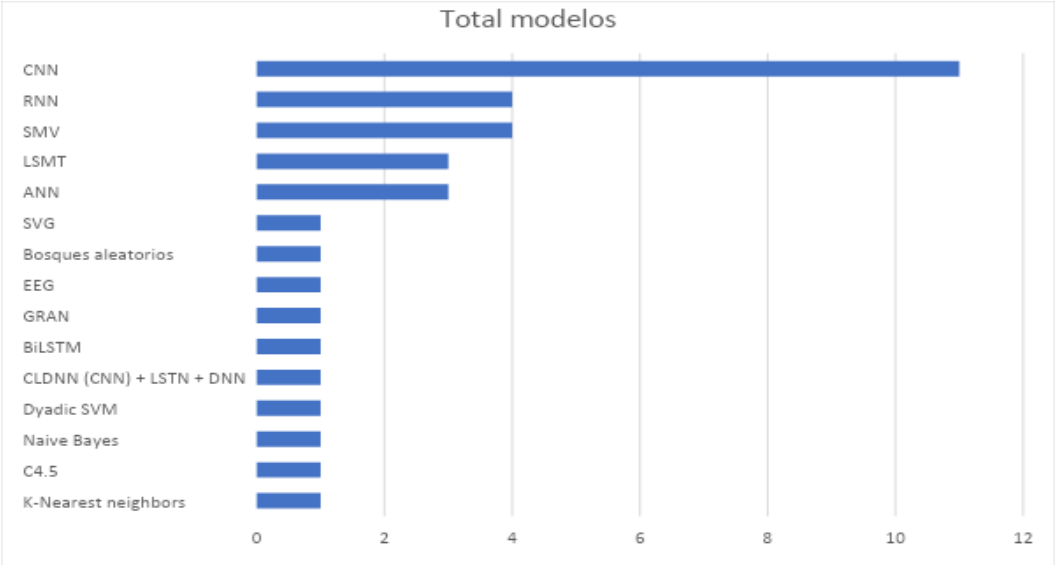


Figura 1. **Porcentaje de Uso Modelos**

2.3.2. Preguntas N°2

¿Qué tipos de características se extraen de las señales de audio para poder identificar emociones en la música?

La gran mayoría de los estudios seleccionados como estudios primarios destacaron las características extraídas para la detección de emociones en la música, estas características en

combinación con los modelos propuestos en el punto anterior nos dan un punto de partida para el desarrollo del proyecto. En la tabla 4 se presentan los estudios primarios que abordan a la segunda pregunta de investigación y proporcionan información pertinente para la investigación dividida en tres columnas la primera columna es la identificación del estudio, la segunda columna el nombre del estudio, la tercera las características extraídas por el estudio para la detección de emociones en la música

Tabla 5. Estudios primarios Pregunta 2

N°	Características extraídas	Información relevante
E1	-Se usan frameworks como Marsyas, MIR Toolbox y PsySound para extraer características. -Con PsySound se pueden extraer características como sonoridad, nitidez, aspereza, fuerza de fluctuación, tono y ritmo -MIR ToolBox se usa específicamente para la extracción de características musicales como tonalidad, timbre, pitch, etc. -Marsyas es un framework para el desarrollo de procesamiento de audio que permite la extracción de características como tempo, MFCC y características espectrales -Con estas tres herramientas se extraen 3 tipos de características las cuales son las características de audio estándar, características MIDI y características de audio melódico	-Se combina información de diferentes fuentes como son el audio, MIDI y Lyrics -Se usa un dataset similar al de MIREX con 5 clusters con diferentes emociones entre si -PsySound se utiliza para el análisis de grabaciones de sonido
E2	Se utilizan espectrogramas en esta investigación en el cual se extraen 2 características típicas de este enfoque las cuales son las MFCC y características de Mel (Mel-Frequency Cepstral Coefficients)	-Se usa un dataset de 1000 canciones por cada canción el dataset provee un fragmento de una canción de 45 segundos y un par de valores de valencia y arousal -Con el dataset anterior se obtienen 30498 espectrogramas -La CNN se entrena con 40000 pasos lo que es aproximadamente 85 epochs
E3	-Se utilizan frameworks y herramientas como son PsySound y Marsyas para la extracción de características en la música -Se para la extracción de características se usan MLR, SVR y BoostR	-Se utilizaron algoritmos como DWCH y spectral contrast -Se basan en el diagrama de Thayrel -Se extraen un total de 114 características

N°	Características extraídas	Información relevante
	-Con PsySound se extraen 44 características las cuales incluyen loudness, level, pitch mulplicity y dissonance -Con Marsyas se extraen 30 características como textura timbral, contenido rítmico y características de pitch -Con Spectral contrast se extraen 12 características Y con DWCH se extraen 28 características	
E4	Se extraen características MFCC y características de MEL	Se usa la arquitectura InceptionV3 para la red neuronal
E5	Se da a entender que se extraen características con la red CNN, pero no se especifica o dicen que tipo de características se extrajeron	Se utiliza un sistema hibrido de capas convolucionales con memorias largas a corto plazo para la extracción de características y una capa de fusión de vectores para entrenar la red usando capas completamente conectadas
E6	-Se extrajeron características relacionadas al tono y melodía: Pitch - Se extrajeron características relacionadas con el ritmo y percusión: El tempo, patrones rítmicos y energía percusiva - Se extrajeron características relacionadas con la dinámica y expresividad Volumen, ataque y decaimiento - Se extrajeron características relacionadas con el timbre e instrumentación -Espectro de frecuencia: es la distribución de energía en diferentes bandas de frecuencia -Armónicos y parciales: Componentes adicionales en el espectro que contribuye al timbre	Al ser más que todo una encuesta no se extraen características en sí, pero se da exponer cuales son las importantes para detectar emociones en la música
E7	-Se usa características de audio estándar las cuales son extraídas utilizando herramientas como MIRtoolbox, OpenSmile y jAudio -MIRtoolbox se encarga de extraer características como el timbre, tonality y pitch	-Se crea una base de datos de música emocional de música turca desde cero compuesta de 124 fragmentos de 30 segundos de música turca tradicional

N°	Características extraídas	Información relevante
	-El JAudio se encarga de características como harmonic change detection function (hcdf), mode, inharmonicity, tonal, chromogram, key clarity, tempo, y fluctuation -OpenSmile es capaz de extraer numerosas características low-level descriptors (LLD) las cuales son usualmente obtenidas por espectro de tiempo-corto como son energy, MFCC, Mel-based features, pitch y spectral features	-Se usa una CNN de una dimensión para extraer características que consiste de una o más capas convolucionales. -Cada capa tiene un numero de filtros para producir nuevos mapas de características
E8	-Características espectrales -Características temporales las cuales son características que se basan en el análisis temporal de señales de audio las cuales incluyen el tempo, duración de notas, silencias, la tassa de cambios de amplitud y variaciones rítmicas -Características armónicas: Las cuales se enfocan en la estructura armónica de la música y las cuales incluyen la detección de acordes. Estimación de tonalidad, relación de frecuencias armónicas -Características emocionales: La cuales se relacionan directamente con las emociones expresadas en la música pueden incluir medidas de valencia, arousal, expresividad emocional e intensidad	
E9	Se extraen características acústicas como intensity, timbre, pitch y rhythm en cada pieza de música y las clasifican de la siguiente manera: -Características de intensidad la cual da indicación del grado de loudness or calmness de la música y da pistas para medir el estrés en la composición de la música y tiene dos importantes entidades que identificar esto como son la amplitud y la energía baja -Características de timbre es determinado por el perfil harmónico o la fuente del sonido y el cual tiene se computa con los siguientes valores: zero crossing, spectral irrgularity y roughness -Características de tono esta característica corresponde a la bajeza o alteza relativa que se puede escuchar en una canción. Los dos componentes que se consideran a los efectos de la taxonomía del estado de ánimo son	Se buscan extraer características relacionadas entre sí como características espectrales y harmónicas Se consideran 8 diferentes humores con música india y occidental

N°	Características extraídas	Información relevante
	el tono en sí y la inarmónica. -Características de ritmo: El ritmo es repetitivo lo cual se relaciona con lo periódico y frecuencia por lo tanto sigue los siguientes factores para determinar el humor en una pieza de música: Beat/Beat Spectrum, Tempo, Fluctuations	
E10	Se extraen características como intensity la cual se menciona que es una característica esencial en la detección del estado de ánimo de la música y se utiliza la suma del espectro de la señal y espectro de subbandas para calcular la intensidad de cada fotograma El timbre en formato spectrum y se utilizan características como brillo, ancho de banda, rolloff y flujo espectral para discriminar diferentes estados de ánimo y se utilizan características basadas en la forma espectral y el contraste espectral basado en octavas para representar el timbre en la música Tempo el cual se menciona que está estrechamente relacionado con la respuesta del estado del ánimo y se proponen 5 características como fuerza del ritmo, regularidad del ritmo. Tempo promedio, frecuencia de onsets y pico de correlación promedio	Estas características se extraen de los clips de música mediante técnicas como el muestreo, la transformada de Fourier rápida (FFT), el análisis de envolvente de amplitud y la detección de onsets. Las características se calculan en cada fotograma o en todo el clip de música, y se utilizan para construir conjuntos de características que se utilizan en el sistema propuesto de detección de estados de ánimo.
E11	Se hace pie en que se debe tener las características necesarias para que este modelo funcione y se pueda detectar la emoción por lo que se seleccionan cinto características importantes según el autor como son el tono, intensidad, formato. LPCC y MFCC	Se buscan extraer características importantes y compararlas los valores obtenidos con los valores más acordes dependiendo de la canción que se esté evaluando. Para lograr lo anterior extraen ciertas características para una emoción determinada ya que creen que extraer una sola característica y determinar la emoción a partir de la misma sería algo no muy de fiar y los valores de una sola característica puede equivocadamente representar la emoción de la canción.
E12	En si no extraen características de la música, sino que se usan características denominadas DASM12, RASM12, PSD24 Y PSD30 que son características extraídas de los datos obtenidos en	

N°	Características extraídas	Información relevante
	el EEG y que se utilizan en el proceso de clasificación emocional basado en EEG	
E13	Se extraen diferentes características de la música como son Mel-Frequency Cepstrum Coefficient (MFCC), Zero Crossing Rate (ZCR) y pitch para el diseño de características de timbre	Se extrajeron características adecuadas para el proyecto se utilizó la transformada de Fourier rápida (FFT) para procesar las muestras y determinar sus características espectrales.
E14	Se extraen características como cruce por cero, el centro espectral, el flujo espectral, la rolloff espectral, los coeficientes cepstrales de frecuencia MFCC y características cromáticas	Se menciona la extracción de características usando herramienta de análisis de audio y recuperación como Essentia y Marsyas Ambas herramientas calculan las estadísticas sobre las características extraídas como la media, varianza y estadísticas de orden superior estas características se utilizan para construir vectores de características que representan cada segmento de música analizado
E15	Se extraen características como media y derivación estándar de acústicas de bajo nivel Características como MFCC, características spectral como flux, centroid, kurtosos, rolloff y características de voz Caracteristicas de raw audio: Log mel-band, specifically, log mel-band Metricas: RMSE	Se extraen un total de 260 características La media y desviación estándar tiene frames de 60 a 50 ms
E16	Se describe el método para extraer características de audio que incluye la segmentación final de audio y la separación de voz se extraen características como características de bajo nivel (LLD) y de alto nivel (HSF), frecuencia de mel (MFCC), tasa de cruce por cero, ancho de banda espectral, atenuación espectral y flujo espectral por último se describe el modelo de clasificación de audio basado en redes neuronales de múltiples capas	Se menciona que se realizan pruebas utilizando dataset reales
E17	En este método se extraen características de las letras utilizando el método Term Farequency-Inverse Document Frequency (TF-IDF) y Word2vec	El TF-IDF es una técnica utilizada para medir la importancia de un término en un documento dentro de una colección de documento se calcula multiplicando dos factores la frecuencia de termino y la frecuencia inversa de documento

N°	Características extraídas	Información relevante
	El TF-TDF se utiliza para asignar peso a cada termino en función de su importancia relativa en un documento especifico en comparación con la colección completa de documentos Word2vec es un modelo de aprendizaje utilizado para representar palabras como vectores números en un especio segmentico	
E18	Para realizar el pre procesamiento y utilizarlas para reconocer las emociones que pueden surgir en la música, se utiliza: -Una capa convolucional unidimensional (1D-CNN) -Una capa de conexión completamente conectada en el tiempo (TD-FC) -Unidad recurrente bidireccional de compuerta (BiGRU) Estas capas se aplican para poder extraer las características y aprender información temporal de la señal de audio	

A partir de la información recopilada se puede determinar que:

Cada estudio se extraen características específicas para poder alimentar la o las redes neuronales utilizadas para lograr el objetivo de detectar emociones en la música. Algunos estudios incluso no llegan a extraer características de la música, sino que usan otros medios como partituras, espectrogramas incluso usan ligeros cambios faciales cuando se está escuchando una canción en específico. Cada estudio de investigación puede tener una o muchas características extraídas para la detección de emociones. En la figura 2 se muestran cuantos estudios extrajeron ciertas características destacando que las características espectrales y pitch fueron de las características más usadas en los proyectos siendo la primera usada en 9 de los 18 estudios primarios y la segunda usada en 6 de los 18 estudios primarios, también se encontraron características menos comunes como LDD, Calmness y dissonance que fueron usados solo en 1 de 18 estudios primarios los cuales seguramente se usaron para una tarea en específico. Por último, se tiene un apartado de “No” donde eran proyectos que no extraían características y en su lugar usaban espectrogramas o partituras

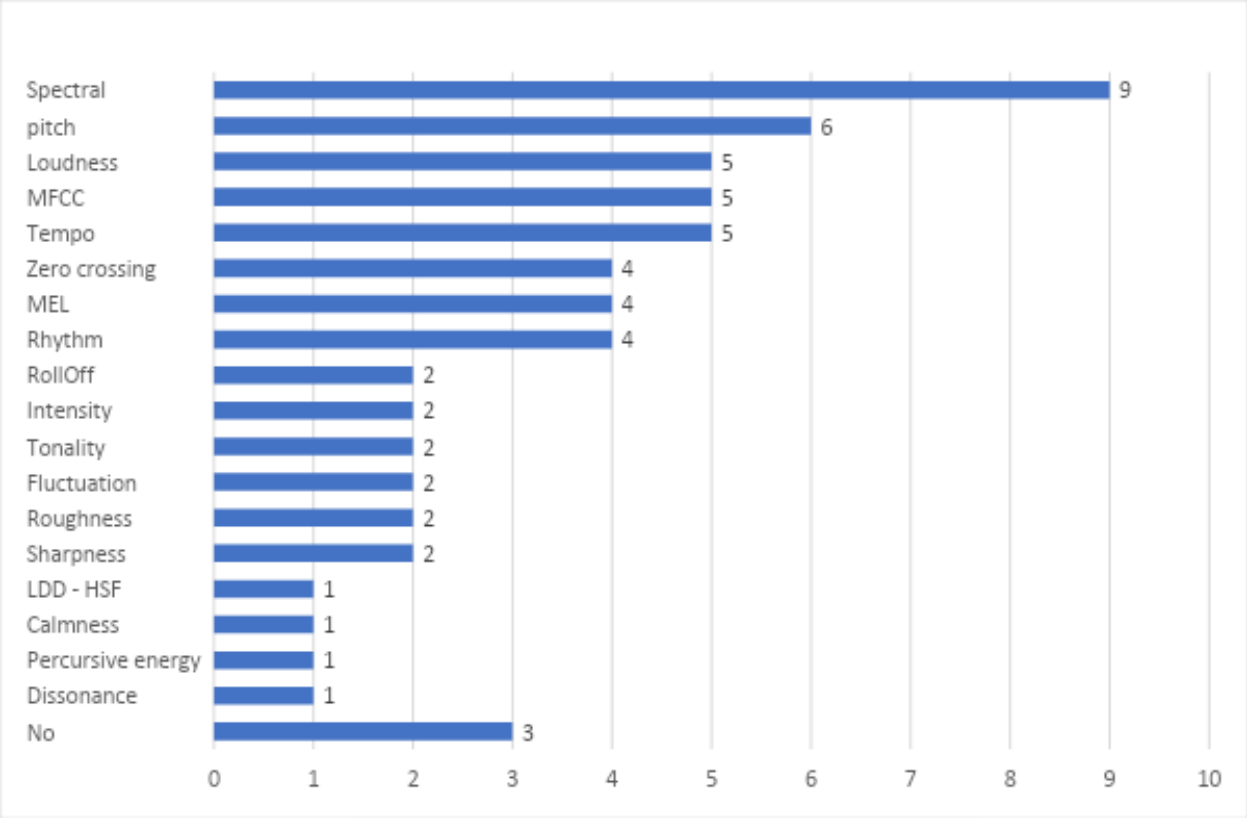


Figura 2. Características por estudios

2.3.3. **Pregunta N°3**

¿Qué investigaciones se han realizado en base a los enfoques de clasificación y regresión y cual ha dado mejores resultados en el campo de la detección de emociones en la música

Cada estudio primario seleccionado puede tener un enfoque de clasificación o regresión cada uno tiene tanto sus ventajas como desventajas por ejemplo en el enfoque de clasificación se puede tener una interpretación más clara de los datos ya que la clasificación proporciona resultados discretos y categorizados, asignando una etiqueta emocional a cada instancia de música lo cual facilita la interpretación de los resultados pero también tiene sus desventajas como que se puede perder la información continuamente ya que al asignar una etiqueta emocional discreta a cada instancia se puede perder información como la intensidad o grado de emoción.

Es por eso que en la tabla N°5 se busca ver cuántos proyectos han utilizado estos enfoques y a partir del enfoque seleccionado cual ha dado mejor resultado.

La tabla se divide de la siguiente forma en la primera columna se coloca la ID del estudio, la segunda y tercera columna son los enfoques para detectar emociones en la música los cuales si el estudio lo menciona será marcado una (X), la última columna son notas e información importante la cual se usó junto al enfoque para poder detectar emociones en la música.

Tabla 6. Estudios primarios Pregunta 3

N°	Clasificación	Regresión	Notas
E1	✓		-Se usó el algoritmo Relief con el fin de generar un peso para cada característica en función de cual es más importante para la clasificación -El proceso de clasificación se validó con un repetido estratificado de 20 repeticiones (10-fold cross validation)
E2	✓		-El proceso de clasificación se hace utilizando un 10-fold cross validation el cual afirma que supera procesos similares hechos con redes SVM
E3		✓	-Se tienen dos criterios para escoger un dataset las canciones deberán estar distribuidas uniformemente en cada cuadrante emocional basándose en el plano de Thayrel y cada fragmento musical deberá expresar una emoción dominante -Se entrenan tres algoritmos de regresión MLR, SVR y BoostR
E4	✓		Se dividen los datos en tres carpetas Training (64%), validation (16%), y test (20%)

N°	Clasificación	Regresión	Notas
E5	✓		Se utiliza la técnica de validación cruzada con k=5 siendo k el numero de interacciones que tiene el modelo ya que según el autor toma de 1 a 2 días el entrenamiento del mismo específicamente con los modelos que usan capas de endebidos Se usa una distribución de 80% para entrenamiento y 20% para validación sin conjunto de pruebas ya que ese rol lo toma el conjunto de validación Las métricas usadas para este proyecto fuera la precisión y perdida para el entrenamiento y el numero y tamaño de las clases
E6	✓	✓	-En la clasificación se busca asignar una etiqueta emocional discreta a cada pieza musical, se entrenan los modelos de clasificación utilizando un dataset etiquetados con emociones y por último se utilizan algoritmos de aprendizaje automático para predecir la emoción a partir de características extraídas - Por el lado del enfoque de regresión busca predecir las respuestas emocionales continúas asociadas en la música como la valencia o la excitación (Russell) en este enfoque se entrena el modelo usando un dataset con puntuaciones emocionales continuas y se utilizan puntuaciones a partir de características extraídas

N°	Clasificación	Regresión	Notas
E7	✓		-Se utilizó una clasificación con 10-fold cross validation -El dataset es dividido aleatoriamente en 10 partes iguales después la data es estratificada para obtener la misma distribución clase aproximadamente -Después nueve partes son separadas para seleccionar la data de entrenamiento las demás partes son para calcular la probabilidad de error este proceso se repite 10 veces
E8			No se menciona que enfoque como clasificación o regresión son más efectivos o utilizados para la detección de emociones ya que se centra en dar un vistazo general al estado actual y desafíos en el campo del MER
E9	✓		Se usa una red neuronal multicapa con un algoritmo de Back propagation La red se inicializa primero configurando todos sus pesos para que sean pequeños números aleatorios y la entrada se aplica desde stereomood.com que consiste de las cuatro características extraídas de las canciones: tono, ritmo, intensidad y timbre que representan varios estados de ánimo que han sido aceptado por la gente de todo el mundo. Por lo tanto, hay son 4 capas en la entrada y 8 capas para representar 8 puntuaciones de estado de ánimo en la salida Más tarde, las características de las canciones de ambos

N°	Clasificación	Regresión	Notas
			Los géneros indios y occidentales se prueban dados como insumos para la red neuronal y se clasifican en diferentes estados de ánimo en función de las puntuaciones de estado de ánimo en la salida. El estado de ánimo más alto la partitura define el estado de ánimo de la canción.
E10	✓		El objetivo del estudio detectar automáticamente el estado de ánimo de una pieza musical lo cual implica asignar una categoría específica a clip de música Se menciona la construcción modelo para estimar automáticamente el estado de ánimo inferido en cada clip de música por lo cual se debe clasificar el estado de ánimo en diferentes categorías o clusters Se usa un clasificador difuso para reconocer el estado de ánimo de la música
E11	✓		No se hace mucho enfoque en el preprocesamiento del de los datos
E12	✓		Se usa el SVM para realizar la clasificación de características extraídas por el EEG en específico el DASM12 y se usa un enfoque de validación cruzada con la validación de 10 pasos (10-fold cross validation)
E13	✓		El enfoque propuesto se compara con métodos de aprendizaje poco profundos como máquinas de vectores de soporte (SVM) y bosques aleatorios

N°	Clasificación	Regresión	Notas
E14		✓	El modelo fue implementado usando WeKaDeeplearning4j y se llevaron bastantes experimentos con diferentes conjuntos de características y segmentaciones Estos vectores de características se utilizan posteriormente para entrenar y construir modelos de redes neuronales recurrentes que pueden predecir las emociones asociadas a cada segmento de música.
E15		✓	Se utilizó un enfoque de regresión los cuales se usan para estimar valores numéricos continuos en contraste a la clasificación que asigna etiquetas discretas a las instancias en este caso se busca predecir valores de valencia y arousal para representar la intensidad y valencia emocional que se asignan a los segmentos de música
E16		✓	Se tiene como objetivo mejorar la precisión del reconocimiento y clasificación de las emociones en la música Lo que se hizo para lograr esto se usa un método que combina una red de aprendizaje profundo con una red de atención dispersa explicita para el procesamiento de dataset Se usan técnicas de segmentación detallada y otros métodos de pre procesamiento para proporcionar un dataset de entrada de alta calidad para el modelo de clasificación

N°	Clasificación	Regresión	Notas
E17	✓		Se fusionan los resultados de salida de la red y el modelo y se obtienen los tipos de emociones mediante el clasificador Softmax Se utilizó el dataset SoundTrack
E18		✓	Se usa una capa completamente conectada maxout (MFC) para tener un enfoque de regresión de las dimensionales de valencia y arousal Este método se evalúa mediante el dataset MediaEval Emotion Music

A partir de la información recopilada se puede determinar que:

El enfoque de clasificación es de los más utilizados en los estudios primarios seleccionados siendo que 12 de los 18 estudios escogidos usan este enfoque mientras que 6 de los estudios primarios usan el enfoque de regresión. Cada estudio usa un dataset o dataset de prueba diferente para evaluar sus características extraídas y probarlas con su red o redes neuronales, los resultados se pueden en la tabla N°5 en conjunto de cual de cuál de los dos enfoques da mejor resultado para la detección de emociones en la música

2.3.4. **Pregunta N°4**

¿Cuál es la precisión y eficiencia de los enfoques de clasificación y regresión en la detección de emociones en la música?

Tabla 7. Resultados Pregunta 3

N°	Resultados de la investigación
E1	Se extraen 26 características las cuales se prueban con diferentes algoritmos de aprendizaje supervisado el mejor rendimiento con características de audio estándar del 44.3% (F-measure) pero si usa un enfoque multi-modal mejora a un 62% con una precisión del 81%
E2	Se hace un 10-fold-cross validation con una CNN el cual se compara con un SVM el promedio de la CNN VS SMV es de 0.724 contra 0.385 afirmando que el modelo es superar a otros modelos, pero a su vez afirmando que el modelo aún tiene muchas cosas por mejorar
E3	Al hacer uso de diferentes herramientas se obtienen resultados del 58.3% en arousal y 28.1% en valence y accuracy del 70%
E4	Para comprobar los datos se tienen 2558 canciones las cuales son introducidas a la red neuronal donde se intenta que la red sea capaz de mostrar buenos resultados se obtienen un accuracy de 0.97%, precisión del 0,98%, recall del 0,98% y F1-Score del 0,98 como máximo en 3 iteraciones
E5	El modelo denominado FRBI presenta los peores resultados con 62.13% comparado con otros modelos uni y multimodales, el modelo FRBE presenta resultados mejores con 86.20% de precisión en la etapa de validación y modelo unimodales presentan una precisión 91%
E6	No hay resultados ya que se trata como una encuesta de algoritmos
E7	Se calcula el rendimiento de la LSTM + DNN con varios datasets las cuales se obtienen al alimentar diferentes fuentes de información a las CNN los cuales son los log-mel filterbank y las MFCC El mejor resultado en accuracy fue de 87.09% a 91.93% sin aplicación de CFS
E8	No hay resultados ya que se trata como una encuesta de algoritmos
E9	Se obtuvo un rendimiento máximo al 94.44% en su mejor momento De las ocho emociones que se evaluaron la que tuvo una mejor eficiencia es calm con 17 canciones y 94.11% y la que tuvo menos eficiencia fue sad con 9 canciones y 77.78% de eficiencia

N°	Resultados de la investigación
E10	Se hace una evaluación con una base de datos compuesta de 800 clips de música con un accuracy promedio al 86.3% y un recall promedio al 84.1%
E11	La cinco emociones que se buscan identificar tienen una precisión mayor al 68% siendo felicidad con la menor precisión del 68% y tristeza con el 90% precisión en este modelo
E12	Se obtienen resultados bastante buenos ya que la precisión del algoritmo fue del 82% en la detección de emociones evaluadas La clasificación de las emociones de alegría, tristeza, ira y placer se obtienen resultados entre el 80% y 90% esto indica una capacidad considerable para distinguir entre diferentes estados emocionales en la música usando este algoritmo
E13	Los resultados muestran que el método de aprendizaje propuesto en conjunto con la estructura residual de un 1D CNN con el Inception-GRU da un porcentaje de precisión del 84% dando un rendimiento mayor en la clasificación y reconocimiento de emociones musicales
E14	Los resultados muestran un porcentaje de 85% el cual fue la efectividad del modelo el autor afirma que los resultados exceden los resultados obtenidos por algoritmos SVG con enfoque en regresión
E15	Se logró un RMSE de 0.2022 para arousal y 0.268 para valencia los cuales fueron los resultados con mejor reporte para este dataset
E16	Los resultados mostraron que la precisión de reconocimiento del método propuesto es del 0,71% con emociones felices en la música y de 0,69 para emociones tristes en la música lo que indica que tienen una buena capacidad de reconocimiento y clasificación de emociones musicales
E17	Se muestra con un dataset seleccionados que la precisión de la clasificación promedio es de 0.848 lo cual es mejor que otros métodos de comparación
E18	El documento a pesar que es de la misma persona que diseño “Music Emotion Classification Method Based on Deep Learning And improved Attention Mechanism “no muestra resultados, pero afirma que tiene un buen rendimiento

En base a la información recopilada se generan dos graficas:

La primera grafica se muestra en la figura 1, donde se presenta estudios primarios que se realizaron usando el enfoque de clasificación y que logran cumplir con el objetivo de la pregunta de investigación. En aproximadamente el 59% de los estudios (E1, E5, E10, E11, E12, E13, E17), se observó una eficiencia de reconocimiento de emociones del 90% y 80%, Seguido de una eficiencia del 100% y 90% el cual se encuentra el 25% de los estudios (E4, E7, E9). Por último, dos proyectos (E2, E6) tienen una eficiencia de reconocimiento de emociones del 80% y 70%. En total se tuvieron 12 estudios los cuales usaron el enfoque de clasificación de los 18 estudios primarios seleccionados

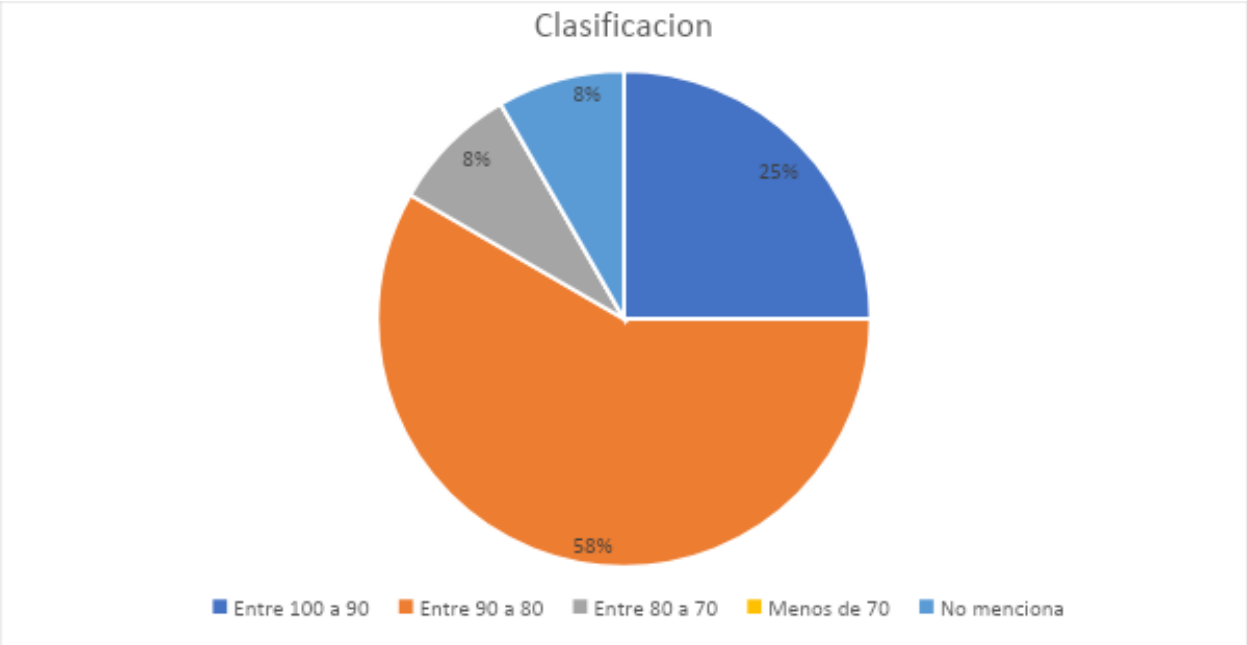


Figura 3. Estudios enfocados en Clasificación, Fuente: Propia

En la segunda tabla se presentan los estudios primarios que se realizaron usando el enfoque de clasificación y que logran cumplir con el objetivo de la pregunta de investigación. Los estudios E14 y E16 tienen una eficiencia de reconocimiento de emociones del 90% y 80% específicamente

85% (E14) y 80% (E16), el estudio E3 tiene una eficiencia del 70% y por último los estudios E6, E15 Y E18 no presentan un porcentaje o resultados de rendimiento de su investigación

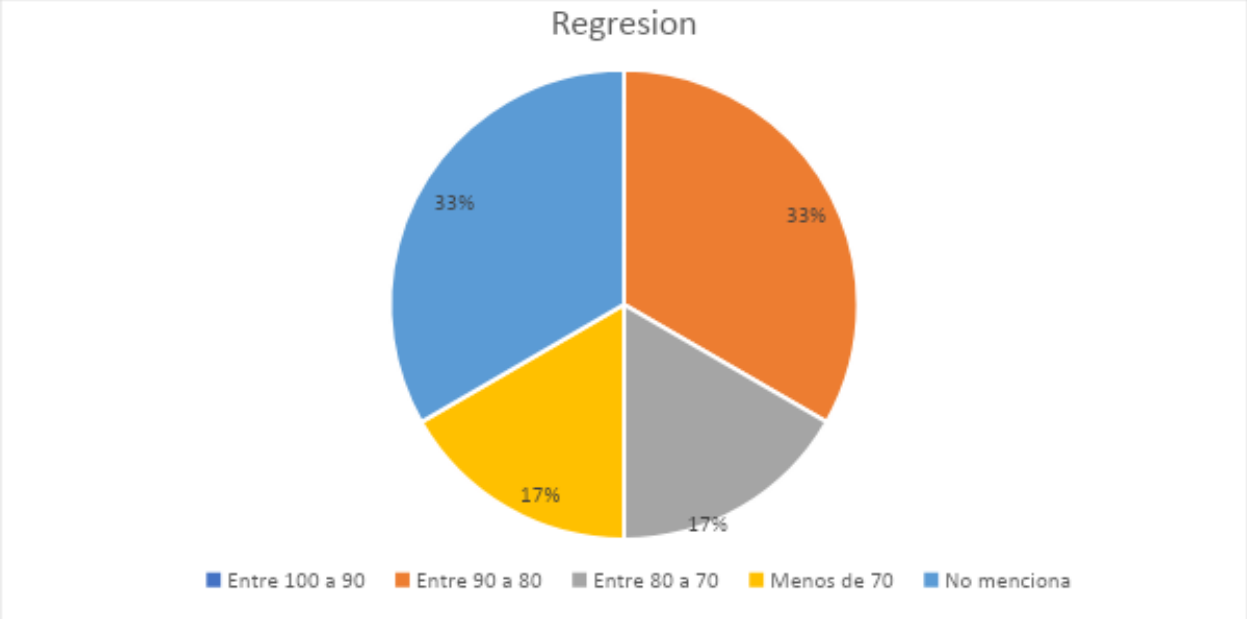


Figura 4. Estudios enfocados en Regresión, Fuente: Propia

A continuación, se presentan hallazgos relacionados con los enfoques de clasificación y regresión en la detección de emociones en la música, basados en los 18 estudios seleccionados como primarios en esta revisión de la literatura.

En primer lugar, se observa que los enfoques de clasificación son ampliamente utilizados en esta área de investigación. De los 18 estudios seleccionados, se encontró que 12 de ellos empleaban enfoques de clasificación como parte de su metodología. Esto nos da a entender que los métodos de clasificación son considerados más efectivos y aplicables en la detección de emociones en la música, además se encontró que los enfoques de clasificación demostraron ser más efectivos en comparación con los enfoques de regresión. Varios estudios muestran resultados cercanos al 100% de efectividad en la detección de emociones en la música. Por otro lado, los enfoques de regresión los cuales lograron una efectividad máxima del 85%. Aunque este valor no es malo,

pero sugiere que los enfoques de clasificación son más prometedores en términos de precisión y capacidad para identificar emociones en la música, pero es importante destacar que este análisis está sujeto a estudios seleccionados en esta investigación y puede existir otras investigaciones que logren mejores resultados pero que no fueron seleccionadas o incluidas para esta revisión.

2.4. CONCLUSIONES DE LA REVISIÓN

En esta revisión de la literatura se proponen cuatro preguntas de investigación relacionadas al campo del MER que tienen como objetivo indagar que tipos redes, enfoques, características y métricas son usadas en este campo de investigación diverso y dinámico. Los estudios revisados revelan que la clasificación es un enfoque más usado a comparación de la regresión, lo cual sugiere que asignar etiquetas emocionales discretas a las piezas musicales facilita la interpretación de los resultados con el uso de diferentes evaluaciones como la validación cruzada, precisión, recall, F1-Score, etc. Destaca la necesidad de una evaluación en cada estudio resaltando la importancia de una elección cuidadosa de los métodos de evaluación según el proyecto que se quiera desarrollar. Además, que los resultados varían según el estudio algunos alcanzando el 100% hasta otros con valores inferior al 50% de precisión esta variabilidad destaca los diferentes desafíos y tareas que se pueden presentar en el desarrollo del proyecto y resalta la importancia de una buena configuración de hiperparametros y un buen conjunto de datos para probar y evaluar el modelo. En general esta revisión de la literatura nos ha dado una buena base de información para desarrollo del proyecto la cual explica los diferentes desafíos que se pueden afrontar, las buenas prácticas que se pueden aplicar y los resultados que se pueden esperar según la red que se quiera utilizar.

3. CAPÍTULO 3 - APLICACIÓN DE LA METODOLOGIA CRISP-DM, PARA IDENTIFICACIÓN DE EMOCIONES EN LA MÚSICA

En esta sección se describe la metodología usada para este proyecto denominada CRISP-DM que es un acrónimo para Cross-Industry Standard Process for Data Mining el cual es un modelo estándar y abierto que describe el proceso típico de minera de datos [45] y el cual se aplicara para alcanzar los objetivos propuestos con anterioridad.

Esta metodología consta de seis fases la primera fase es el entendimiento del negocio se trata de entender los que se quiere lograr desde la perspectiva del negocio se colocan objetivos y limitaciones que deben ser equilibrados adecuadamente, la segunda fase corresponde al entendimiento de datos el cual requiere que se obtenga información que será útil para el desarrollo del proyecto, la tercera fase se denomina preparación de datos en la cual se decide con la información anteriormente recolectada en donde y para que se va a utilizar en el proyecto teniendo en cuenta la relevancia y la calidad de la información para cumplir con los objetivos de la minería de datos, la cuarta fase se denomina como modelado en la cual donde se tendrá que seleccionar específicamente la técnica modelo que sería utilizada para el desarrollo del proyecto, la quinta fase se denomina evaluación en el cual se evaluara que tan bien el modelo se desarrolla en un ambiente de operativo y si es capaz de cumplir los objetivos impuestos de la minera de datos por último la sexta fase se denomina despliegue en donde con el conocimiento obtenido en las fases anteriores se debe documentar y presentar los hallazgos obtenidos en las fases anteriores [45]

3.1. FASE 1 COMPRENSIÓN DEL NEGOCIO

3.1.1. EVALUACIÓN DEL CONTEXTO

La detección de emociones en la música constituye un campo de estudio que abarca diferentes áreas, las cuales buscan entender cómo ciertas composiciones musicales con el uso de acordes, cambios de ritmo y letras pueden influir profundamente en nuestro estado emocional, cognitivo y/o social a su vez pudiendo generar respuestas variadas y significativas en diferentes momentos y situaciones. Las diferentes investigaciones dirigidas en esta área se han beneficiado bastamente de los avances tecnológicos de análisis de datos y aprendizaje automático además la masiva recopilación de datos musicales combinada con técnicas avanzadas de machine learning como el aprendizaje profundo ha permitido explorar de manera más detallada y precisa la relación de música y emociones. Para afrontar este problema se emplean normalmente dos enfoques distintos los cuales son: el enfoque categórico el cual se centra en identificar patrones significativos en fragmentos de música en base a características musicales y técnicas ML los cuales asignan etiquetas emocionales a las fragmentos de canciones permitiendo una clasificación en base categorías predefinidas por emociones, por otro lado el enfoque de regresión se enfoca a la predicción de emociones por medio de la intensidad o grado de una emoción particular en un fragmento musical ofreciendo una comprensión más detallada de que tanto influencia una emocione en una pieza musical. Con el uso de técnicas ML y aprendizaje profundo en conjunto con una CNN se busca identificar la o las emociones evocadas que mejor representen un fragmento de música.

3.1.2. **OBJETIVOS DEL NEGOCIO**

Según lo mencionado previamente en el punto anterior se busca “Desarrollar y evaluar un modelo basado en técnicas ML y de aprendizaje profundo”. El modelo diseñado busca predecir emociones presentes en un fragmento de música, siguiendo un enfoque.

3.1.3. **RECURSOS**

Se cuentan con 2 frameworks para la extracción de características del dataset SoundTrack el primero es OpenSmile el cual es un software de código abierto que se usa para el análisis de audio y voz y se usa para extraer características espectrales, tono, timbre, etc. y librosa el cual al igual que OpenSmile se usa para analizar música y audio y proporciona una serie de herramienta y funciones para la extracción de características de alto y bajo nivel como el roll off o el MFCC. Se cuenta con un computador de mesa con una tarjeta gráfica GTX 1060, un procesador i3 7100 a 3.8Ghz con 8 GB de RAM para desarrollar el modelo, aunque también se puede usar ambientes de desarrollo como Google Collab o Jupyter para el desarrollo si la máquina anteriormente mencionada no es capaz de desarrollar el proyecto

3.1.4. **COSTOS Y BENEFICIOS**

Al realizar el proyecto se obtendrá un modelo ML el cual tomó como ejemplo trabajos como [8, 16, 40-41,43-45] que sea capaz de identificar y predecir las emociones presentes en fragmentos de música esto daría un mejor entendimiento sobre cómo la música afecta las emociones humanas y porque algunas canciones pueden ser muy influyentes para algunas personas

Tabla 8. Recursos Proyecto

Items	**Recursos**		**Total**
	Director	**Estudiantes**	
Recursos humanos director/a del proyecto/ co-directora	2500000	15360000	17860000
Hardware: Equipos	0	2800000	2800000
Software: Python, windows 10, Office 2021, Google Colab, Visual Studio Code	0	232130	232130
Renta de informes, libros, artículos	0	25000	25000

Items	Recursos		Total
	Director	Estudiantes	
Salidas o viajes	0	0	0
Documentación	0	0	0
Publicaciones artículos científicos	0	0	0
Otros	0	0	0
Total			

3.1.5. **RIESGOS Y CONTINGENCIAS**

A continuación, se presenta la tabla 8 donde se detallan los posibles riesgos y la propuesta de contingencias que pueden presentarse durante el proyecto.

Tabla 9. Riesgos y contingencias

Riesgos	Descripción	Contingencia
Costo computacional y recursos disponibles	Al ser un proyecto ML pueden requerir un costo computacionales considerables ya que se necesita probar y entrenar el modelo bastantes veces para obtener los mejores resultados posibles al no tener una máquina capaz de manejar esto puede producir un entrenamiento lento y prolongado	Se pueden usar ambientes de desarrollo como Google Colab o Jupyter para desarrollar el modelo si la máquina no cumple con las expectativas de rendimiento, estos ambientes de desarrollo no toman recursos locales así que el tiempo que tomará para el entrenamiento dependerá del dataset
Sobreajuste del modelo	Un sobreajuste significa que el modelo se ajusta demasiado a los datos de entrenamiento y no predice correctamente nuevos datos que entran al modelo esto puede causar un problema de rendimiento cuando se enfrenta a datos	Se pueden aplicar diferentes técnicas ML conocidas como la regularización de datos, capas como el dropout, validación cruzada, regularización, etc.

	reales	
Calidad de los datos	Al ser un dataset hecho por un tercero se corre el riesgo que los datos puedan estar desbalanceados, la duración de los fragmentos no sea la misma, etc. lo cual pueda hacer que el modelo aprenda patrones incorrectos	Hacer una revisión de los datos en el preprocesado si el dataset está balanceado especialmente las etiquetas emocionales de cada fragmento, hacer técnicas de segmentado, data augmentation y revisión si no hay datos faltantes que puedan afectar el rendimiento final del modelo
Complejidad del modelo	Las arquitecturas para modelos del REM pueden ser complejas y difíciles de interpretar esto puede hacer complicado de entender como el modelo toma decisiones y puede dificultar cómo se puede corregir un error	Examinar diferentes trabajos que utilicen redes iguales o similares y probar los parámetros y capas con el dataset, la prueba y error es un punto de partida para evaluar cómo se comporta el modelo, cómo toma decisiones y como identifica patrones

3.1.6. **TERMINOLOGÍA**

Los siguientes conceptos son usados en distintos proyectos [1-8, 16, 40-41,43-45] son comúnmente usados para referirse a métodos o redes usados para la detección de emociones en la música los cuales son:

- Reconocimiento de emociones en la música (MER): Es el proceso de predecir emociones humanas de señales de audio usando técnicas de inteligencia artificial el cual tiene un amplio rango de áreas que se pueden aplicar como la psicología, medicina, educación y entretenimiento
- Red Neuronal Convolucional (CNN): Es un tipo de red neuronal que se utiliza comúnmente en tareas de análisis de imágenes y reconocimiento de patrones. Esta se diseña para procesar datos en forma de matrices o tensores lo que la hace ideal para

analizar imágenes, videos y en este caso audio también. Este tipo de red se basa en capas convolucionales, capas pooling y capas totalmente conectadas que permite aprender patrones locales y globales de datos de entrada. (Saha, 2018)

- Conjuntos de data entrenamiento (train), validación (validation), pruebas (test): El conjunto de entrenamiento son todos los datos que el modelo ve y aprende en su proceso de entrenamiento, el conjunto de validación es usado para evaluar el modelo pero esta es una evaluación frecuente normalmente se usan estos datos para mejorar y actualizar los híper parámetros del modelo así que el modelo ocasionalmente ve estos datos pero nunca aprende de ellos así que este conjunto afecta el modelo pero indirectamente a diferencia del conjunto de entrenamiento, el conjunto de prueba provee piezas de información con las que se puede evaluar el modelo una vez es completado [33]
- Preprocesado de datos: Son técnicas de la preparación de datos que describe cualquier tipo de procesamiento realizado “raw data” para ser preparada o usada en otro proceso, este proceso de transformación convierte los datos en un formato que es más fácil y efectivo para procesos de minado de datos, ML u otras tareas de ciencia de datos, estas técnicas son normalmente usadas en las primeras etapas de desarrollo para asegurar resultados más precisos. [34]
- Datos originales (raw, source, atomic o primary data): Son datos que no han sido procesados para el uso la cual requiere una selectiva extracción, organización y algunas veces análisis y darle formato para la presentación o alimentación de un modelo ML algunas veces estos datos terminan en bases de datos el cual deja los datos más accesibles para un mejor procesado y análisis [35]
- Validación cruzada: Es un método para evaluar el rendimiento de un modelo ML al dividir el conjunto de datos de entrenamiento y prueba en n-segmentos de manera iterativa

- Verdadero positivo - VP: Es un resultado el cual el modelo predice correctamente la clase positiva [36]
- Verdadero negativo - VN: Es un resultado el cual el modelo predice correctamente la clase a la cual no pertenece el resultado [36]
- Falso positivo - FP: Es un resultado el cual modelo predice de manera incorrecta la clase el cual el resultado pertenece el resultado [36]
- Falso negativo - FN: Es un resultado el cual el modelo predice de manera incorrecta la clase a la cual no pertenece el resultado [36]

3.1.7. OBJETIVOS DE LA MINERÍA DE DATOS

Especificar y desarrollar un modelo de detección de emociones en la música mediante técnicas ML y de aprendizaje profundo, utilizando una CNN que sea capaz de analizar fragmentos musicales del dataset SoundTrack con el propósito de igualar o superar los resultados previamente obtenidos en investigaciones anteriores [8, 40-41] que emplearon este dataset.

3.1.8. CRITERIOS DE ÉXITO DE LA MINERÍA DE DATOS

Con la red y dataset seleccionados se debe obtener una precisión igual o superior al 75% para alcanzar este objetivo se hará uso de métricas como precisión, recall y F1-Score para evaluar el rendimiento del modelo después del entrenamiento y observar como de desempeña cada emoción y también se hará uso de estadísticas como la matriz de confesión para medir la tasa de aciertos y errores cometidos en la predicción de clases.

3.1.9. PLAN DEL PROYECTO

- Fase 1: Comprensión del negocio: En esta fase se busca comprender los objetivos del proyecto lo cual implica identificar las metas y requerimientos del proyecto en el contexto de la detección de emociones en la música se definen los objetivos a alcanzar tanto en el proyecto como en la minería de datos y se establece un plan de proyecto y como la detección de emociones puede generar un valor
- Fase 2 Comprensión de datos: En esta fase se exploran los datos que vienen disponible en el dataset seleccionado. Se analizan las variantes que ofrece el autor del dataset, las características de los datos como etiquetas emocionales, estructura de las canciones y la calidad de los mismos. El objetivo de esta fase es comprender la naturaleza y la calidad de los datos disponibles para determinar qué información es relevante para el desarrollo del proyecto.

- Fase 3: Preparación de los datos: En esta fase se realizan tareas de limpieza, transformación de datos, se hace un preprocesado de datos musicales para eliminar datos incompletos o irrelevantes para el modelo, además se hace una tarea importante la cual es seleccionar y extraer las características de cada fragmento de música para el desarrollo, entrenamiento y prueba del modelo

- Fase 4: Modelado: En esta fase se empieza a construir el modelo en base a la metodología CRISP-DM se hacen uso de técnicas ML y de aprendizaje profundo con una red neuronal convolucional - CNN se entrena el modelo con los datos preprocesados y se ajustan los híper parámetros para obtener el mejor resultado posible del entrenamiento.

- Fase 5 Evaluación: En esta fase se evalúa la efectividad y rendimiento del modelo utilizando métricas como precisión, F1-SCORE, recall, accurancy, etc. para medir el desempeño del modelo y comparar los resultados obtenidos con trabajos previos para validar la efectividad del modelo
- Fase 6 Despliegue: En esta fase se realiza el plan de despliegue de la solución para de forma terminar con el modelo y la metodología CRISP-DM

3.2. **FASE 2 ENTENDIMIENTO DE LOS DATOS:**

3.2.1. REPORTE DE BÚSQUEDA INICIAL Y SELECCIÓN DE DATOS

El objetivo principal de este proyecto es crear un modelo ML desde cero y lograr que pueda identificar emociones en la música para lograr esto se debía tener una base de datos musical etiquetada la cual servirá como entrada para alimentar el modelo, se realizó una búsqueda exhaustiva en diferentes artículos científicos y repositorios para descubrir cuáles eran los dataset más utilizados en este tipo de proyectos, según la investigación que se realizó se descargaron algunos dataset mencionados en los artículos como SoundTrack, AMG 1608, DEAM, emotify, etc. Para seleccionar el dataset se uso como base los artículos escogidos en la revisión de la literatura y se examino cual o cuales datasets eran los mas usados en este tipo de problema, uno de los mas utilizados era el dataset SoundTrack360 ya que era un dataset pequeño de 360 canciones, etiquetado con puntuaciones hechas por expertos y balanceado ya que todas las etiquetas emociones tenían el mismo número de canciones además que comprendía diferentes géneros usados en películas como música clásica, jazz, rock, electrónica, etc. El cual esta disponibles en formato csv a través del sitio web https://osf.io/p6vkg/.

Para poder usar estos datos se tiene que acceder al sitio web de OSF y descargarlos de forma normal los datos están abiertos al público y no hay que registrarse o autenticarse con algún

correo para acceder a los datos se encuentran 2 datasets denominados mean_ratings_set1.csv y mean_ratings_set 2.csv. El set1 consta de 360 canciones con etiquetas emociones como felicidad, enojo, ira, tristeza, etc. y una etiqueta emocional para cada canción además que este set sirvió de base para hacer el set2 el cual consta con menos canciones en total 110 pero añade 2 etiquetas emocionales al del set1 como son belleza y liking. En este proyecto se usó el set1 que a pesar de que puede ser menos preciso que el set2 hay algunos artículos científicos cómo [3, 8, 40] que han dado buenos resultados con este dataset.

3.2.2. ESTRUCTURA Y CONTENIDO DE LOS DATOS

El dataset SoundTrack en su versión 1, tiene un total de 360 fragmentos musicales, con una duración variable que oscila entre 15 segundos a 1 minuto. Este dataset se distingue por presentar valoraciones emocionales que abarcan doce distintas categorías como son: felicidad, ira, tristeza, ternura, tensión, miedo, energía y valencia. Estas valoraciones se derivan de la evaluación realizada por un grupo de individuos que califican cada canción en una escala de 1 a 9.

valence	energy	tension	anger	fear	happy	sad	tender	TARGET
5.37	4.99	3.32	1.06	1.03	5.5	1.36	2.81	Happy
5.04	2.55	2.95	1.08	1.4	1.47	5.46	3.87	Sad
5.93	2.89	2.17	1	1.06	3.11	2.82	5.69	Tender
2.58	4.32	5.76	3.34	5.288	1.02	2	1.03	Fear
2.81	5.02	5.87	5.38	4.7	1.03	1.79	1.15	Anger
3.1	4.46	5.24	2.319	3.18	1.78	1.91	1.28	Surprise
5.69	4.2	2.84	1.04	1.12	4.25	2.2	3.96	High Valence
1.8	4.62	6.24	4.3	5.87	1.017	1.253	1.01	Low Valence
4.14	5.67	4.59	1.83	2.29	3.45	1.52	1.33	High Energy
4.8	2.3	3.03	1.1	1.54	1.45	4.72	3.708	Low Energy
2.57	4.92	5.97	4.16	5.45	1.04	1.48	1.02	High Tension
5.45	2.88	2.48	1.01	1.28	2.54	3.37	5.05	Low Tension

Figura 5. Ejemplo de las etiquetas en el dataset original, Fuente: Propia

En este rango el número 1 representa una emoción apenas perceptible o ausente en el fragmento musical mientras que el 9 denota una emoción fuertemente presente o altamente perceptible en el fragmento musical. Esta evaluación pretende etiquetar emocionalmente cada canción, sirviendo como una posible referencia emocional asociada a la misma. Claramente no todas las canciones obtienen una puntuación perfecta en cada una de las emociones mencionadas. Por ende, la valoración emocional con la mayor puntuación es la emoción que más se percibe en el fragmento musical. En este dataset compuesto por 360 canciones, se establece una distribución equilibrada mediante una división de doce grupos, los cuales cada grupo está compuesto por 30 canciones, cada una con su propia etiqueta emocional. Esta estratificación asegura un equilibrio en la representación de las distintas emociones a lo largo del dataset.

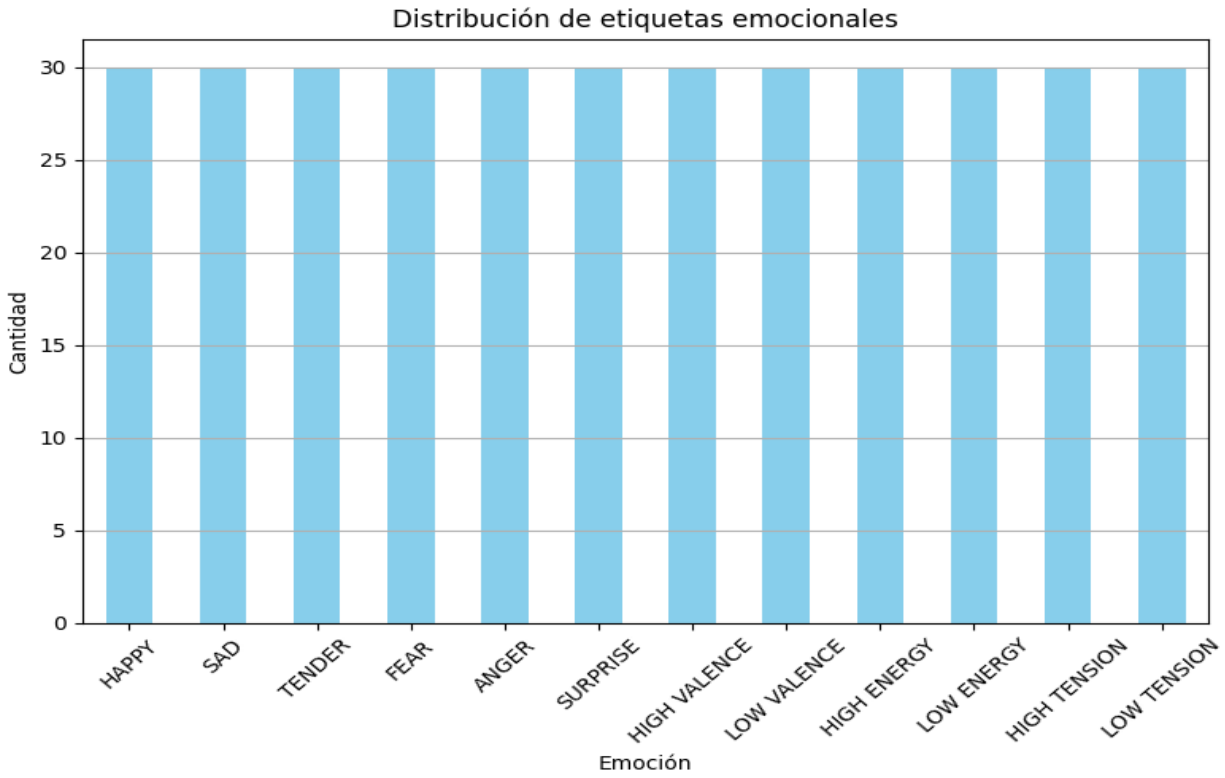


Figura 6. Distribución de etiquetas del dataset Soundtracks, Fuente: Propia

Etiquetas especiales: Al explorar el dataset que contiene las etiquetas emocionales se observa que existe una columna denominada “target” que tiene 12 etiquetas distintas las cuales 6 pueden ser consideradas etiquetas específicas como son happy, sad, tender, fear, anger y surprise. Las otras 6 etiquetas pueden ser consideradas como etiquetas especiales ya que abarcan diferente rango de emociones que pueden estar relacionados al modelo de Russell. Por ejemplo, las etiquetas high valence son etiquetas asociadas a la música o sonidos que pueden generar emociones alegres o positivas [22] como felicidad, alegría, euforia, amor [27]. En contraste las etiquetas low valence son sonidos o música que pueden generar sentimientos tristes o negativos como tristeza, melancolía, desesperación. Asimismo, las etiquetas como high energy representa música o sonidos que pueden ser ruidosos y/o rápidos [22] como es el caso del metal o heavy metal y se relacionan emociones como la euforia, excitación, entusiasmo, etc. Por otra parte, las etiquetas low energy se relaciona a sonidos suaves o muy lentos como la música clásica y se relacionan emociones como la calma, serenidad, tranquilidad. finalmente tenemos las etiquetas high y low tension que especifican el grado de anticipación o tensión en la música, las cuales contribuyen a la atmósfera emocional que la pieza quiere transmitir

3.3. **FASE 3 PREPARACIÓN DE LOS DATOS**

3.3.1. **CONSTRUCCIÓN, LIMPIEZA Y TRANSFORMACIÓN DE DATOS**

Dado que los datos que provee el dataset SoundTrack puede parecer pequeño para abarcar un problema como es el de detectar emociones en fragmentos de música se emplearon diversas técnicas de enriquecimiento y transformación de datos las cuales contribuyen para la expansión

y mejora del dataset la cual proporciona una base más sólida para el entrenamiento del modelo, lo primero que se hizo fue verificar la duración que tenía cada fragmento de música ya que el modelo trabaja mejor con fragmentos de duración fija por lo que si se usan fragmentos de duración variada se puede tener el riesgo de perder información importante de los fragmentos de datos con menor duración a comparación a fragmentos con el doble o triple de duración lo cual podría generar inconsistencias al momento del procesamiento, análisis y evaluación de datos dando lugar a posibles problemas de rendimiento del modelo. Para solucionar esto se optó en dividir todos los fragmentos de música en segmentos de 15 segundos cada uno si un fragmento era por ejemplo de 25 segundos se dividía en dos segmentos de 15 segundos donde el primer fragmento tendría los primeros 15 segundos de música del fragmento original y el segundo tenía los 10 segundos restantes más 5 segundos donde no se escucharía nada ya que se rellenará con ceros [44] como se ve en la figura 7

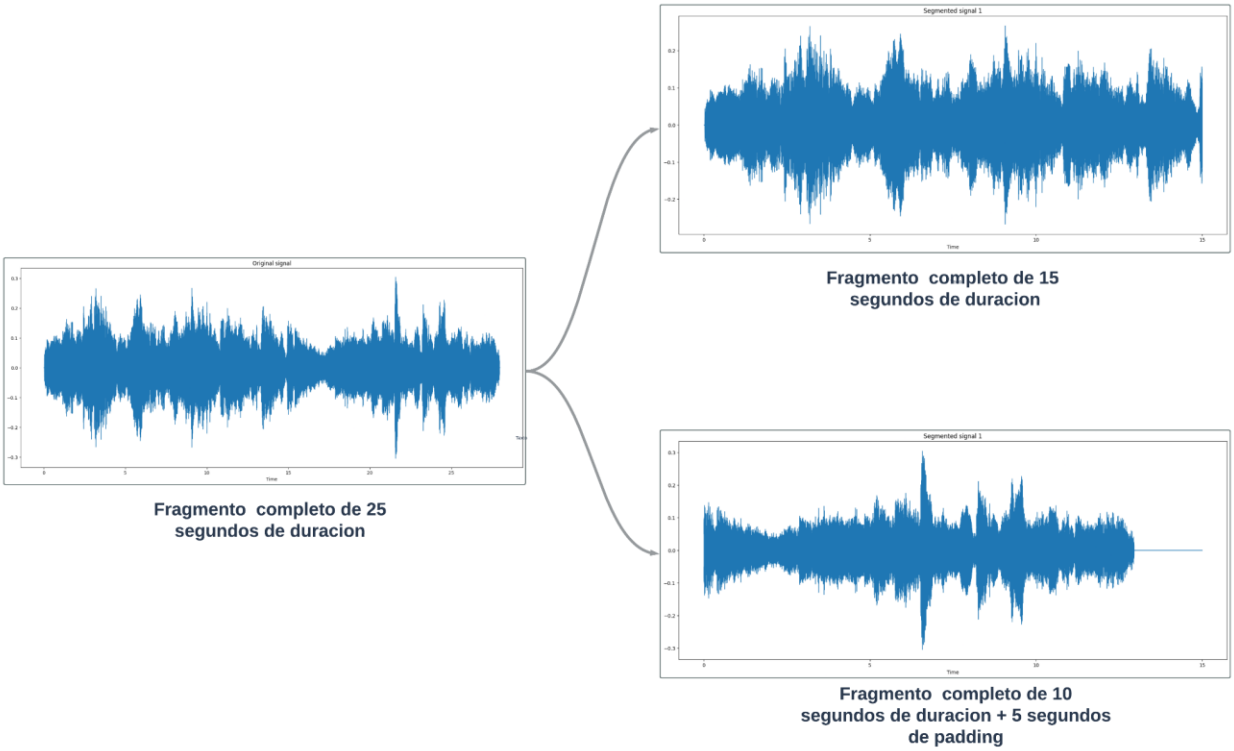


Figura 7. Construcción de fragmento, Fuente: Propia

Se obtiene un total de 633 fragmentos de música con segmentación, el siguiente paso después que cada fragmento de música se segmenta y tuvieran una duración fija se dividió el nuevo dataset de la misma forma que se divide un dataset en prueba, validación y test. Se optó por dividir el dataset en un 70% para test y 30% para validación y se usó la técnica de aumento de datos o data augmentation que consiste en generar datos artificiales usando transformaciones, manipulaciones o replicaciones para aumentar la cantidad y diversidad del dataset original pero se utilizó solo en el dataset de entrenamiento por diferentes razones como primero evitar el sobre ajustamiento del modelo, segundo para que el modelo cuando este en la fase de entrenamiento tenga diferentes datos de los cuales pueda aprender y por último para que el modelo aprenda y sea capaz de patrones más complejos dicho esto se aplican cinco tipos de transformaciones a cada fragmento de audio. El primero es el desplazamiento de tiempo el cual según [42] la idea es cambiar la posición del audio de izquierda a derecha aleatoriamente si se cambia a la izquierda los primeros segundos del fragmento tendrán un sonido en blanco ósea no se escucharía nada lo mismo sucedería si se hace el mismo procedimiento a la derecha solo que no se escucharía nada al final del fragmento

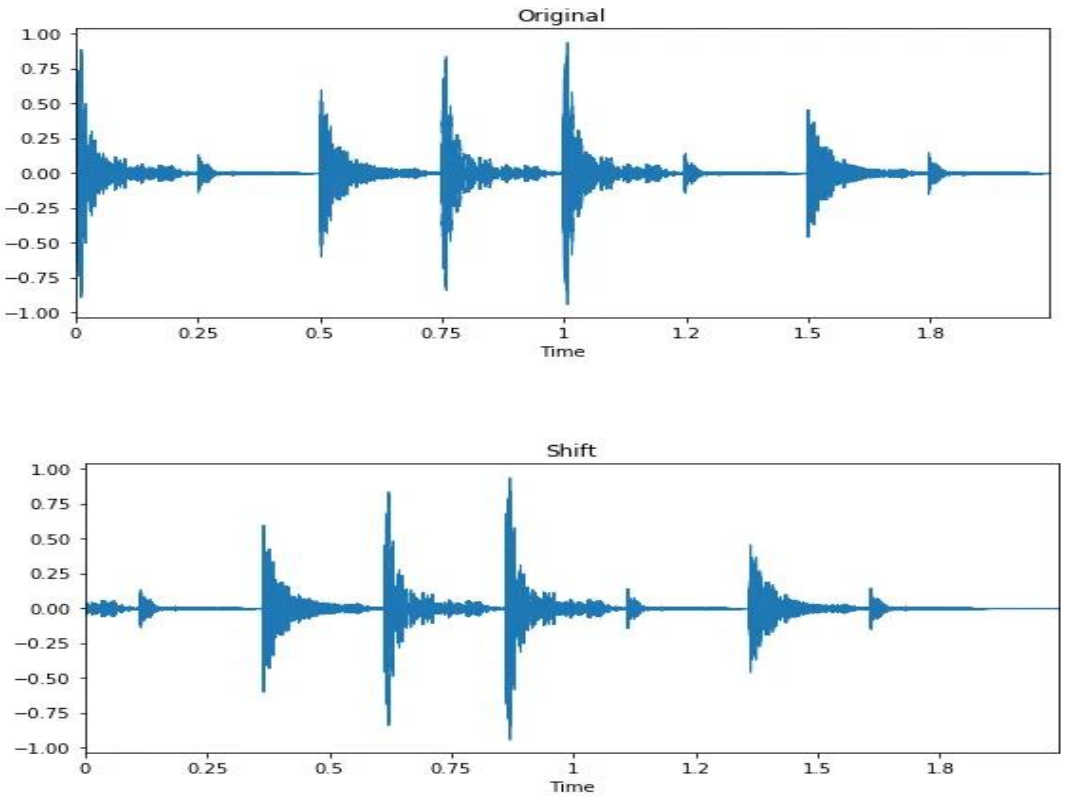


Figura 8. Desplazamiento del tiempo, Fuente: Propia

El segundo procedimiento consistió en incorporar ruido a cada fragmento lo cual implica introducir una señal aleatoria no deseada que puede distorsionar la señal de audio original [42] como se ilustra en la figura 9

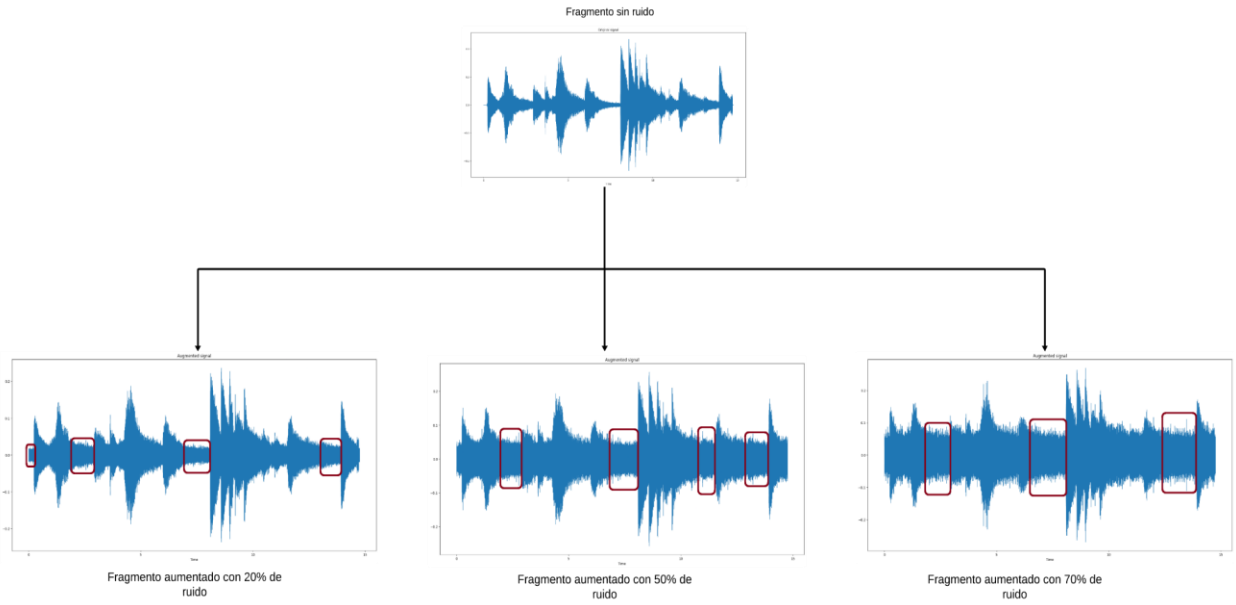


Figura 9. Incorporación del ruido, Fuente: Propia

cuando se aplica ruido a un fragmento de audio se añade un sonido de estática a lo largo de todo el fragmento, como muestra en los recuadros resaltados en rojos de la figura 9 se puede observar cómo partes del fragmento original con menor intensidad experimentan un cambio perceptual al aplicar niveles de ruido del 20%, 50% y 70% haciendo que estas secciones adquieran mayor amplitud a comparación de la forma original, el tercero es el estiramiento de tiempo el cual se utilizó la librería librosa.effects el cual consiste en alterar el paso percibido de tiempo de una señal de audio sin modificar aspectos como el tono original del audio o el tempo pero modificando la duración del fragmento ya sea alargando y haciendo que el fragmento suene lento si se usa un valor muy bajo o disminuyendo la duración y haciendo que suene rápido si se usa un valor alto [36].

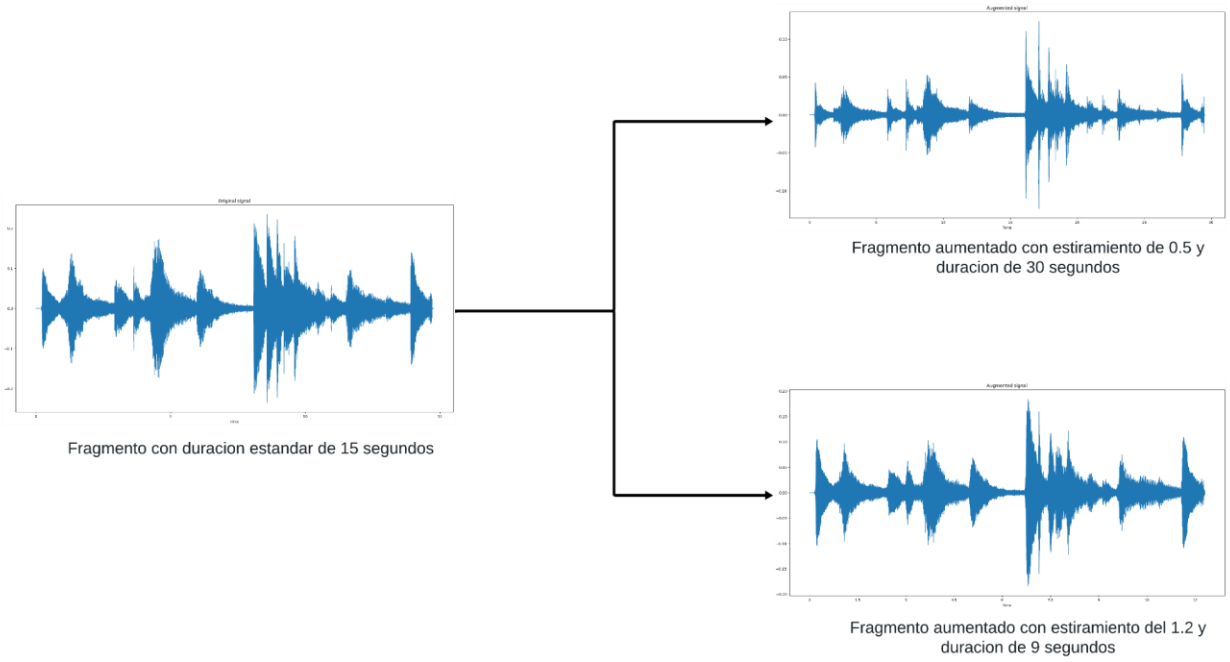


Figura 10. Estiramiento del fragmento, Fuente: Propia

Como se ve en la figura 10 si aplicamos un estiramiento muy bajo entre 0.1 y 0.6 el fragmento va a aumentar su duración haciendo que en comparación a el fragmento original este se vea estirado y suene lento. Si usan valores altos en este caso 1.2 se ve como algunas partes se ven contraídas a comparación del fragmento original, usar esta técnica puede ser un poco complicado ya que los fragmentos deben ser con una duración igual o menor a 15 segundos así que usar estos valores bajos puede complicar la tarea de segmentación que se hizo con anterioridad por lo que en este caso se usó estiramientos con valores de 1.1 a 1.2 para tener fragmentos entre 11 a 9 segundos. El cuarto fue el escalamiento de tono el cual al igual que la transformación anterior se utiliza la librería librosa.effectts el cual consiste en cambiar el tono del fragmento sin cambiar la velocidad del sonido original del audio. Por último, se usó la ganancia aleatoria el cual consiste en cambiar la amplitud del audio moviéndola de arriba a abajo.

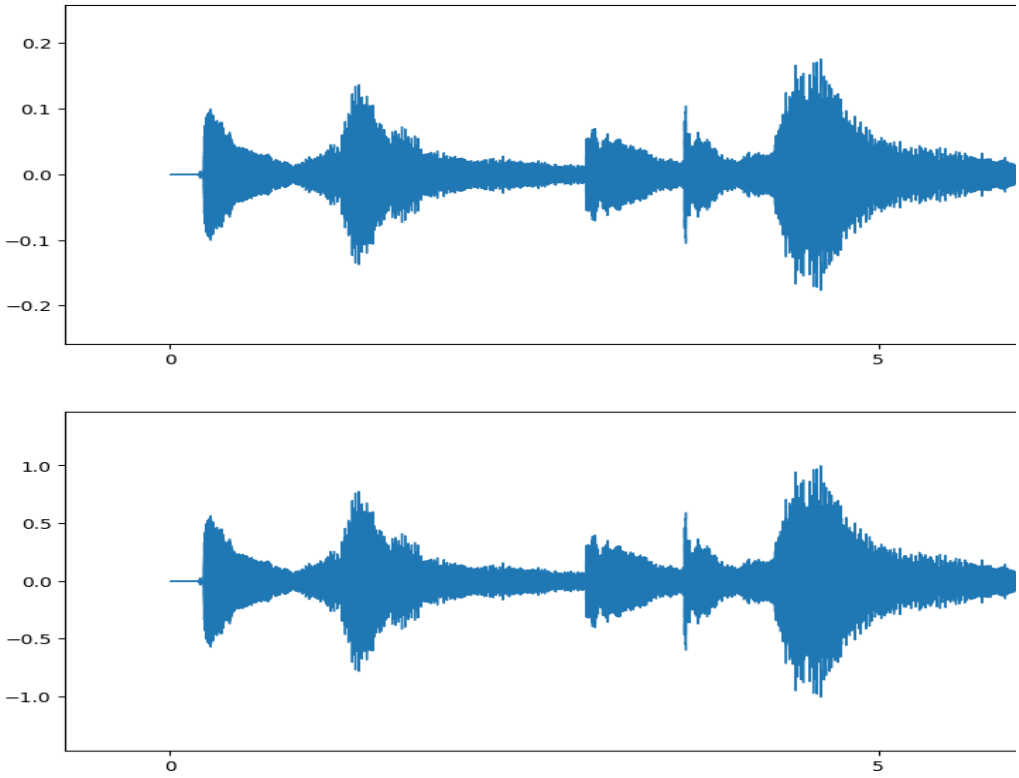


Figura 11. Ganancia aleatoria a un fragmento, Fuente: Propia

Como se ve en la figura 11 se aplica la ganancia aleatoria a un fragmento de audio a simple vista no se ve ningún cambio en el gráfico ya que tenemos la misma duración y forma que el gráfico original pero si nos fijamos en el eje Y veremos que hay un aumento en la amplitud del audio pasando de -0.2 a 0.2 a -1.0 a 1.0 haciendo que nuestro audio tenga más volumen o suene más fuerte que el original ya que se usó una ganancia mínima de 3 y máxima de 12 para formar este audio aumentado si usáramos ganancias más bajas o incluso negativas haríamos que el audio perdiera amplitud y a su vez que perdiera más volumen que el original.

Al hacer estas transformaciones a cada fragmento al conjunto de entrenamiento anteriormente dividido se aumenta considerablemente la cantidad de elementos del dataset dando un total de 3165 fragmentos de música para el conjunto de entrenamiento. Después se debe extraer

características de cada dataset con fragmentos de audio tanto segmentado, transformado y original como se sabe hay tres grupos principales de características en la música como son las características de alto, medio y bajo nivel cada nivel tiene ciertas características que son útiles para la detección de emociones pero cómo podemos saber cuáles son las más importantes con anterioridad se hizo una revisión de la literatura la cual uno de sus objetivos específicos era investigar cuál o cuáles eran las características más utilizadas en el MER donde se obtuvieron un total de 54 características con la librería pandas y librosa de Python se extrajeron estas características de cada fragmento, cuando la extracción de características era completada se añadía a un archivo CSV denominado features_train_aug.csv para el conjunto de entrenamiento y features_val.csv para el conjunto de validación. Por último en el dataset original se utilizaban puntuaciones dadas por sujetos de prueba y etiquetas emocionales de cada canción estas son necesarias para el entrenamiento del modelo pero al tener datos separados, aumentados y segmentados ya no se podía utilizar de la forma que venían originalmente en el dataset así que lo que se hizo en este caso fue utilizar como referencia el nombre de cada canción en el dataset original y comparar tanto con los datos de validación y datos de entrenamiento se comparan por medio del nombre de la canción si el nombre coincidía se tomaba la celda completa del csv del dataset original y se agregaba a un nuevo csv que contiene los datos de los conjuntos de datos de validación y entrenamiento por separado denominados target_data_test.csv para entrenamiento y target_data_val.csv para validación

3.4. **FASE 4 MODELADO**

3.4.1. **SELECCIÓN DE TÉCNICA DE MODELADO**

La elección de una red para detectar emociones en la música es compleja ya que implica determinar la red más adecuada para trabajar con el dataset previamente procesado por lo que

se lleva a cabo una revisión de la literatura donde se investigan las diferentes redes neuronales más utilizadas y que han tenido éxito en proyectos MER. La revisión abarca diversos enfoques y técnicas que incluyen redes RNN [6,9,15,16,19], SVM [2,3,4,7,13], combinaciones como redes CDLNN [8, 9,18], k-NN [2], CRNN [9], ANN [7,10,11] incluso proyectos donde experimentan con espectrogramas. Pero la elección más común en proyectos del MER y las cuales tienen mayor éxito son las redes CNN [3,4,5,9,14,16,17,18,19] ya que son atractivas debido a la eficacia en el análisis de señales secuenciales como el procesamiento de audio y música, además que es capaz de aprender características jerárquicas automáticamente lo que la hace capaz de capturar patrones complejos y relevantes en el dataset y posterior preprocesado

3.4.2. UNIÓN DE CONJUNTO DATOS

Previamente se realizó una división inicial del dataset con el propósito de aplicar la técnica de aumento de datos al conjunto de entrenamiento. Con esto se generó cuatro conjuntos de datos dos para el conjunto de entrenamiento el cual consiste de características extraídas de fragmentos de música y etiquetas emocionales y dos para el conjunto de validación que contiene característica y etiqueta de datos sin ninguna técnica, Esta fase implica primero la reintegración de los conjuntos de datos que contienen características extraídas de los conjuntos de entrenamiento y validación la cual se denomina con la variable “X_features” para después hacer lo mismo con los conjuntos de etiquetas emociones la cual se denomina con la variable “Y_features” de esta forma tener dos conjuntos de datos para después unir estos conjuntos y tener un solo dataset denominado “Data” para posteriormente llevar a cabo una nueva división de datos la cual garantiza la presencia de datos segmentados y aumentados tanto en los conjuntos de entrenamiento como en los de validación para cuando se lleve a cabo el proceso de entrenamiento del modelo.

3.4.3. PREPROCESAMIENTO DE DATOS

Para que los datos fueran aptos para el modelo primero se tuvo que seguir una serie de pasos ordenados para que el modelo pudiera aceptar y trabajar con los datos con los que se quería entrenar el modelo

3.4.4. ETIQUETAS ONE-HOT O ETIQUETAS DUMMY

La codificación one-hot es un método común para lidiar con datos categóricos en ML que debe ser usado en el pre-procesado de datos ya que los modelos ML requieren entradas de variables numéricas este enfoque crea una nueva columna por cada valor único en la columna de la categoría original donde el uno representa en este caso que la etiqueta o categoría emocional está presente y el cero representa que la etiqueta o categoría emocional está ausente [28]. Para aplicar este método se usan la librería de Pandas con dos métodos el primero es DataFrame el cual en este caso se denomina como df el cual toma como parámetro la columna “TARGET” del dataset denominado anteriormente como “Data” de esta forma solo se toman las etiquetas emocionales después se utiliza el método get_dummies denominado como df_encode para volver todas las etiquetas del DataFrame df en etiquetas one-hot dando así un total de 12 columnas cada una representando una etiqueta emocional con 1620 filas cada una representando un fragmento de música

Tabla 10. Etiqueta emocional

TARGET - ETIQUETA EMOCIONAL
HAPPY
SAD
TENDER

FEAR
ANGER
SURPRISE

Tabla 11. Etiqueta target procesada

HAPPY	SAD	TENDER	FEAR	ANGER	SURPRISE
1	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	1

Una vez obtenido las etiquetas one-hot se procede a eliminar la columna “TARGET” del dataset “Data” anteriormente usadas para crear estas etiquetas

3.4.5. DIVISIÓN EN CONJUNTOS DE ENTRENAMIENTO Y VALIDACIÓN

En este paso se divide el dataset obtenido previamente en dos subconjuntos más pequeños denominados entrenamiento (X_train) y validación (X_val). Este tipo de división es necesaria ya que el conjunto de entrenamiento se utilizará para entrenar el modelo y el conjunto de validación se usará después que el entrenamiento haya finalizado para ver que tan bien el modelo responde a datos que no haya visto o estudiado cabe resaltar que el modelo puede presentar problemas de sobre ajustamiento ya que puede adaptarse o rendir de manera excelente con los datos de entrenamiento pero cuando ve o intenta analizar datos nuevos le cuesta predecir o evaluar estos datos [30, 33] este problema se puede resolver con diferentes métodos como regularización o

capas dropout en el modelo utilizando otros métodos como la normalización o estandarización de los datos. Para este modelo se usó una división común del 80% del dataset para el conjunto de entrenamiento el cual se encargará de entrenar, ajustar pesos y parámetros del modelo y el 20% del dataset es para el conjunto de validación el cual se utiliza para ajustar hiperparametros del modelo, evaluar su rendimiento mientras se hace el entrenamiento del mismo y se utiliza para evaluar el rendimiento final del modelo después de pasar por las etapas de entrenamiento, este conjunto puede dar una evaluación independiente ya que el modelo nunca estudia estos datos durante el proceso de entrenamiento. Para hacer la división en el dataset actual primero con anterioridad se dividió en dos variables llamadas “Data” y “df_encode” donde en “Data” se colocan todas las características extraídas de las canciones y la valoración emocional dado por los datasets Y_train_music y Y_val_music excluyendo la columna “TARGET” ya está contiene la etiqueta emocional. En “df_encode” se colocan las etiquetas one-hot que se crearon con anterioridad usando las etiquetas emocionales de la columna “TARGET”. La división de datos se hace de la siguiente forma usando la librería train_test_split usando las variables “Data” y “df_encode” anteriormente creadas y con un test_size de 0.2 se da 80% de los datos de “Data” y “df_encode” a las variables X_train y Y_train y 20% a las variables denominas X_val y Y_val.

3.4.6. **OPTIMIZADOR**

Los optimizadores son algoritmos que ajustan los parámetros del modelo mientras se lleva a cabo el proceso de entrenamiento para minimizar la función de pérdida y mejorar la precisión [31]. Existen diferentes optimizadores para problemas de ML como el SGD, RMSprop y Adam cada uno con sus ventajas y desventajas para este problema se escoge usar el optimizador Adam ya que es un optimizador bien conocido por adaptar características de los optimizadores RMSprop y SGD lo cual resulta en una convergencia de datos más rápida y eficiente en comparación a los dos métodos anteriormente mencionados, además que este optimizador es

capaz de ajustar su tasa de aprendizaje de forma automática por cada peso de red individualmente a diferencia del SGD que tiene una tasa de aprendizaje constante a lo largo del entrenamiento [31]

3.4.7. **FUNCIÓN DE COSTE**

Al ser este un problema de multi-clase con uso de etiquetas one-hot la función de pérdida elegida es la categorical_crossentropy ya que es comúnmente escogida para problemas multi clase y se desea medir la discrepancia entre las distribuciones de probabilidad que genera el modelo y las etiquetas reales en este caso las etiquetas one-hot

3.4.8. **MÉTRICAS**

Para el proyecto se utilizan diferentes métricas para medir el rendimiento del modelo durante y después del entrenamiento, Durante el entrenamiento del modelo hace uso de dos métricas muy comunes para problemas de ML como es el accuracy o precisión ya que es fácil de entender, implementar y es una métrica que permite evaluar el rendimiento del modelo de manera simple la fórmula para esta función es sencilla ya que es el número de predicciones correctas hechas por el modelo sobre el número total de predicciones

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

Figura 12. Formula de la precisión correctos vs toral de clases**, [12]**

O también como el total de verdaderos positivos más verdaderos negativos sobre la suma total de verdaderos positivos (TP) verdaderos negativos (TN) falsos positivos (FP) y falsos negativos (FN) esta fórmula es más utilizada cuando se está generando la matriz de confusión

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

Figura 13. Formula de la precisión, Matrix de confusión, **[12]**

También se hace uso del lost o pérdida el cual mide qué tan bien el modelo está realizando predicciones en comparación a las etiquetas reales el objetivo en este caso es mantener esta métrica tan baja como sea posible y al ser un problema categórico usando una función de coste de categorical_crossentropy se obtiene la siguiente fórmula

3.4.9. **CONSTRUCCIÓN DEL MODELO**

Al ser un problema de ML se debe crear una red neuronal desde cero para el problema que se debe afrontar se tomaron como ejemplo ciertas redes descritas en trabajos previos [8,16,20,40-41,43,45] se eligieron y se experimentó con diferentes hiperparametros a lo largo del desarrollo buscando el mejor resultado posible:

Capa de entrada: Es la primera capa del modelo y su función es recibir los datos de entrada y pasarlos al resto de la red para su procesamiento, para poder usar correctamente esta capa se necesita saber el número de neuronas de entrada esto viene dado por número de total de características que tiene el dataset se toma como entrada el X_train con el método shape en la posición 1 ya que nos da el número de características exactas que necesita el modelo en este

caso serían 62 donde 53 vienen dadas por la extracción de características de los fragmentos de música y 9 vienen del dataset original sin contar la columna TARGET ya que es la etiqueta emocional.

Capas convolucionales Conv1D: Son una parte importante en el procesamiento de las características extraídas de los fragmentos de música. Estas capas aplicar filtros a lo largo de la secuencia temporal de las características el cual permite identificar patrones relevantes para detectar emoción/emociones en cada canción [1].

En diferentes proyectos ML que usan técnicas DL que buscan detectar emociones en la música es muy común que se usen capas Conv2D o capas convolucionales de dos dimensiones con espectrogramas en formato de imagen ya que en este formato cada pixel tienen tiene coordenadas en dos dimensiones la cual es altura (height) y ancho (width) esto sirve como entrada para la red CNN de dos dimensiones y le será útil para que aprenda características locales y globales de la imagen mediante el uso de filtros pero al estar usando características extraídas de la música como características Chroma, Espectrales, etc. Hace que el uso la capa Conv2D sea imposible ya que estas son características unidimensionales o de una dimensión por lo que se tiene que usar capas que acepten datos unidimensiones como la Capa Conv1D la cual es capas de aprender patrones secuenciales y características a lo largo del tiempo de este tipo de datos usando filtros. El uso de filtros en secuencias temporales hace que la Conv1D se enfoque en encontrar elementos importantes que puedan ser representativos como tonos, ritmos, cambios de amplitud y otros elementos importantes que influyen en la expresión emocional de la canción. Para este modelo se usaron 2 de estas capas con filtros de 64 para el primera capa y 128 neuronas para la segunda, un batch_size de 3 y con activación ReLU se optó por usar un regularizador L2 en cada capa para evitar el overfitting al 0.2 pero los resultados mostraron que el rendimiento del modelo disminuye al paso de las épocas así que se decidió retirarlo y dejar las capas de la misma forma antes de la regularización

Capa MaxPooling1D: Para cada capa conv1D se usa una capa MaxPooling esto con el fin de reducir la dimensionalidad y extraer las características más importantes para su posterior uso en capas densas.

Capa Flatten: Esta capa se usa para aplanar la salida de las capas convolucionales en un vector unidimensional que pueda ser usada en las capas completamente conectadas o mejor conocidas como densas [23]

Capa dropout: se utilizan una capa por cada capa densa que se tiene en el modelo para regularizar la red neuronal y reducir/evitar el sobreajuste lo que hace esta capa es apagar aleatoriamente una fracción de neuronas durante el entrenamiento lo que ayuda un poco a la generalización del modelo [26]

Capa completamente conectada: Estas capas realizan operaciones de aprendizaje no lineal en la información que proviene de las capas anteriores [1] para este proyecto estas capas son responsables de combinar y procesar las características extraídas por las capas convolucionales y pooling para realizar la clasificación de las emociones/emociones pertenecientes a el fragmento de música. A lo largo del desarrollo del proyecto se experimentaron con diferente número de estas capas empezando con solo 1 como capa de salida con 12 neuronas representando el total de emociones que se tiene en el proyecto hasta llegando a usar 4 cada una con igual o diferente número de neuronas con el fin de obtener el mejor resultado posible al final se utilizaron 2 capas densas cada una con 512 neuronas las cuales cumplen con la tarea de procesar las características extraídas y aprender representaciones de la misma para que se pueda predecir la emoción que más se representa en la canción

Capa de salida: Es la última capa del modelo y su función es generar la predicción del modelo. la estructura de esta capa para este problema consta del número total de emociones a predecir que son 12 y un activador softmax el cual es indicado para problemas de clasificación multiclase

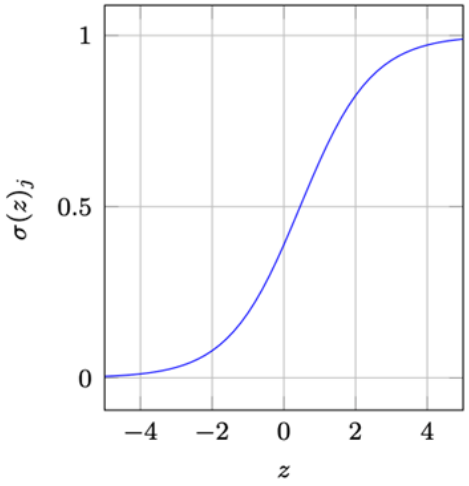


Figura 14. Formula de la precisión**,** [13]

3.5. **CONSTRUCCIÓN DEL MODELO EXPERIMENTACIÓN**

Al tratarse de un problema de aprendizaje profundo, la red neuronal convolucional con la que se experimentó paso por diferentes cambios significativos los cuales se hicieron con el fin de alcanzar el mejor resultado posible por lo que en esta sección se detallarán las diferentes modificaciones que se realizaron en la red neuronal abarcando específicamente la adaptación de hiperparametros como la incorporación o eliminación de ciertas capas ocultas

Ajustes de hiperparametros y capas: Durante esta fase se llevaron a cabo diferentes interacciones para tratar de optimizar el rendimiento del modelo lo cual llevó a la modificación de los hiperparametros que tenían las capas ocultas del modelo tales como el número de neuronas, regularizadores en ciertas capas, las tasas de aprendizaje del optimizador y el tamaño de lotes

además que se realizaron ajustes en la configuración de las capas ocultas con la inclusión o eliminación de capas convolucionales, densas, dropout y maxpooling

Evaluación del rendimiento del modelo: Para evaluar el rendimiento del modelo después del entrenamiento se utilizaron diferentes métricas las cuales permiten tener una visión de cómo se desempeñó el modelo. Se utilizaron gráficos para mostrar la precisión y perdida a lo largo de las épocas para el conjunto de entrenamiento y validación en el caso de la validación para determinar si se tenía una problema de sobre ajustamiento, se utilizó la matriz de confusión para analizar la precisión a detalle de cada etiqueta emocional, por último se utilizó las curvas ROC las cuales proporcionan información sobre la sensibilidad del modelo y que además se caracteriza por no verse afectado por un sesgo de clase, este gráfico representa la tasa de falsos positivos (FP) sobre la tasa de verdaderos positivos (VP)

Para este cada modelo se usó el dataset hecho con anterioridad con una división de 80% para entrenamiento y 20% para validación se busca el mejor resultado posible para el entrenamiento del modelo así que se hacen diferentes cambios tanto a los hiperparametros como a las capas del modelo.

Para todas las pruebas se hace uso de una activación reLu para las capas del modelo, un optimizador de tipo Adam y una pérdida de categorical cross-entropy por último se entrena el modelo en épocas de 50,100,150 y 300 cada una.

En cada experimento se prueba cómo va el modelo capas MaxPooling, dropout y con regularización L2 para evitar el sobreajuste

Modelo 1: La arquitectura del modelo es la siguiente

- Una capa de entrada con input_shape de 62 características
- Una capa convolucional oculta
- Dos capas MaxPooling
- Una capa flatten
- Dos capas completamente conectadas

Tabla 12. Resultados de entrenamiento modelo 1

N° Neuronas / Etapas	50	100	150	300
Entrada de 16 Neuronas, Convolucional oculta de 32 neuronas, completamente conectadas de 64 neuronas	Accuracy : 0.32 Loss : 1.68 Val_acc : 0.40 Val_loss : 1.55	Accuracy : 0.38 Loss : 1.57 Val_acc : 0.45 Val_loss : 1.44	Accuracy : 0.44 Loss : 1.42 Val_acc : 0.47 Val_loss : 1.34	Accuracy : 0.58 Loss : 1.08 Val_acc : 0.62 Val_loss : 1.09
Entrada de 32 Neuronas, Convolucional oculta de 64 neuronas, completamente conectadas de 128 neuronas	Accuracy : 0.42 Loss : 1.46 Val_acc : 0.47 Val_loss : 1.36	Accuracy : 0.54 Loss : 1.21 Val_acc : 0.54 Val_loss : 1.20	Accuracy : 0.62 Loss : 0.99 Val_acc : 0.63 Val_loss : 1.07	Accuracy : 0.82 Loss : 0.50 Val_acc : 0.68 Val_loss : 1.11
Entrada de 64 Neuronas, Convolucional oculta de 256 neuronas, completamente conectadas de 512 neuronas	Accuracy : 0.44 Loss : 1.67 Val_acc : 0.41 Val_loss : 1.65	Accuracy : 0.52 Loss : 1.51 Val_acc : 0.48 Val_loss : 1.55	Accuracy : 0.58 Loss : 1.35 Val_acc : 0.54 Val_loss : 1.48	Accuracy : 0.80 Loss : 0.98 Val_acc : 0.66 Val_loss : 1.37

Para este primer modelo se consiguen buenos resultados con las 3 configuraciones de neuronas como se ve en la figura 15 en la que se destaca es la segunda configuración ya que se obtiene

una precisión del 82% en el conjunto de entrenamiento y 68% en el conjunto de validación en la figura 16 podremos apreciar que en todas las configuraciones el conjunto de validación al principio de cada interacción tiende a tener más precisión que el conjunto de entrenamiento especialmente en la primera configuración este no es un síntoma malo pero es curioso ya que el modelo no está estudiando estos datos directamente.

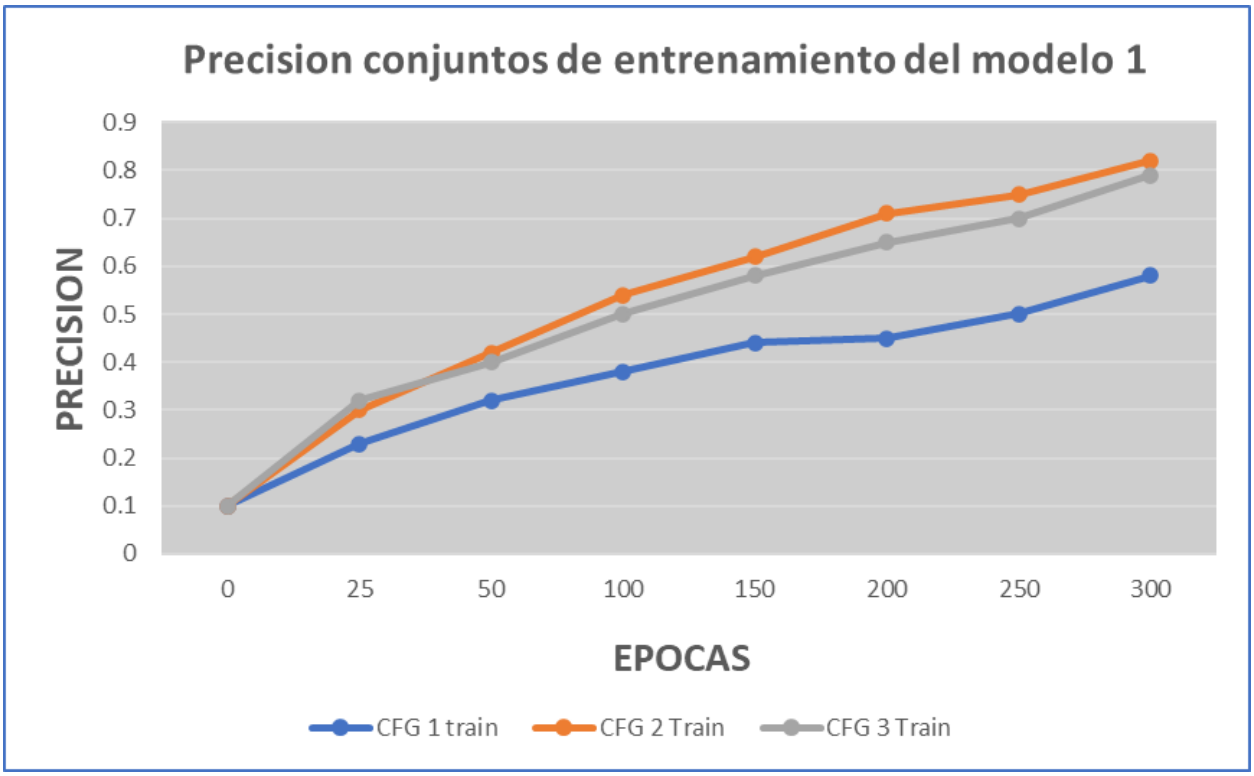


Figura 15. Rendimiento del modelo 1 en el conjunto de entrenamiento, Fuente: Propia

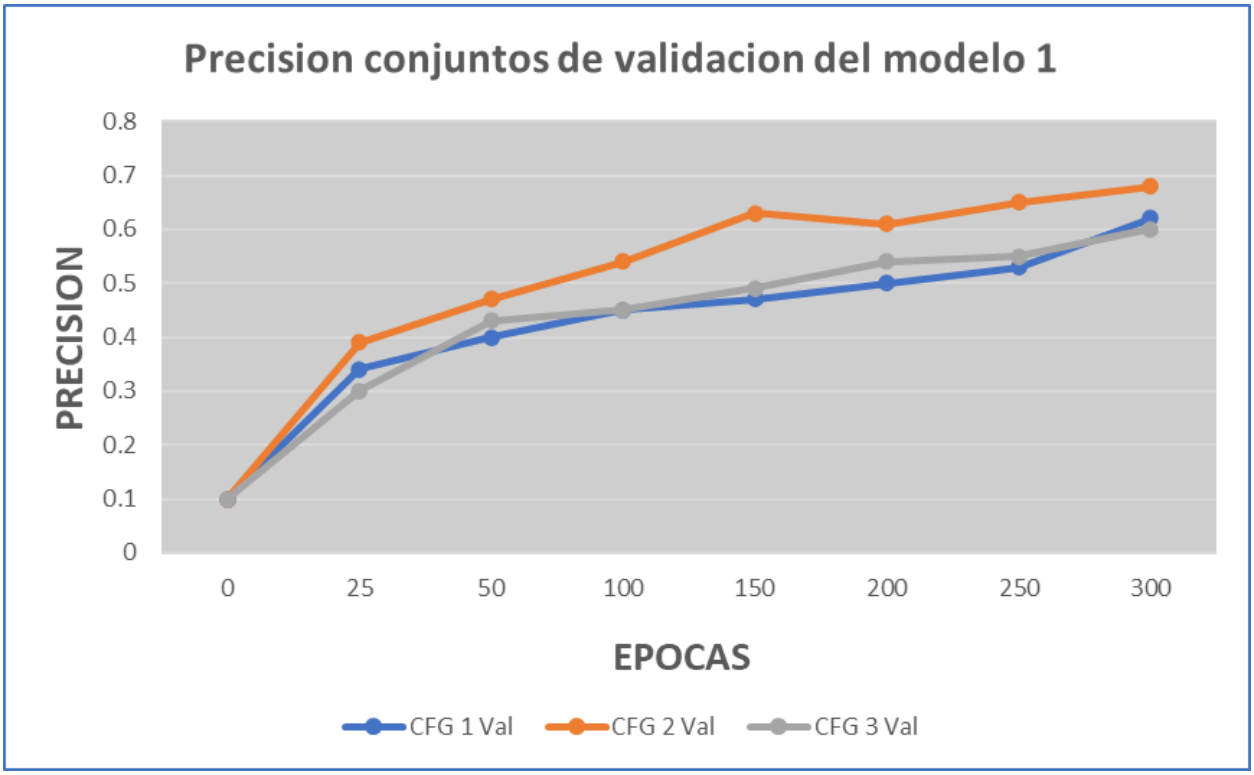


Figura 16. Rendimiento del modelo 1 en el conjunto de validacion, Fuente: Propia

Eventualmente el conjunto de entrenamiento sobrepasa en precisión al conjunto de validación en la primera configuración en la época 300 en la segunda 100 y en la tercera en la época 50 también podemos notar que en la segunda y tercera configuración el conjunto de validación se queda atrás y un poco estancado en la precisión con respecto al conjunto de entrenamiento mientras se entrena el modelo esto puede ser un síntoma de sobre ajustamiento ya que si observamos las pérdida de la tabla 11 observaremos que las pérdidas en el conjunto de validación no bajan de 1.00 mientras que la pérdida del conjunto de entrenamiento bajan hasta 0.50. Si detallamos las estadísticas obtenidas de la matriz de confusión (CM) y las etiquetas

emocionales en el entrenamiento de las tablas 12-13 podremos ver que tan bien el modelo se desempeñó en el entrenamiento

Tabla 13. Precisión obtenida en las 3 CFG del modelo 1

Emocion	Precision CFG 1	Precision CFG 2	Precision CFG 3
Happy	0.67	0.67	0.60
Sad	0.28	0.58	0.41
Tender	0.52	0.76	0.60
Fear	0.83	0.83	0.75
Anger	0.50	0.69	0.61
Surprise	0.38	0.65	0.58
High Valence	0.55	0.81	0.95
Low Valence	0.29	0.41	0.45
High Energy	0.82	0.86	0.82
Low Energy	0.68	0.68	0.56
High Tension	0.42	0.54	0.67
Low Tension	0.69	0.75	0.93

Si comparamos la precisión obtenida en las tres configuraciones de la tabla 12 de las etiquetas emocionales podremos observar que la configuración 2 tiene los mejores resultados ya que etiquetas como “Fear”, “High energy” y “High valence” empiezan con buenos resultados de precisión en la configuración 1 y tienen a mejorar o ser estables en la configuración 2 estas etiquetas pueden guardar una relación entre sí, ya que en películas del género del terror se buscan generar emociones fuertes (high valence) con la combinación de música que tenga

sonidos fuertes o intrigar al espectador (high energy) esta relación puede hacer que el modelo le sea más fácil de identificar estas etiquetas.

Por otra parte etiquetas como “sad” y “low valance” empiezan como etiquetas débiles o difíciles de identificar para el modelo en la segunda configuración se mejora su precisión pero siguen siendo las etiquetas más difíciles de identificar esto es curioso ya que estas etiquetas pueden tener una relacion ya que “sad” esta relacionada a canciones tristes o melancólicas mientras “low valence” esta relacionada a emociones que generan o sean relacionados a la tristeza

Tabla 14. Recall obtenida en las 3 CFG del modelo 1

Emocion	Recall CFG 1	Recall CFG 2	Recall CFG 3
Happy	0.81	0.68	0.69
Sad	0.87	0.95	0.92
Tender	0.45	0.60	0.62
Fear	0.55	0.57	0.56
Anger	0.43	0.72	0.76
Surprise	0.39	0.59	0.68
High Valence	0.43	0.64	0.52
Low Valence	0.57	0.83	0.73

High Energy	0.56	0.63	0.50
Low Energy	0.63	0.73	0.77
High Tension	0.52	0.89	0.80
Low Tension	0.61	0.70	0.69

En términos de recall varía según la configuración etiquetas como “sad” y “low energy” tienen buenos valores de recall en las tres configuraciones a pesar de que en la precisión era todo lo contrario. En la configuración 2 se encuentra el mejor recall siendo un 95% para la etiqueta “sad” lo cual indica que el modelo es consistente en identificar estas emociones por otra parte las etiquetas como “Surprise” y “Fear” muestran los recalls más bajos entre 57% a 59%

Tabla 15. F1-Score obtenido en las 3 CFG del modelo 1

Emocion	F1-Score CFG 1	F1-Score CFG 2	F1-Score CFG 3
Happy	0.73	0.67	0.64
Sad	0.42	0.72	0.57
Tender	0.48	0.67	0.61
Fear	0.66	0.65	0.65
Anger	0.46	0.70	0.68
Surprise	0.39	0.62	0.63
High Valence	0.48	0.72	0.67
Low Valence	0.38	0.55	0.56
High Energy	0.67	0.73	0.62

Low Energy	0.65	0.70	0.65
High Tension	0.46	0.68	0.73
Low Tension	0.65	0.72	0.79

En la tabla 14 en la que se detalla los F1-Score para las distintas configuraciones se evidencia que la CFG 2 genera resultados superiores oscilando entre 55% y 73% en diversas emociones. Este rango de F1-score da a entender que la CFG 2 logra un mejor equilibrio entre precisión y recall.

Los resultados combinados de precisión, recall y F1 nos dan una idea que la CFG 2 es la configuración más ideal para la tarea que se intenta desarrollar en términos de rendimiento y la cual se puede tomar como base en las siguientes arquitecturas para mejorar el rendimiento del modelo

Tabla 16. Resumen de los datos obtenidos en la matriz de confusión del modelo 1

Emocion	CM - VP CFG1	CM - FP CFG 1	CM - FN CFG 1	CM - VP CFG2	CM - FP CFG2	CM - FN CFG2	CM - VP CFG3	CM - FP CFG3	CM - FN CFG3	Etiquetas reales
Happy	22	11	5	22	11	10	20	13	9	33
Sad	8	21	1	17	12	1	12	17	1	29
Tender	13	12	16	19	6	14	15	10	9	25
Fear	24	5	20	23	6	18	22	7	14	29
Anger	13	13	17	18	8	7	16	10	5	26
Surprise	11	18	17	19	10	13	17	12	8	29

High Valence	12	10	16	18	4	10	21	1	19	22
Low Valence	7	17	5	10	14	2	11	13	4	24
High Energy	18	4	14	19	3	11	18	4	18	22
Low Energy	17	8	10	17	8	6	14	11	4	25
High Tension	13	18	12	17	14	2	21	10	5	31
Low Tension	20	9	13	22	7	10	27	2	12	29

La mejor precisión en general dada por la CM en este modelo fue de 69% en la CFG 2 según la tabla 15 esta CFG también da buenos resultados en la predicción correcta de etiquetas (VP) por lo que en algunas etiquetas las predicciones incorrectas (FP) son bajas como en la etiqueta “high energy” 3 FP,” high valance” 4 FP y “fear” 6 FP con esto podemos sacar unas conclusiones:

- El modelo en cualquier configuración predice mejores emociones que tienden a ser fuertes o positivas

- El modelo le cuesta predecir o aprender emociones que están relacionadas a emociones tristes o melancolicas

Modelo 2: Para el segundo modelo se usará como base la segunda configuración del modelo 1 ya que fue la mejor configuración (CFG) que dio en rendimiento general del modelo (precision y loss) y en métricas de evaluación ademas que se le agregara una nueva capa completamente conectada para evaluar el rendimiento de este modelo

- Capa de entrada convolucional con 64 neuronas iniciales input_shape de 62
- Capa oculta convolucional con 128 neuronas iniciales
- 2 capas MaxPooling
- Capa flatten
- 3 capas densas con 128 neuronas iniciales
- 3 capas dropout
- Regularizador L2 al 0.02 en las capas densas
- Capa de salida densa con 12 neuronas

Tabla 17.Resultados del entrenamiento del modelo 2

N° Neuronas / Etapas	50	100	150	300
Entrada y convolucional oculta de 64	Accuracy : 0.28 Loss : 1.49 Val_acc : 0.32	Accuracy : 0.38 Loss : 1.18 Val_acc : 0.42	Accuracy : 0.42 Loss : 0.88 Val_acc : 0.43	Accuracy : 0.75 Loss : 0.44 Val_acc : 0.62

Neuronas, completamente conectadas de 128 neuronas (CFG 1)	Val_loss : 1.48	Val_loss : 1.38	Val_loss : 1.26	Val_loss : 1.22
Entrada y convolucional oculta de 128 Neuronas, completamente conectadas de 256 neuronas (CFG 2)	Accuracy : 0.43 Loss : 1.42 Val_acc : 0.47 Val_loss : 1.33	Accuracy : 0.58 Loss : 1.07 Val_acc : 0.57 Val_loss : 1.08	Accuracy : 0.66 Loss : 0.91 Val_acc : 0.63 Val_loss : 1.02	Accuracy : 0.86 Loss : 0.38 Val_acc : 0.75 Val_loss : 0.94
Entrada de 256 Neuronas, completamente conectadas de 512 neuronas (CFG 3)	Accuracy : 0.39 Loss : 1.59 Val_acc : 0.43 Val_loss : 1.54	Accuracy : 0.51 Loss : 1.39 Val_acc : 0.53 Val_loss : 1.40	Accuracy : 0.62 Loss : 1.16 Val_acc : 0.63 Val_loss : 1.18	Accuracy : 0.75 Loss : 0.88 Val_acc : 0.71 Val_loss : 1.12

Para este modelo se percibe una mejora sutil a comparación a las configuraciones del modelo 1 en términos de precisión ya que el mejor resultado fue de 86% de precisión en el conjunto de entrenamiento con una pérdida del 0.38 y una mejora en el conjunto de validación también con 75% en precisión y 0.94 en pérdida esto da a entender que la modificación de hiperparametros aumentando las neuronas de las capas convolucionales y ocultas más la adicción de otra capa completamente conectada mejora la capacidad de aprendizaje del modelo con el conjunto de entrenamiento a pesar de que con el conjunto de validación no haya una mejora tan grande como se ve en las gráficas de la figura 17-18

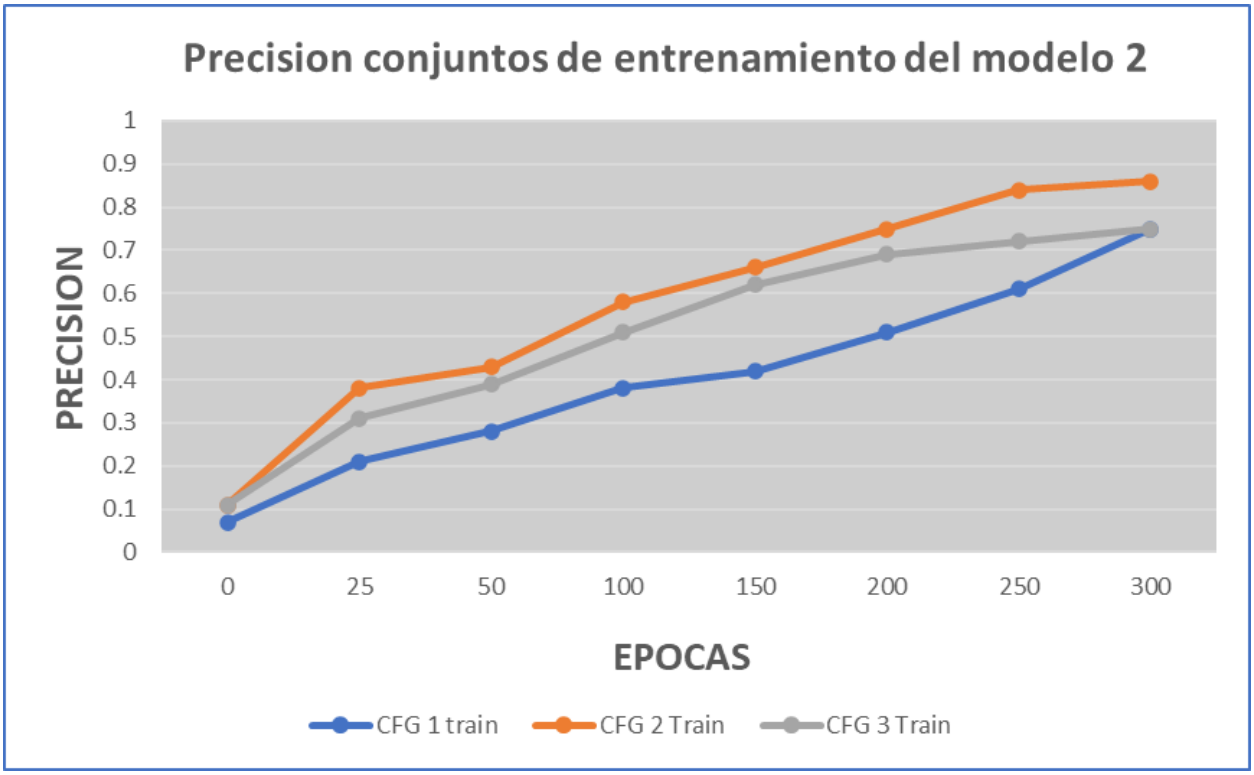


Figura 17. Rendimiento del modelo 2 en el conjunto de validación, Fuente: Propia

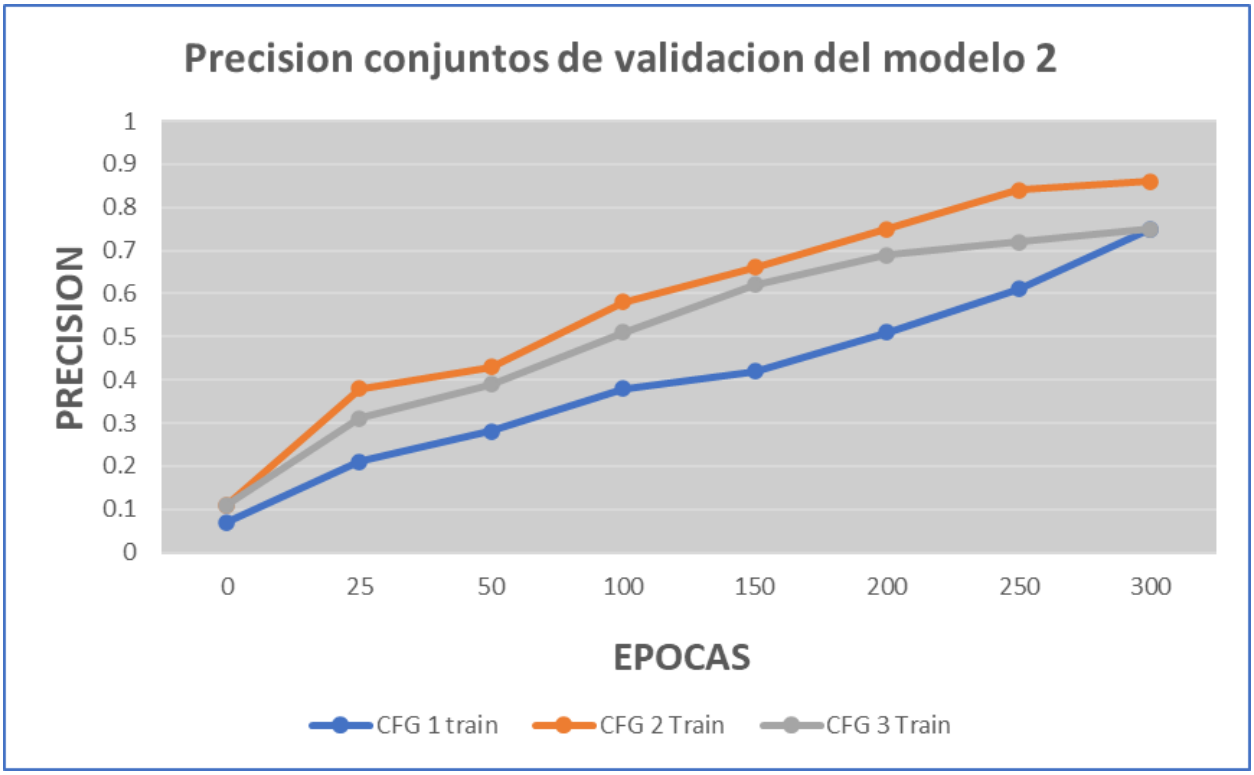


Figura 18. Rendimiento del modelo 2 en el conjunto de validación, Fuente: Propia

El conjunto de validación tiene el mismo comportamiento que en el modelo anterior ya que empieza teniendo una precisión superior al conjunto de entrenamiento, pero al paso de las épocas se ve superado y en algunos casos como en la CFG 2 se queda un poco estancado pudiendo ser un síntoma de sobre ajustamiento si se entrenara con más épocas el modelo, pero en todo caso esta configuración fue la mejor en este modelo incluso superando a las configuraciones del modelo anterior

Tabla 18. Precisión por emoción obtenidas durante el entrenamiento del modelo 2

Emocion	Precision CFG 1	Precision CFG 2	Precision CFG 3
Happy	0.70	0.76	0.54
Sad	0.75	0.82	0.65
Tender	0.84	0.76	0.56
Fear	0.90	0.90	0.89
Anger	0.92	0.88	0.96
Surprise	0.68	0.66	0.69
High Valence	1.00	0.95	0.86
Low Valence	0.50	0.50	0.54
High Energy	0.77	0.82	0.86
Low Energy	0.80	0.84	0.64
High Tension	0.64	0.58	0.61
Low Tension	0.79	0.75	0.72

La precisión en todas las CFG del modelo presentan una buena mejora con respecto al modelo anterior ninguna etiqueta baja del 50% de precisión lo cual quiere decir que el modelo aprende y predice de mejor una mejor manera que en interacciones pasadas

Si miramos de manera general la tabla 17 nos daremos cuenta que “anger” y “high valence” siguen siendo etiquetas emocionales fáciles de predecir para el modelo con precisiones bastante buenas entre 88% a 100%, etiquetas como “sad” que eran problemáticas para el modelo pasado en cualquier CFG pasan de tener un pobre desempeño 28% de precisión a 75% en la CFG y 82% en la CFG 2 pero otra etiqueta problemática la cual era “low valence” sigue teniendo precisiones bajas en comparación a las demás etiquetas

Tabla 19. Recall por emocion obtenidas durante el entrenamiento del modelo 2

Emocion	Recall CFG 1	Recall CFG 2	Recall CFG 3
Happy	0.82	0.89	0.86
Sad	0.88	0.80	0.76
Tender	0.78	0.76	0.61
Fear	0.74	0.68	0.59
Anger	0.69	0.79	0.67
Surprise	0.76	0.68	0.77
High Valence	0.63	0.63	0.54
Low Valence	0.75	0.75	0.68
High Energy	0.77	0.78	0.73
Low Energy	0.74	0.72	0.72
High Tension	0.90	0.90	1.00
Low Tension	0.88	0.88	0.77

En la tabla 18 podremos ver los recalls obtenidos en las CFG de este modelo si los comparamos con la tabla 13 tabla recall de la CFG 2 observaremos que ciertas etiquetas tuvieron una mejora como las etiquetas “happy”, “tender”, “fear” y “low tension” entre la CFG 1 y 2. Por otra parte etiquetas como “sad” y “low valence” sufren una caída en comparación a la CFG 2 del modelo 1 donde presentaban un buen rendimiento y en este pierden hasta un 10% de recall esto puede indicar que aunque subió la precisión se tuvo sacrificar un poco de la capacidad del modelo para predecir las instancias verdaderas (VP) en estas etiquetas

En general a pesar de las caídas en algunas etiquetas en el recall se tuvieron buenos resultados entre 59% a 100% entre las 3 CFG

Tabla 20. F1-Score por emoción obtenidas durante el entrenamiento del modelo 2

Emocion	F1-Score CFG 1	F1-Score CFG 2	F1-Score CFG 3
Happy	0.75	0.82	0.67
Sad	0.81	0.81	0.70
Tender	0.81	0.76	0.58
Fear	0.81	0.77	0.71
Anger	0.78	0.83	0.79
Surprise	0.72	0.67	0.72
High Valence	0.77	0.76	0.67
Low Valence	0.60	0.62	0.60
High Energy	0.77	0.80	0.79
Low Energy	0.77	0.77	0.68
High Tension	0.75	0.70	0.76
Low Tension	0.75	0.81	0.75

En la tabla 19 se pueden observar los F1-score de este modelo la CFG 2 fue la configuración más equilibrada en este caso mantiene un buen equilibrio entre precisión y recall en varias etiquetas emocionas lo cual indica un buen y constante rendimiento no obstante la CFG 1 también muestra un buen rendimiento en algunas etiquetas pero en otras como “happy” o “sad” le cuesta un poco más en predecir.

Tabla 21. Resumen de los datos obtenidos en la matriz de confusión del modelo 2

Emocion	CM - VP CFG1	CM - FP CFG 1	CM - FN CFG 1	CM - VP CFG2	CM - FP CFG2	CM - FN CFG2	CM - VP CFG3	CM - FP CFG3	CM - FN CFG3	Etiquetas reales
Happy	23	10	5	25	8	3	18	15	3	33
Sad	22	7	3	24	5	6	19	10	6	29
Tender	21	4	6	19	6	6	14	11	9	25
Fear	26	3	9	26	3	12	26	3	18	29
Anger	24	2	11	23	3	6	25	1	12	26
Surprise	20	9	6	19	10	8	20	9	6	29
High Valence	22	0	13	21	1	12	19	3	16	22
Low Valence	12	12	4	12	12	4	13	11	6	24
High Energy	17	5	5	18	4	5	19	3	7	22
Low Energy	20	5	7	21	4	8	16	9	6	25
High Tension	20	11	2	18	13	2	19	12	0	31
Low Tension	23	6	3	22	7	3	21	8	6	29

En la matriz de confusión se logra un desempeño del 75% de precisión en general en la CFG 2 en comparación al CFG 2 del modelo hay una mejora del 7% si comparamos los VP del modelo anterior con este notaremos que el modelo ha tenido una mejora en la mayoría de etiqueta a excepción de “low valence” obviamente al haber una mejora en la predicción del modelo hay menos FP y también hubo una buena reducción de los FP por parte del modelo lo cual se pudo ver en la [tabla modelo 2 recall] ya que el recall en general subió

En general esta interacción fue muy buena ya que el modelo mejoró tanto en rendimiento general como en las evaluaciones de rendimiento el único problema que se pudo notar fue un pequeño sobre ajustamiento

Modelo 3: Para este modelo se tomó como referencia la red utilizada en [8] usando cuatro capas convolucionales con los hiperparametros y se agregan tres capas completamente conectadas con los hiperparametros de la CFG 2 modelo 2

- Capa convolucional de entrada con 64 neuronas y 10 filtros
- Tres capas ocultas convolucionales con 128 neuronas y 3 filtros
- Tres capas MaxPooling
- Una capa flatten
- Tres capas ocultas completamente conectadas inicialmente con 256 neuronas
- Tres capas DropOut al 0.5

Tabla 22. Resultados generales del entrenamiento del modelo 3

N° Neuronas / Etapas	50	100	150	300
Entrada de 64 Neuronas con 10 filtros, tres capa oculta convolucional de 128 y tres capas completamente conectadas de 256 neuronas (CFG 1)	Accuracy : 0.27 Loss : 1.79 Val_acc : 0.30 Val_loss : 1.70	Accuracy : 0.39 Loss : 1.59 Val_acc : 0.44 Val_loss : 1.54	Accuracy : 0.47 Loss : 1.43 Val_acc : 0.43 Val_loss : 1.47	Accuracy : 0.72 Loss : 1.05 Val_acc : 0.58 Val_loss : 1.40
Entrada de 128 Neuronas con 10 filtros, tres capa oculta convolucional de 256, dos capas completamente conectadas de 256 neuronas y una capa completamente conectada de 512 neuronas (CFG 2)	Accuracy : 0.35 Loss : 1.67 Val_acc : 0.38 Val_loss : 1.62	Accuracy : 0.47 Loss : 1.51 Val_acc : 0.46 Val_loss : 1.50	Accuracy : 0.56 Loss : 1.32 Val_acc : 0.51 Val_loss : 1.43	Accuracy : 0.84 Loss : 0.77 Val_acc : 0.66 Val_loss : 1.40
Entrada y tres oculta de 256 Neuronas con 10 filtros,, tres capas completamente conectadas de 512 neuronas (CFG 3)	Accuracy : 0.36 Loss : 1.64 Val_acc : 0.41 Val_loss : 1.64	Accuracy : 0.52 Loss : 1.39 Val_acc : 0.44 Val_loss : 1.51	Accuracy : 0.69 Loss : 1.12 Val_acc : 0.60 Val_loss : 1.31	Accuracy : 0.91 Loss : 0.60 Val_acc : 0.71 Val_loss : 1.41

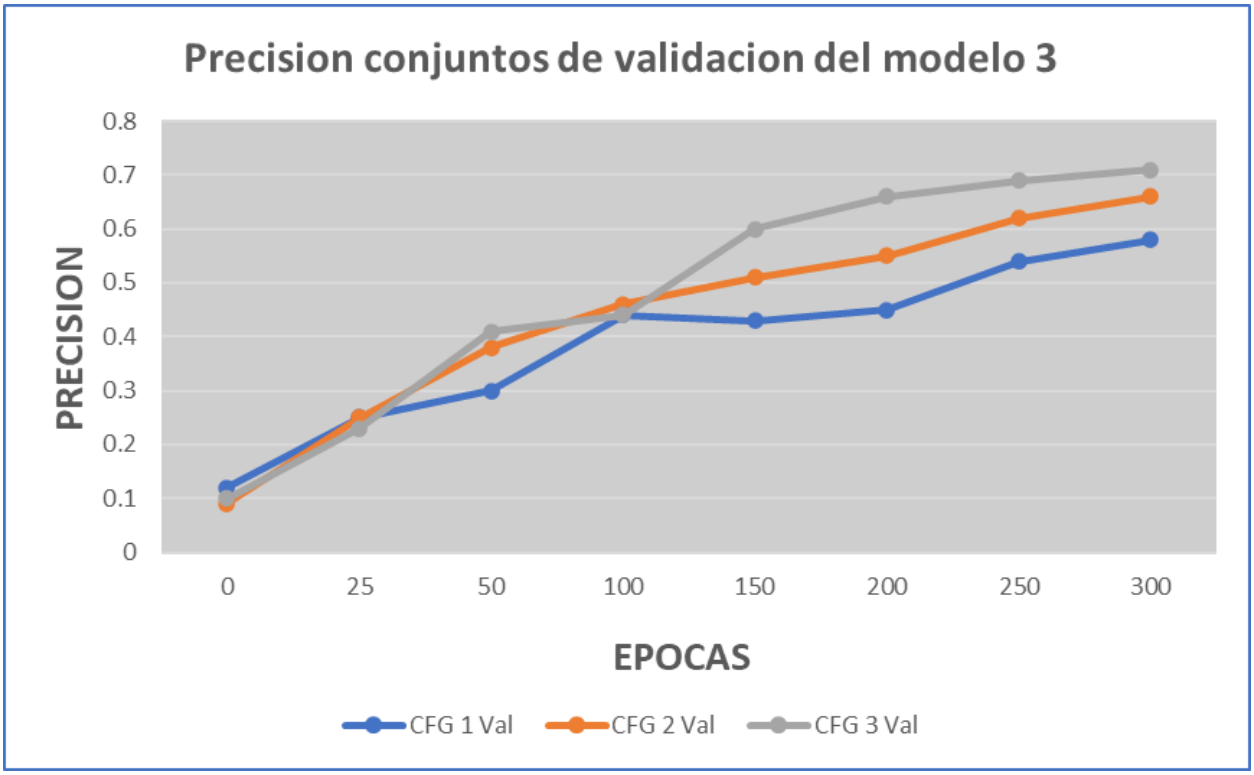


Figura 19. Rendimiento del modelo 3 en el conjunto de entrenamiento, Fuente: Propia

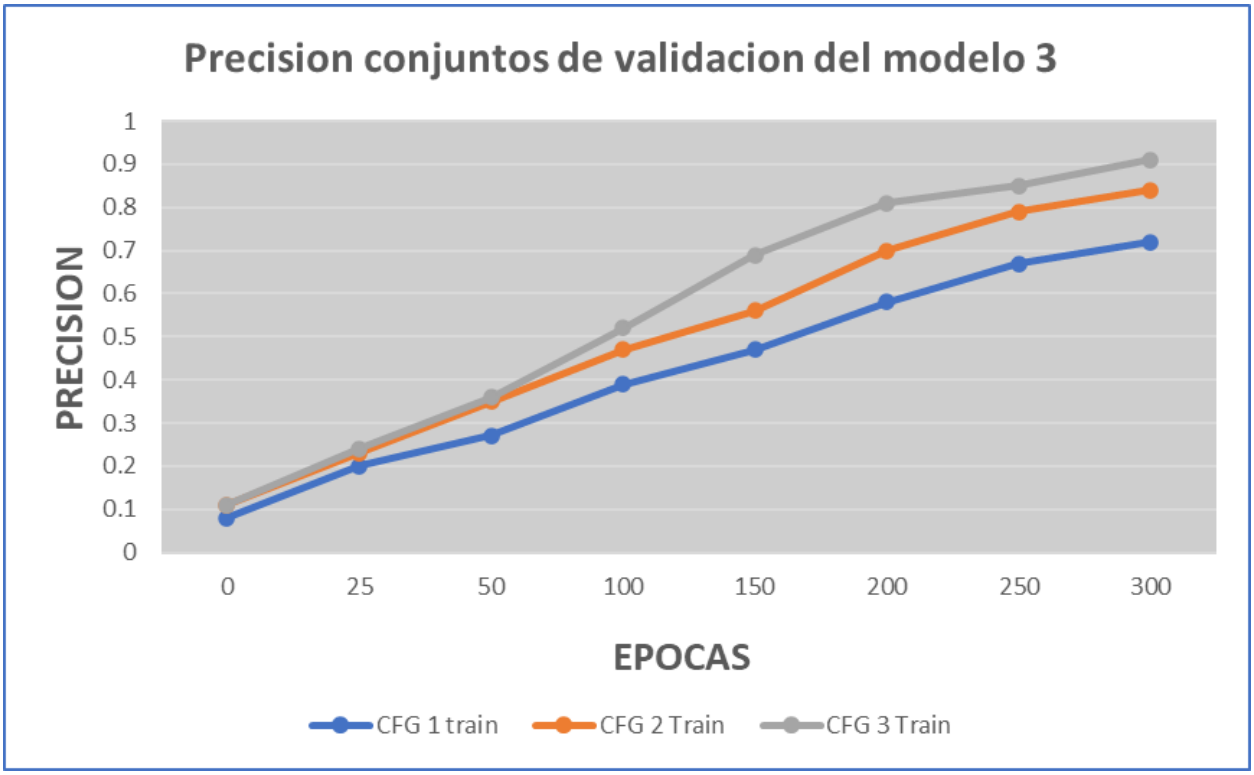


Figura 20. Rendimiento del modelo 3 en el conjunto de validación, Fuente: Propia

Este modelo presenta ligeras mejoras a lo visto en el modelo 2 CFG 2 en el conjunto de entrenamiento en las gráficas de la figura 19 se llega a nota una precisión del 91% en la CFG 3 de este modelo pero eso sería lo único positivo en este caso ya que en la tabla 22 se identifica que la pérdida en el mejor de los casos aumenta de 0.38 a 0.60 y el conjunto de validación tampoco mejora mucho ya que la precisión baja en todas las configuraciones en el mejor de los casos 71% es lo mejor que se obtuvo contra 75% de la CFG 2 del modelo 2 incluso la pérdida de la valoración tiende a subir en la mejor CFG 3 el da a entender que el modelo se tiende o puede sobre ajustarse si sigue entrenando

Tabla 23. Precisión por emoción obtenidas durante el entrenamiento del modelo 3

Emocion	Precision CFG 1	Precision CFG 2	Precision CFG 3
Happy	0.67	0.58	0.76
Sad	0.41	0.48	0.45
Tender	0.44	0.44	0.64
Fear	0.79	0.86	0.86
Anger	0.69	0.69	0.73
Surprise	0.65	0.62	0.59
High Valence	0.77	0.86	0.77
Low Valence	0.45	0.66	0.79
High Energy	0.73	0.86	0.77
Low Energy	0.60	0.80	0.76
High Tension	0.61	0.51	0.58
Low Tension	0.58	0.68	0.90

En la tabla 22 se muestran las precisiones obtenidas para este modelo los resultados en general de la configuración no muestran una mejoría con respecto a la CFG 2 del modelo algunas etiquetas en la CFG 3 como “low tension” y “low valance” mejoran lo cual es curioso ya que estas etiquetas han mostrado un rendimiento débil en los demás modelos

Tabla 24.Recall por emoción obtenidas durante el entrenamiento del modelo 3

Emocion	Recall CFG 1	Recall CFG 2	Recall CFG 3
Happy	0.71	0.79	0.68
Sad	0.70	0.58	0.87
Tender	0.47	0.68	0.62
Fear	0.56	0.59	0.68
Anger	0.58	0.52	0.56
Surprise	0.79	0.81	0.81
High Valence	0.45	0.59	0.85
Low Valence	0.61	0.66	0.82
High Energy	0.67	0.52	0.63
Low Energy	0.71	0.77	0.73
High Tension	0.65	0.76	0.75
Low Tension	0.62	0.86	0.76

En la tabla 23 que muestra los recalls es el mismo comportamiento que en la precisión solo algunas etiquetas mejoran su recall incluso alguna empeoran su rendimiento en comparación a CFG 2 del modelo 2

Tabla 25. F1-Score por emoción obtenidas durante el entrenamiento del modelo 3

Emocion	F1-score CFG 1	F1-score CFG 2	F1-score CFG 3
Happy	0.69	0.67	0.71
Sad	0.52	0.52	0.59
Tender	0.46	0.53	0.62
Fear	0.66	0.70	0.75
Anger	0.63	0.60	0.63
Surprise	0.72	0.70	0.68

High Valence	0.57	0.70	0.80
Low Valence	0.52	0.66	0.80
High Energy	0.70	0.65	0.69
Low Energy	0.65	0.78	0.74
High Tension	0.63	0.61	0.65
Low Tension	0.60	0.77	0.83

En la tabla del F1-score 24 se puede determinar que el modelo más equilibrado es el CFG 3 pero realmente no hay una mejora en comparación a CFG 2 modelo 2 las causas de los problemas que pudo tener este modelo se discutirán después de la tabla de los resultados de la matriz de confusión 25

Tabla 26. Resumen de los datos obtenidos en la matriz de confusión del modelo 3

Emocion	CM - VP CFG1	CM - FP CFG 1	CM - FN CFG 1	CM - VP CFG2	CM - FP CFG2	CM - FN CFG2	CM - VP CFG3	CM - FP CFG3	CM - FN CFG3	Etiquetas reales
Happy	22	11	8	19	14	6	19	14	6	33
Sad	12	17	5	14	15	10	14	15	10	29
Tender	11	14	12	11	14	5	11	14	5	25
Fear	23	6	18	25	4	17	25	4	17	29
Anger	18	8	13	18	8	16	18	8	16	26
Surprise	19	10	5	18	11	4	18	11	4	29
High Valence	17	5	21	19	3	12	19	3	12	22

Low Valence	11	13	7	16	8	8	16	8	8	24
High Energy	16	6	8	19	3	17	19	3	17	22
Low Energy	15	10	6	20	5	6	20	5	6	25
High Tension	19	12	10	16	15	5	16	15	5	31
Low Tension	17	12	10	20	9	3	20	9	3	29

El modelo sigue teniendo una buena precisión dada por la CM en este caso 71% en la CFG 3 pero aun así pierde 4% en comparación al CFG 2 del modelo 2 en general las predicciones son iguales o un poco mejores en algunas etiquetas que el modelo anterior pero en etiquetas como “happy” o “sad” el modelo tuvo una caída en los VP pasando de más de 24 VP a menos de 20 en ambas etiquetas esto claramente aumenta la cantidad de FP en el modelo y también hubo un aumento en los FN en general esto se puede deber a ciertas razones

- La complejidad del modelo ya que en sí este es el modelo con más capas ocultas de los tres modelos hasta el momento

- Los filtros en la capa de entrada pudieron modificar el rendimiento del modelo

- El sobre ajustamiento como se vio en la CFG el modelo en cierto punto no mejoraba en el conjunto de validación sino que empeoraba mediante pasaban las épocas

- Es posible que la configuración de hiperparametros para esta red no fueron los correctos aunque en la CFG se usaron los mismos que en [8]

Modelo 4: Para este modelo se removieron 2 capas convolucionales dejando originalmente las capas de 64 y 128 neuronas, se agrego una capa densa de 512 neuronas y se aumentó la capa densa restante a 512 neuronas la arquitectura quedaría de la siguiente manera:

- Una capa convolucional de 64 neuronas
- Una capa convolucional de 128 neuronas
- Dos capa Densas de 512 neuronas

Tabla 27. Resultados generales del entrenamiento del modelo 4

N° Neuronas / Etapas	50	100	150	300
Entrada de 64 Neuronas con 10 filtros, capa oculta convolucional de 256 neuronas y tres capas completamente conectadas de 256 neuronas	Accuracy: 0.44 Loss: 1.59 Val_acc: 0.37 Val_loss : 1.75	Accuracy: 0.52 Loss: 1.42 Val_acc : 0.47 Val_loss : 1.55	Accuracy : 0.68 Loss : 1.16 Val_acc : 0.55 Val_loss : 1.52	Accuracy : 0.94 Loss : 0.57 Val_acc : 0.64 Val_loss : 1.63
Entrada de 128 Neuronas con 10 filtros, capa oculta convolucional de 128 neuronas y tres capas completamente conectadas de 512 neuronas	Accuracy : 0.60 Loss : 1.06 Val_acc : 0.58 Val_loss : 1.11	Accuracy : 0.83 Loss : 0.48 Val_acc : 0.70 Val_loss : 1.00	Accuracy : 0.92 Loss : 0.24 Val_acc : 0.74 Val_loss : 0.94	Accuracy : 0.97 Loss : 0.14 Val_acc : 0.78 Val_loss : 1.23
Entrada de 128	Accuracy : 0.50	Accuracy : 0.73	Accuracy : 0.85	Accuracy : 0.96

Neuronas con 10 filtros, capa oculta convolucional de 256 neuronas,dos capas completamente conectadas de 512 neuronas y una capa completamente conectada de 1024 neuronas	Loss : 1.40 Val_acc : 0.50 Val_loss : 1.41	Loss : 0.98 Val_acc : 0.60 Val_loss : 1.36	Loss : 0.68 Val_acc : 0.73 Val_loss : 1.11	Loss : 0.34 Val_acc : 0.75 Val_loss : 1.23

En este modelo se presentan los mejores resultados de todo los experimentos de hiperparametros hasta el momento en términos de precisión se llega a alcanzar entre un 94% a 97% en las 3 configuraciones (CFG) de red consiguiendo pérdidas entre 0.57 y 0.14 siendo la más baja obtenida en la CFG 2 esto muestra que haber tomado como base la CFG 2 del modelo 2 y haber modificado un poco más los hiperparametros de esta red hace que la red tuviera un desempeño superior pero se también afirma algo que se sospechaba en el modelo 2 y 3 que es sobre ajustamiento ya que si se ve la CFG 2 de este modelo y la figura 21y 22 se podrá apreciar mejor

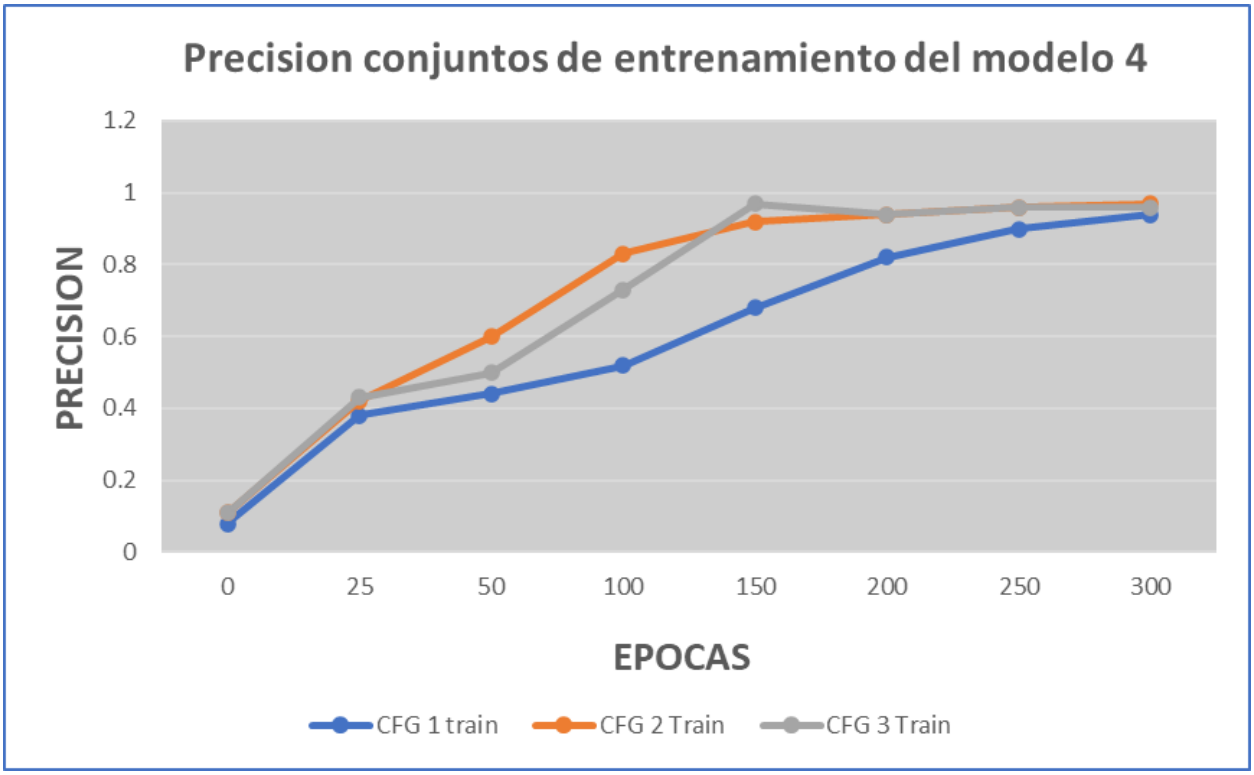


Figura 21. Rendimiento del modelo 4 en el conjunto de entrenamiento, Fuente: Propia

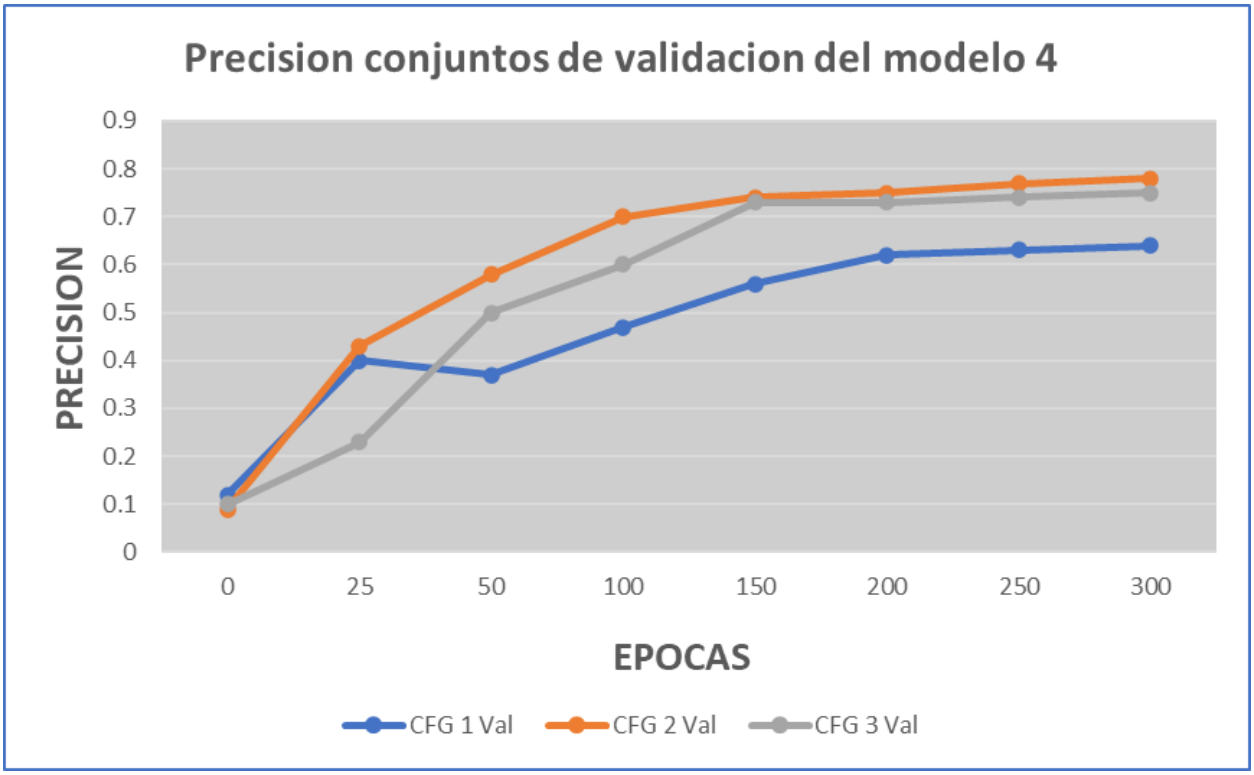


Figura 22. Rendimiento del modelo 4 en el conjunto de validación, Fuente: Propia

Se puede observar en todas las gráficas que a pesar que la presión sube en el conjunto de entrenamiento en todas las configuraciones el conjunto de validación se queda estancado después de la época 150+ entre 60% a 75% de precisión y si vemos las pérdidas en la tabla 26 veremos que hasta la época 150 la pérdida baja pero al final de la época 300 esta empieza a subir esto significa que el modelo se está sobre ajustando a los datos de entrenamiento que a pesar de usar capas DropOut y técnicas como la regularización no se ha podido contener del todo este problema

Tabla 28 Precisión por emoción obtenidas durante el entrenamiento del modelo 4

Emocion	Precision CFG 1	Precision CFG 2	Precision CFG 3
Happy	0.85	0.76	0.75
Sad	0.62	0.62	0.62
Tender	0.84	0.64	0.68
Fear	0.79	0.90	0.89
Anger	0.84	0.92	0.88
Surprise	0.72	0.72	0.68
High Valence	0.86	0.81	0.81
Low Valence	0.62	0.79	0.92
High Energy	0.84	0.86	0.90
Low Energy	0.76	0.92	0.56
High Tension	0.74	0.64	0.67
Low Tension	0.90	0.86	0.75

Las predicciones de etiquetas emocionales en este caso han mejorado en la mayoría de etiquetas en comparación al modelo 2 CFG 2 a excepción de la etiqueta “sad” que cae de 82% a 62% en las 3 CFG de este modelo en general a pesar de no haber alcanzado una etiqueta con 100% de precisión como en anteriores modelos la precisión se encuentra entre 79% a 92% a excepción de la etiqueta “sad” lo cual es aceptable

Tabla 29. Recall por emocion obtenidas durante el entrenamiento del modelo 4

Emocion	Recall CFG 1	Recall CFG 2	Recall CFG 3

Happy	0.75	0.80	0.73
Sad	0.94	0.94	0.85
Tender	0.70	0.72	0.80
Fear	0.72	0.62	0.74
Anger	0.81	0.75	0.88
Surprise	0.66	0.75	0.80
High Valence	0.83	0.81	0.69
Low Valence	0.93	0.90	0.56
High Energy	0.75	0.67	0.62
Low Energy	0.76	0.71	0.70
High Tension	0.95	0.95	1.00
Low Tension	0.74	0.86	0.91

El recall también muestra mejoría en la mayoría de sus etiquetas teniendo un recall entre 62% a 95% siendo el menor la etiqueta fear la cual igual que en modelo 2 CFG 2 se mostraba con una precisión alta y un recall bajo a comparación con las demás etiquetas emocionales

Tabla 30. F1-Score por emocion obtenidas durante el entrenamiento del modelo 4

Emocion	F1-score CFG 1	F1-score CFG 2	F1-score CFG 3
Happy	0.80	0.78	0.74
Sad	0.75	0.75	0.72
Tender	0.76	0.68	0.74
Fear	0.75	0.73	0.81
Anger	0.83	0.83	0.88
Surprise	0.68	0.74	0.74

High Valence	0.84	0.82	0.75
Low Valence	0.75	0.84	0.69
High Energy	0.78	0.76	0.74
Low Energy	0.76	0.81	0.62
High Tension	0.83	0.76	0.80
Low Tension	0.81	0.91	0.83

Según la tabla del F1-Score para este modelo 21 la configuración que obtuvo un mejor rendimiento en el entrenamiento fue la CFG 1 si tomamos esta configuración y la comparamos con lo que era nuestro mejor modelo con configuración (modelo 2 CFG 2) nos daremos cuenta que efectivamente la relación precision-recall o la detección de emociones del modelo ha mejorado con este nuevo modelo por último la tabla de la CM generada en este modelo [mum ref] la etiqueta “sad” es la débil de predecir para este modelo ya que tiene 20 FP lo cual indica que es la etiqueta más complicada de aprender para el modelo como también se vio en el modelo 1 y modelo 2

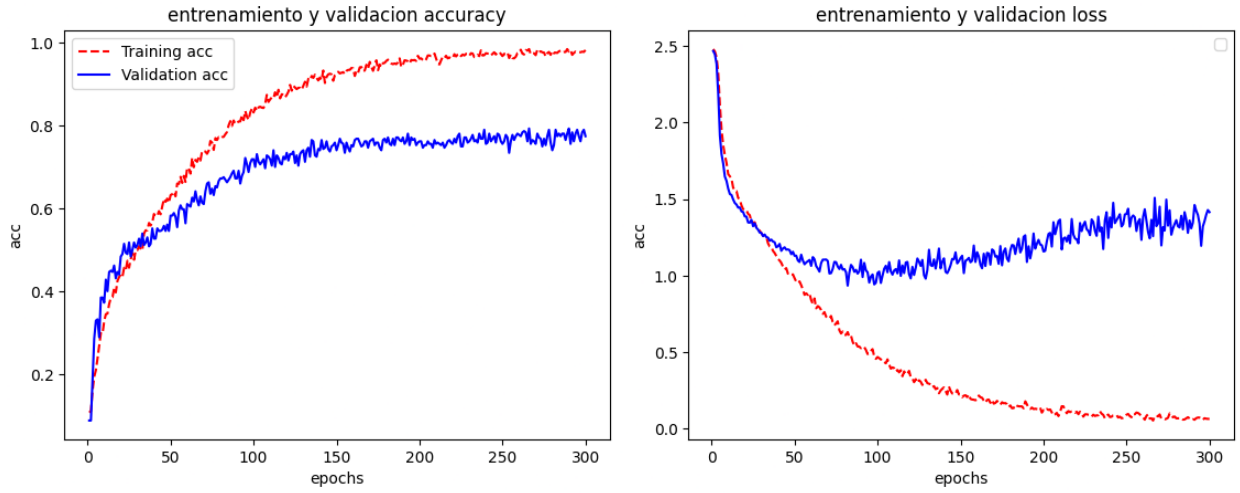
Tabla 31. Resumen de los datos obtenidos en la matriz de confusión del modelo 4

EEmocion	CM - VP CFG1	CM - FP CFG 1	CM - FN CFG 1	CM - VP CFG2	CM - FP CFG2	CM - FN CFG2	CM - VP CFG3	CM - FP CFG3	CM - FN CFG3	EEtiquetas reales
Happy	26	7	5	25	8	6	25	8	9	33
Sad	9	20	2	18	11	1	18	11	3	29
Tender	18	7	8	16	9	6	17	8	4	25
Fear	25	4	11	26	3	15	26	3	9	29
Anger	22	4	14	24	2	8	23	3	3	26

Surprise	23	6	10	21	8	7	20	9	5	29
High Valence	20	2	7	18	4	4	18	4	8	22
Low Valence	17	7	4	19	5	2	22	2	17	24
High Energy	18	4	12	19	3	9	20	2	12	22
Low Energy	15	10	5	23	2	9	14	11	6	25
High Tension	17	18	5	20	11	1	21	10	0	31
Low Tension	25	4	7	25	4	1	22	7	2	29

A pesar de que se consiguió el objetivo de conseguir el 75% de precisión del modelo para detectar emociones en la música aún se tiene el problema de sobre ajustamiento si lo dejamos de esta forma cuando ingresemos nuevos datos el modelo detectara una emoción pero es posible que la emoción detectada sea errónea ya que el modelo se sobreajuste a los datos del entrenamiento por eso se busca aplicar otra técnica lo cual la validación cruzada con el mejor modelo y configuración obtenido en estas interacciones

Validación cruzada: Después de hacer diferentes experimentos con distintos modelos CNN y se pudo conseguir un modelo que rindiera aceptablemente con el dataset, pero tenía un pequeño sobre ajustamiento después que pasaba cierta interacción como se ve en la siguiente imagen



Se puede observar que la precisión en el conjunto de validación se queda estancada entre 70% a 80% de precision mientras que la pérdida empieza a subir casi en la misma época 150+, anteriormente se realizó un proceso de aumento de datos sobre el dataset base que se probó en el experimento del modelo actual + el dataset original que fue acertado ya que al haber trabajado solo con los 360 datos originales podríamos haber tenido un problema de sobre ajustamiento mucho más grande, si volviéramos a hacer este proceso sobre los datos actuales podríamos tener ciertos problemas como variaciones no realistas de los datos como lo menciona [42] o podríamos utilizar más recursos [39, 42] o incluso limitar la propia tarea que estamos intentando realizar [39]

Una de las formas más comunes para solucionar este problema es el uso de la validación cruzada algunos trabajos como [2-3,7-10,12,15-16,21] hacen uso de esta técnica en lugar de la típica división de conjuntos de train-val-test así que eso es lo que se intenta hacer aquí. Lo que se hace a continuación es tomar las características y puntuaciones (nuestra variable X) y dividirlas en 5 conjuntos los cuales cada conjunto tendrá un fold o un pliegue como se ve en la siguiente imagen

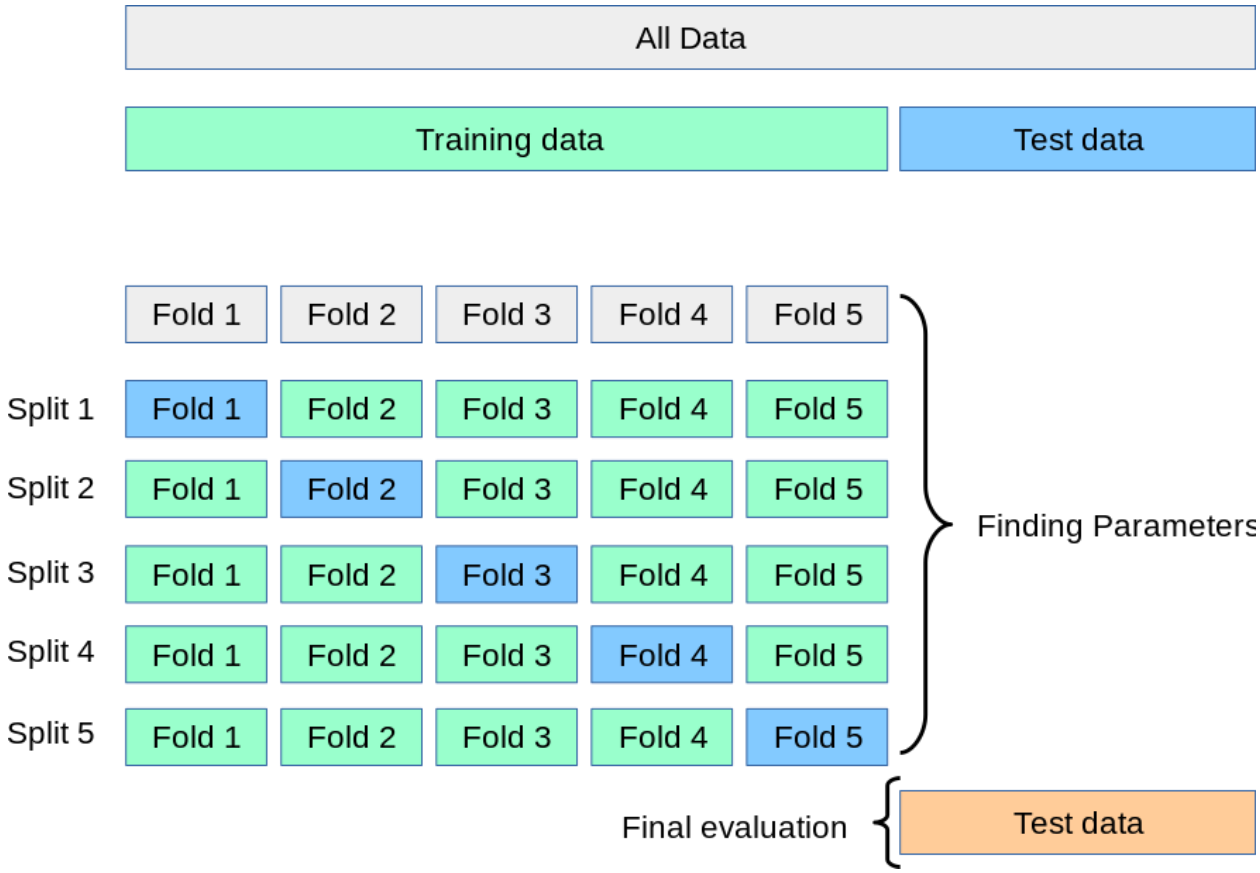
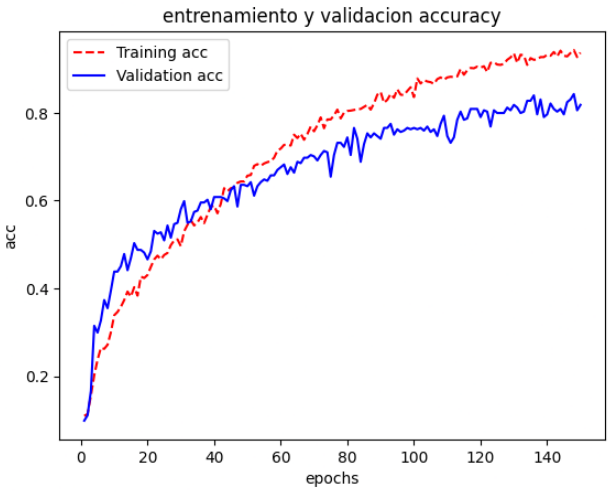


Figura 23. Ejemplo de partición de conjunto de datos durante la validación cruzada, Fuente: Propia

Usaremos el mismo modelo que usamos en la cuarta interacción con los mismos hiperparametros y capas de entrada, salida y ocultas con 150 etapas ya que desde esa etapa el modelo empieza a tener un sobre ajustamiento. En este caso tendremos 5 diferentes interacciones con el modelo la primera interacción tomará el primer pliegue o fold como conjunto de pruebas y los otros 4 pliegues como conjunto de entrenamiento esto se repetirá hasta que todos los pliegues hayan sido utilizados como conjunto de pruebas esto nos ayuda a tener una evaluación más robusta y completa del dataset y nos ayuda a determinar el rendimiento del

modelo ya que se evalúan diferentes partes del modelo y no una parte aleatoria como se hacía anteriormente

Interacción 1: Para esta primera interacción observamos un comportamiento similar a el entrenamiento previo del modelo sin validación cruzada como se ven en las figuras 21 y 22 se nota una diferencia mínima entre la precisión y pérdida en ambos conjuntos. De hecho, en las primeras 40 épocas, el conjunto de validación tiende a tener una precisión ligeramente más alta que el conjunto de entrenamiento. Sin embargo, esta diferencia tiende a superarse después de la época 50 en adelante. Esto sugiere que en conjunto de entrenamiento mejora su capacidad para clasificar sus muestras, aunque el conjunto de validación no queda rezagado del todo. Las gráficas muestran que ambos conjuntos tienden a estabilizar su precisión en valores cercanos a 75% a 85% y en pérdidas en 0.3-1.4



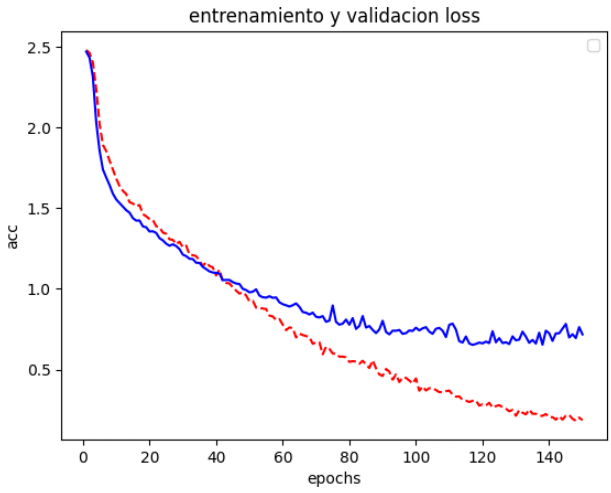


Figura 24. Rendimiento y perdida de los conjuntos de entrenamiento y validación durante la interacción 1, **F**uente: Propia

A pesar de que la precisión en los conjuntos de validación y prueba no muestran diferencia al final del entrenamiento de la primera interacción en la gráfica de pérdida se nota que después de la época 80+ se empieza a estancar y tiende a subir la pérdida un poco similar a lo observado en 24 por lo que sugiere que sigue teniendo un sobre ajustamiento con esta primera parte del dataset que puede ser resuelto con el paso de las interacción ya que este primer entrenamiento sienta las bases para las próximas fases de entrenamiento y ajusta los pesos que se utilizaran en las siguientes interacciones. Observando la matriz de confusión y estadísticas por emocion generadas en esta interacción tenemos

Tabla 32. Resumen métricas obtenidas en el entrenamiento de la interacción 1

Emocion	Precision	Recall	F1-Score	MC - VP	MC - FP	MC - FN	Reales
Happy	0.67	0.86	0.74	22	11	4	33
Sad	0.66	0.95	0.77	19	20	1	39
Tender	0.76	0.79	0.77	19	6	5	25
Fear	0.90	0.70	0.79	26	3	11	29
Anger	0.85	0.85	0.85	22	4	4	26
Surprise	0.69	0.77	0.73	20	9	6	29
High	1.00	0.63	0.77	22	0	13	22

Valence							
Low Valence	0.83	0.77	0.80	20	4	6	24
High Energy	0.96	0.64	0.77	21	1	12	22
Low Energy	0.56	0.82	0.67	14	11	3	25
High Tension	0.81	0.96	0.87	25	6	1	31
Low Tension	0.79	0.82	0.80	23	6	5	29

Segun las metricas y matriz de confusión podemos observar que el modelo tiene una precisión en general del 78% si calculamos todas los VP tomados de la diagonal del matriz de confusion y lo dividimos entre todas las instancias del modelo.El en el entrenamiento de etiquetas es rendimiento sobresaliente al lograr la precisión del 100% en la etiqueta high valence ya que de 22 etiquetas reales el modelo es capaz de identificar todas como VP. Sin embargo el recall de esta etiqueta resultó ser la más baja ya que fue del 63% esto indica que se tuvieron 37% de FN predichos por el modelo esto se puede ver en el eje X donde se ve que el modelo predice incorrectamente 9 etiquetas emocionales. Por otra parte la etiqueta Low Energy tuvo la precisión más baja de todas las etiquetas con 56% ya que de 24 etiquetas reales el modelo solo pudo predecir 14 correctamente como VP. En general el modelo con las demás etiquetas presenta un rendimiento aceptable de 66% a 85% de precisión y de 65% a 95% en recall

Interacción 2: en el gráfico de la figura 25 se puede observar los cambios de precisión y pérdida de los conjuntos de entrenamiento y validación, se puede observar como en el conjunto de validación empieza con una precisión muy alta del 99% que al paso de las épocas va disminuyendo hasta 94% esto se puede deber a la modificación de pesos que hay entre una

interacción a otra, por otro lado en el conjunto de entrenamiento es diferente se ve empieza casi en el mismo punto donde termina la última interacción y empieza a subir hasta del 81% al 97% y 98% esto indica que el modelo sigue aprendiendo los patrones que puede encontrar en los datos de entrenamiento de los pliegues antiguos y en el pliegue de datos que no estudio en la interacción pasada.

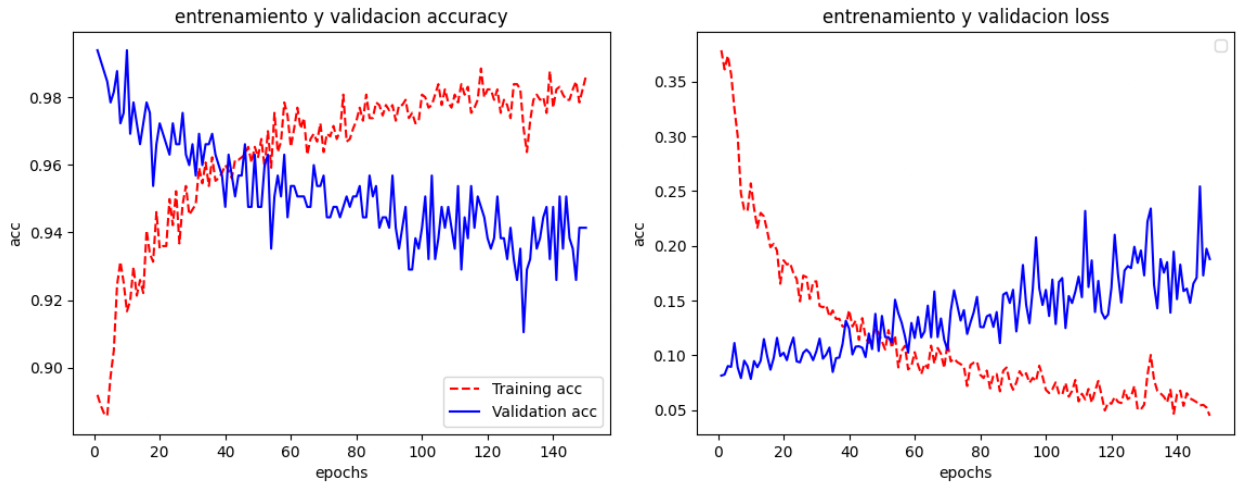


Figura 25. Rendimiento y perdida de los conjuntos de entrenamiento y validación durante la interacción 2, Fuente: Propia

En la gráfica de la pérdida tiene un comportamiento similar al de la precisión dónde en el conjunto de entrenamiento empieza a bajar del 0.35 a 0.05 esto indica que el modelo puede seguir aprendiendo de los datos de entrenamiento pero en el conjunto de validación empieza entre 0.08 y a lo largo de las épocas empieza a subir hasta 0.25 aunque el aumento es poco comparado a la pérdida en la figura 24 este comportamiento más la disminución de la precisión de 98% al 94% puede aun ser un síntoma de sobre ajustamiento ya que el modelo en esta interacción la precisión

en conjunto de entrenamiento llega casi al 100% y puede estar sobre ajustándose a los datos de este conjunto si en futuras interacciones se sigue viendo este tipo de comportamiento en el conjunto de validación. Si vemos los datos que arroja la matriz de confusión y tabla de estadísticas de etiquetas observaremos que hay una mejora con respecto a la interacción pasada

Tabla 33.Resumen métricas obtenidas en el entrenamiento de la interaccion 2

Emocion	Precision	Recall	F1-Score	MC - VP	MC - FP	MC - FN	Reales
Happy	0.96	0.86	0.91	26	1	4	27
Sad	0.96	0.83	0.89	25	1	5	26
Tender	0.94	0.88	0.91	16	1	2	17
Fear	0.90	0.96	0.93	27	3	1	30
Anger	0.72	0.96	0.82	24	9	1	33
Surprise	1.00	0.95	0.98	20	0	1	20
High Valence	0.78	1.00	0.88	25	7	0	32
Low Valence	1.00	0.86	0.93	20	0	3	20
High Energy	0.92	0.96	0.94	24	2	1	26
Low Energy	1.00	0.96	0.98	29	0	1	29
High Tension	1.00	0.89	0.94	33	0	4	33
Low Tension	0.96	0.94	0.95	30	1	0	31

Comparando la CM de la primera interacción con la presente se observa una variación en el número de etiquetas predichas para cada emoción. Este cambio en el número de etiquetas es esperado ya que en cada interacción de la validación cruzada se usa un nuevo conjunto de entrenamiento y prueba como se ilustra en la figura 23. En esta segunda interacción el modelo logra una precisión en general del 92.2% representando una mejora del 14% con respecto al modelo de la interacción anterior.

Al observar las estadísticas por emoción se destaca que el modelo presenta mejores precisiones en casi todas las etiquetas emocionales. Cuatro de ellas presentan una precisión del 100% y la mayoría un recall de 83% a 100% la peor etiqueta en este caso fue la etiqueta “sad” ya que registra 5 FN y 25 VP, es de destacar que la etiqueta “high valence” que tenía una precisión del 100% en la interacción anterior ahora muestra un 78% de precisión sin embargo se observa una mejora en su recall del 63% al 100% este cambio podría deberse a la utilización de un nuevo conjunto de pruebas el cual incluye más o menos instancias para identificar en comparación del conjunto anterior.

Interacción 3: En esta interacción el modelo presenta mejoras notables con respecto al manejo del sobre ajustamiento con respecto a la interacción 1, si se ve la figura de la figura 26 se puede determinar que ambos conjuntos están llegando al 100% en precisión incluso el conjunto de validación empieza al 100% esta interacción pero tiende a bajar hasta 98% al paso de las épocas como en la interacción 2 pero en este caso el conjunto de entrenamiento tiene un precisión máxima de 99% lo que indica un control más efectivo del sobreajuste en comparacion a la primera interacción. Este resultado sugiere que el modelo esta generalizando de manera más efectiva los datos no vistos, evitando la tendencia al sobreajuste que se habia observado previamente

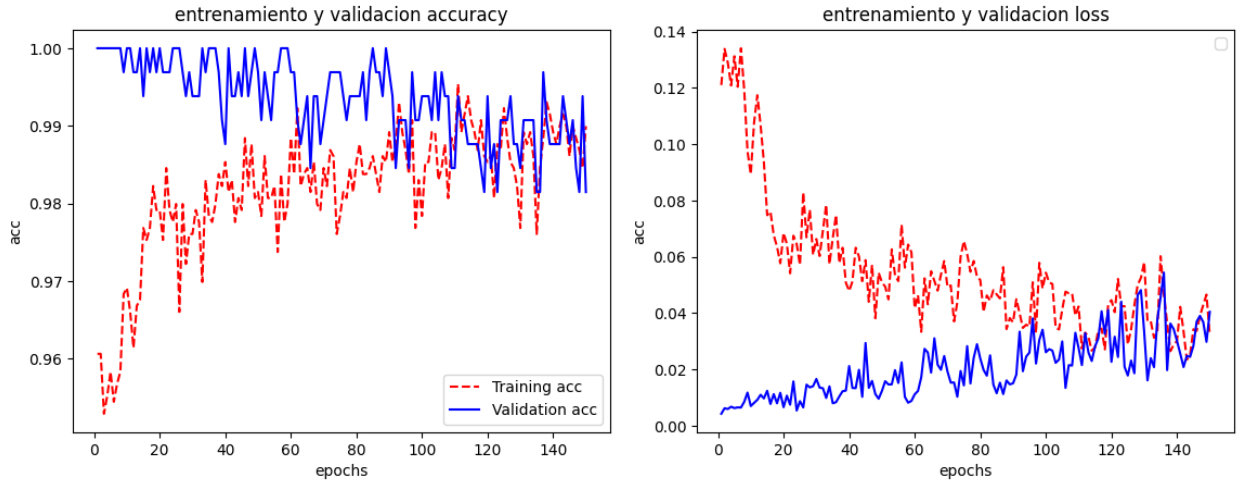


Figura 26. Rendimiento y perdida de los conjuntos de entrenamiento y validación durante la interacción 3, Fuente: Propia

Si vemos la gráfica de pérdida en ambos conjuntos tenemos pérdidas buenas finalizando el entrenamiento con 0.03 y 0.04 indicando que el sobreajuste se ha minimizado con las técnicas utilizadas hasta el momento y que el modelo generaliza bien los datos. SI observamos la CM y estadísticas de esta interacción

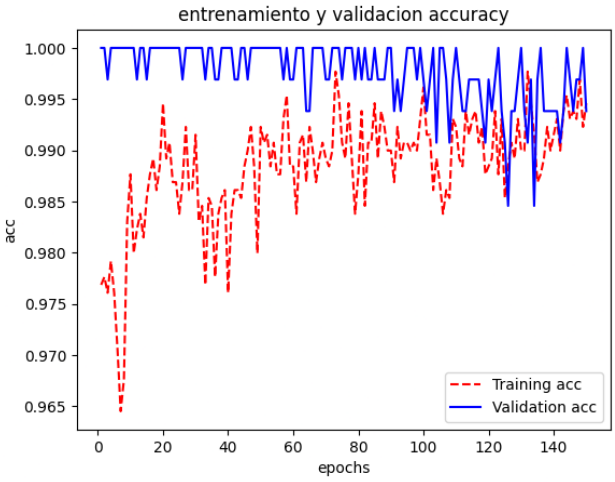
Tabla 34. Resumen metricas obtenidas en el entrenamiento de la interacción 3

Emocion	Precision	Recall	F1-Score	MC - VP	MC - FP	MC - FN	Reales
Happy	1.00	1.00	1.00	26	0	0	27
Sad	0.96	0.95	0.98	23	1	0	26
Tender	1.00	1.00	0.97	38	0	2	17
Fear	0.93	1.00	0.97	28	2	0	30
Anger	0.92	1.00	0.96	24	2	0	33
Surprise	1.00	1.00	1.00	28	0	0	20
High Valence	0.96	1.00	0.98	28	1	0	32
Low Valence	1.00	1.00	1.00	25	0	0	20
High Energy	1.00	0.88	0.94	22	0	3	26
Low Energy	1.00	0.96	0.98	26	0	1	29
High Tension	1.00	1.00	1.00	25	0	0	33
Low Tension	1.00	1.00	1.00	25	0	0	31

En general el modelo vuelve a dar un buen rendimiento en precisión tanto en la predicción de etiquetas emocionales que en esta interacción fue del 98.1 mejorando un 5% con respecto a la interacción anterior dando un total de 7 FP en todo el entrenamiento, si miramos las estadísticas por etiqueta solo 4 etiquetas no tienen un puntaje de 100% en precisión siendo la más baja del 92% de nuevo en este caso es anger al igual que la interacción 2 pero en este caso es más aceptable ya que se tiene una mejora del 20% con respecto a la interacción anterior.En el recall solo 3 etiquetas no alcanzan el 100% siendo la más baja high energy con 88% el cual tuvo 3 FN

en todo el entrenamiento a este punto se puede parar el entrenamiento ya que ambos conjuntos tienen una precisión de casi 100% y una pérdida bastante baja pero se hará una interacción para ver se desarrolla el modelo en esta configuración y observar si hay sobre ajustamiento

Interacción 4 y 5: En estas últimas interacción podemos ver una mejora mínima en los conjuntos de entrenamiento si tomamos como referencia solo las figura 27-28 veremos qué precisión en ambos conjuntos es casi del 100% solo bajando al 99.5% si se sigue entrenando se llegara a un estado donde la precisión ya no baje del 100% pero en este caso ya no es necesario más entrenar ya que se cumplió con objetivo de tener un precisión alta y pérdida baja en ambos conjuntos para evitar el sobre ajustamiento y hacer una buena predicción de las emociones cuando se pasen datos no vistos



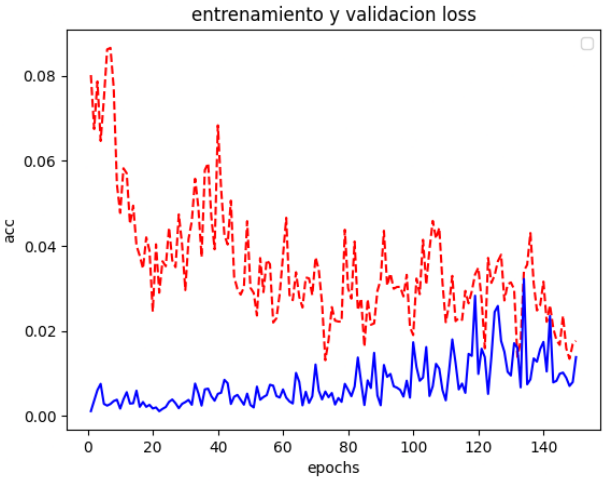


Figura 27. Rendimiento y perdida de los conjuntos de entrenamiento y validación durante la interacción 4, Fuente: Propia

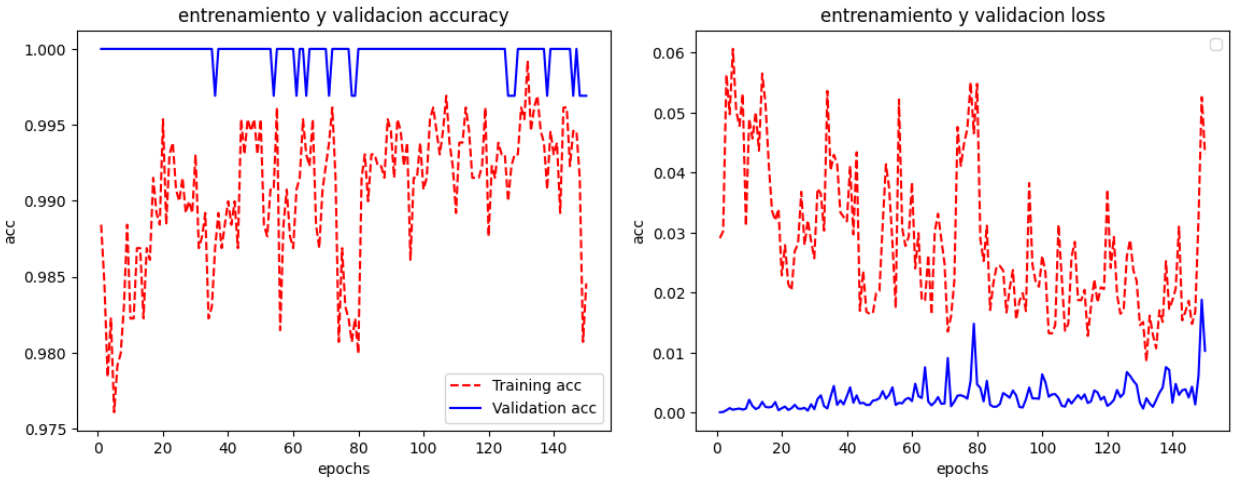


Figura 28. Rendimiento y perdida de los conjuntos de entrenamiento y validación durante la interacción 5, Fuente: Propia

Con la CM y estadísticas de la última interacción tenemos que el modelo una precisión en general del 99.6% solo una mejora del 0.4% con respecto a la tercera interacción y por parte de las etiquetas emocionales solo una etiqueta no logra llegar al 100% el cual es la etiqueta tender con un solo FP dando un 96% de precisión que en cualquier caso es una precision muy buena y en recall todas las emociones dan un total de 100% a excepción de low energy con solo 1 FN.

FASE 5 PRUEBAS Y EVALUACION:

Evaluación de resultados: los datos usados en entrenamiento del modelo arrojan una precisión general del 95%, un recall del 94% y un F1-Score del 93% usando la técnica de validación cruzada con 5 pliegues con 150 etapas para cada pliegue. Estos resultados permiten concluir que el objetivo de la minería de datos de obtener al menos un 75% de precisión en la detección de emociones ha sido logrado con el dataset y modelos propuestos dado que las métricas obtenidas son muy buenas y arrojan bajos FN Y FP.

El siguiente paso para determinar qué tan bueno es el modelo anteriormente creado es probarlo con distintos datasets que hayan tenido efectividad en el campo del MER por ejemplo como los usado en [41] donde se usa el dataset Bi-modal o en [45] donde se usa un dataset similar a Soundtrack con características musicales ya extraídas se hará el mismo procedimiento de pre-procesado y entrenamiento con el modelo y se comparan las métricas de rendimiento de cada dataset buscando un resultado igual o mejor que con dataset SoundTracks 360.

Evaluación del modelo con diferentes Datasets: Como se puede ver en la tabla 35 el modelo tiene un buen rendimiento en los distintos datasets usados para la detección de emociones en la música, logrando resultados mayores al 90% siendo el peor de todos el dataset Bi-modal [40] que tiene una precisión del 92% en comparación a los demas que tienen una precisión superior al 95%

Tabla 35. Resultados de las pruebas de los datasets con entrenamiento normal

Nombre del dataset	Precisión Entrenamiento del modelo	Pérdida entrenamiento del modelo	Precisión valoración del modelo	Pérdida de valoración del modelo
SoundTracks360	97.00%	0.0306	74.69%	1.6512
MER_audio_taffc	98.9999%	0.0150	73.3333%	2.1533
Bi-modal	92.5%	0.0272	74.44%	1.76
Turkish Dataset	99.8%	0.01222	72.50%	2.0893
Hindi Dataset	99.24%	0.0331	80.81%	0.5408

También podremos notar en la tabla 34 que los datasets tienen un comportamiento similar que el dataset SoundTracks que a pesar de tener una precisión de entrenamiento alta y una precisión de variación mayor al 70% se tiene una pérdida de valoración muy alta originalmente datasets como Turkish [16], Hindi [10] y Bi-modal [45] se entrenaron con validación cruzada entre 5 a 10 pliegues lo cual es lo que se muestra en la tabla 35

Tabla 36. Resultado de las pruebas con validación cruzada

Nombre del dataset	Precisión entrenamiento promedio del modelo	Pérdida entrenamiento promedio del modelo	Precisión promedio valoración	Pérdida promedio valoración
SoundTracks360	95%	0.0150	93.19%	0.0200

MER_audio_taffc	98.1%	0.0209	98.9999%	0.0020
Bi-modal	92.5%	0.011	92.5%	0.0119
Turkish Dataset	94.32%	0.0059	93.5%	0.0119
Hindi Dataset	99.64%	0.0041	98.68%	0.0250

En la tabla 35 se muestran los resultados obtenidos en el entrenamiento de los diferentes datasets usando validación cruzada con 5 pliegos como se hizo en la fase 4 se ven resultados más bajos que las que se usan en el entrenamiento normal pero esto es normal ya que el promedio de los resultados obtenidos en la validación se obtienen resultados mayores al 90% siendo de nuevo Bi-modal el dataset con la peor precisión con 92% y el dataset Hindi con el mejor resultado del 99% superando el dataset usado para este proyecto

Determinar pasos a seguir: Con la evaluación del modelo usando diferentes datasets se puede deducir que el modelo debería ser capaz de predecir una o varias emociones que provengan de un fragmento de música ya con los resultados obtenidos en la fase anterior dan precisiones buenas en el entrenamiento normal y pérdidas bajas con la validación cruzada en este caso todas las interacciones con los dataset.

En anexo 1 se puede observar las diferentes MC y resultados por emoción que se obtuvieron en las interacciones pasadas tanto en el entrenamiento normal como en las interacciones de validación cruzada

FASE 6: DESPLIEGUE

Reporte final: Entendimiento del negocio

El modelo construido cumple con los objetivos del negocio propuestos ya que da un buen rendimiento con la utilización de la validación cruzada, la experimentación hecha en la fase 5 del CRISP-DM fue un paso importante para el desarrollo del proyecto ya que a medida que el primer modelo se mejoraba agregando capas o modificando hiperparametros se conseguían mejores resultados mediante pasaban las épocas de entrenamiento a excepción de la interacción 3 ya que él ni el modelo propuesto ni ninguna CFG lograba mejorar los resultados de modelos pasados

Proceso de minería: La librería librosa fue de gran utilidad para el desarrollo de este proyecto ya que se usó para la extracción de características de cada fragmento musical, así como para el proceso de aumento de datos

Resultado de la minería de datos: Los resultados obtenidos en el proceso de minería de datos se justifican al haber alcanzado los objetivos propuestos al haber construido un modelo de detección de emociones con una precisión mayor al 75% y haberlo probado con diferentes datasets y haber alcanzado una cifra igual o mayor a la planteada

Despliegue de los resultados: Los resultados obtenidos durante el entrenamiento del modelo se presentan de diferentes formas durante el entrenamiento del modelo se presentan las estadísticas por época las cuales muestran en entrenamiento y pérdida del modelo en los conjuntos de entrenamiento y validación, al acabar el entrenamiento estas estadísticas son

resumidas en dos gráficas como las que se encuentran en el anexo 1, después se genera una matriz de confusión donde se muestran los VP, FN Y FP que predice el modelo por cada emoción además generando la precisión general del modelo por último se resumen estadísticas por emoción generando precisión, recall y F1-Score basándose en la matriz de confusión

4. ALCANCES Y LIMITACIONES

El desarrollo de aplicaciones que usan técnicas de aprendizaje profundo para detectar emociones en la música tiene un campo de investigación grande y en constante evolución. Esto se apoya en la capacidad de las IA para procesar grandes volúmenes de datos y la capacidad de aprender patrones complejos haciendo posible la creación de herramientas computacionales capaces de identificar las emociones en distintos tipos de formatos.

Este proyecto ha buscado diseñar una herramienta computacional para la detección de emociones en la música. Se busca identificar y aplicar algoritmos de aprendizaje profundo que puedan ser apropiados para esta tarea utilizando diversos conjuntos de datos para entrenar, validar y evaluar la eficiencia del modelo haciendo así una herramienta automatizada que permita analizar distintos géneros musicales en busca de patrones emocionales. Pero hay que destacar que existen limitaciones con la herramienta como que la detección precisa de emociones no es una tarea sencilla debido a la diferentes emociones que pueden surgir o que sentir una persona al momento de escuchar una canción ya que no todas las personas sienten lo mismo al momento de oír la misma canción, además otro aspecto importante a considerar es la disponibilidad y que tan representativos son los conjuntos de datos que se utilizan ya que la calidad y diversidad de estos puede influir importantemente en la capacidad del modelo para generalizar distintos contextos musicales.

5. CONCLUSIONES Y TRABAJOS FUTUROS

5.1 CONCLUSIONES

Revisión de la literatura

Para llevar a cabo el desarrollo del proyecto fue necesario realizar una revisión de la literatura basada en el campo del MER dado que hay muchas redes neuronales disponibles, así como diversas características que combinadas son capaces de realizar esta tarea exitosamente, para ello fue necesario formular cuatro preguntas de investigación para tener una idea de cómo desarrollar el proyecto. Se abordaron temas como que redes neuronales eran las más usadas en el MER, que tipo de características se extraían en este proyecto, que tipo de enfoque usaban y por ultimo las métricas que generaba el proyecto Las respuestas a estas preguntas proporciono información valiosa que dan una base solida para lo que fue el desarrollo del proyecto, así como que desafíos se podrían encontrar durante el desarrollo del proyecto y las buenas prácticas usadas en el campo del MER.

Librerias y Frameworks

Para el desarrollo del proyecto se utilizado especialmente la biblioteca Librosa esencialmente para llevar a cabo la extracción de características de la música tales como el tono, la melodía, características Chroma, espectrales, etc. También haciendo uso de la misma para el proceso de aumento de datos permitiendo la creación de un conjunto de datos más robusto para alimentar el modelo con información diversa y por consecuente mejorar significativamente el rendimiento del mismo.

Metodología CRISP-DM

Esta guía ha sido de gran ayuda para el desarrollo del proyecto, aprovechando los datos obtenidos durante la revisión sistemática. Los procesos y actividades de la metodología han sido diseñados para asegurar el éxito en el desarrollo del proyecto. Además, este guía no solo ha dado pasos específicos, sino que incorpora actividades de evaluación que respaldan la realización de análisis profundo de los resultados obtenidos. De esta manera, se ha logrado un enfoque efectivo que impulsa la ejecución eficaz del proyecto, sino que también promueve la mejora continua a lo largo de todo el proceso de desarrollo.

Modelo para detección de emociones

A pesar de tener características indicadas para cada genero de música del dataset escogido la creación del modelo fue compleja ya que se tuvo que pasar por un largo proceso de evaluación para poder diseñar la mejor configuración de red posible dando resultado en sus peores configuraciones de 65% de precisión hasta en los mejores resultados en 99%.

Se pudo observar que el uso de capas influye mucho en rendimiento del modelo tanto en cuestión de tiempo de ejecución, así como en los resultados finales usar muchas capas no es en este caso una buena opción como se vio en el modelo 3 donde se resalta el sobre ajustamiento del modelo a pesar de tener un buen rendimiento en un conjunto de datos. Modelos más simples como el 2 o 4 dan mejores resultados a pesar de no poder alcanzar esa precisión lo cual hace usar otras técnicas como la validación cruzada la cual fue el punto clave de la experimentación y donde se pudo obtener los mejores resultados de todo el proyecto.

5.2 TRABAJOS FUTUROS

Una de las mejoras que se puede hacer al modelo y que se exploro durante del desarrollo de la revisión sistemática es el uso de una o más redes neuronales como por ejemplo una CRNN o una CDLNN como se vio en la misma revisión la gran mayoría de estos modelos son capaces de hacer la tarea de detectar emociones en la música de forma eficiente el uso combinado de estas redes puede mejorar el rendimiento del modelo, así como mejorar la detección de emociones en general.

Se podría experimentar con otros datasets más robustos donde no se tenga que requerir la técnica de aumento de datos donde se consideren una mayor variedad de géneros musicales y etiquetas emocionales para mejorar la capacidad de generalización del modelo.

Se puede experimentar otras técnicas para solucionar el sobre ajustamiento del modelo a pesar que el uso de la validación cruzada fue efectivo para el modelo existen otras técnicas como la validación cruzada estratificada si el dataset utilizado tiene clases desbalanceadas después del aumento de datos o el uso de Ensemble learning el cual consiste en combinar multiples modelos para mejorar la generalización o incluso usando métodos denominados “Poda” los cuales eliminan características poco importantes para el modelo el cual reduce la complejidad del mismo

6. **BIBLIOGRAFIA**

[1] S. Saha, “A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way | by Sumit Saha | Towards Data Science,” Towards Data Science. Accessed: Mar. 12, 2023. [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

[2] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, *Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis*. 2013.

[3] T. Liu, L. Han, L. Ma, and D. Guo, “Audio-based deep music emotion recognition,” 2018, p. 040021. doi: 10.1063/1.5039095.

[4] yi-hsuan Yang, Y.-C. Lin, Y.-F. Su, and H. Chen, “yangtaslp08 emo regression.” Jul. 2014.

[5] Blazquez Daniel, “Desarrollo de un sistema de detección del nivel de música sobre voz en señales de audio,” UNIVERSIDAD POLITECNICA DE MADRID, Madrid, 2019.

[6] R. Zatarain Cabada, M. L. Barrón Estrada, and H. M. Cárdenas López, “Reconocimiento multimodal de emociones orientadas al aprendizaje,” *Research in Computing Science*, vol. 148, no. 7, pp. 153–165, Dec. 2019, doi: 10.13053/rcs-148-7-12.

[7] R. Panda, R. Malheiro, and R. P. Paiva, “Audio Features for Music Emotion Recognition: A Survey,” *IEEE Trans Affect Comput*, vol. 14, no. 1, pp. 68–88, Jan. 2023, doi: 10.1109/TAFFC.2020.3032373.

[8] S. Hizlisoy, S. Yildirim, and Z. Tufekci, “Music emotion recognition using convolutional long short term memory deep neural networks,” *Engineering Science and Technology, an International Journal*, vol. 24, no. 3, pp. 760–767, Jun. 2021, doi: 10.1016/j.jestch.2020.10.009.

[9] D. Han, Y. Kong, J. Han, and G. Wang, “A survey of music emotion recognition,” *Front Comput Sci*, vol. 16, no. 6, p. 166335, Dec. 2022, doi: 10.1007/s11704-021-0569-4.

[10] A. S. Bhat, V. S. Amith, N. S. Prasad, and D. M. Mohan, “An Efficient Classification Algorithm for Music Mood Detection in Western and Hindi Music Using Audio Feature Extraction,” in *2014 Fifth International Conference on Signal and Image Processing*, IEEE, Jan. 2014, pp. 359–364. doi: 10.1109/ICSIP.2014.63.

[11] Lie Lu, D. Liu, and Hong-Jiang Zhang, “Automatic mood detection and tracking of music audio signals,” *IEEE Trans Audio Speech Lang Process*, vol. 14, no. 1, pp. 5–18, Jan. 2006, doi: 10.1109/TSA.2005.860344.

[12] X. Mao, L. Chen, and L. Fu, “Multi-level Speech Emotion Recognition Based on HMM and ANN,” in *2009 WRI World Congress on Computer Science and Information Engineering*, IEEE, 2009, pp. 225–229. doi: 10.1109/CSIE.2009.113.

[13] Y.-P. Lin, C.-H. Wang, T.-L. Wu, S.-K. Jeng, and J.-H. Chen, “EEG-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Apr. 2009, pp. 489–492. doi: 10.1109/ICASSP.2009.4959627.

[14] X. Han, F. Chen, and J. Ban, “Music Emotion Recognition Based on a Neural Network with an Inception-GRU Residual Structure,” *Electronics (Basel)*, vol. 12, no. 4, p. 978, Feb. 2023, doi: 10.3390/electronics12040978.

[15] J. Grekow, “Music emotion recognition using recurrent neural networks and pretrained models,” *J Intell Inf Syst*, vol. 57, no. 3, pp. 531–546, Dec. 2021, doi: 10.1007/s10844-021-00658-5.

[16] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, and R. Jarina, “Stacked Convolutional and Recurrent Neural Networks for Music Emotion Recognition,” Jun. 2017.

[17] X. Jia, “Music Emotion Classification Method Based on Deep Learning and Explicit Sparse Attention Network,” *Comput Intell Neurosci*, vol. 2022, pp. 1–9, Jun. 2022, doi: 10.1155/2022/3920663.

[18] X. Jia, “Music Emotion Classification Method Based on Deep Learning and Improved Attention Mechanism,” *Comput Intell Neurosci*, vol. 2022, pp. 1–8, Jun. 2022, doi: 10.1155/2022/5181899.

[19] R. Orjesek, R. Jarina, M. Chmulik, and M. Kuba, “DNN Based Music Emotion Recognition from Raw Audio Signal,” in *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, IEEE, Apr. 2019, pp. 1–4. doi: 10.1109/RADIOELEK.2019.8733572.

[20] J.-H. Su, T.-P. Hong, Y.-H. Hsieh, and S.-M. Li, “Effective Music Emotion Recognition by Segment-based Progressive Learning,” in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, Oct. 2020, pp. 3072–3076. doi: 10.1109/SMC42975.2020.9283500.

[21] P. S. Lokhande and B. S. Tiple, “A framework for emotion identification in music: Deep learning approach,” in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, Jun. 2017, pp. 262–266. doi: 10.1109/ICCONS.2017.8250723.

[22] P. Álvarez, J. García de Quirós, and S. Baldassarri, “RIADA: A Machine-Learning Based Infrastructure for Recognising the Emotions of Spotify Songs,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. InPress, no. InPress, p. 1, 2022, doi: 10.9781/ijimai.2022.04.002.

[23] A. Ayesh, “Emotion-Based Machine Learning,” in *Encyclopedia of the Sciences of Learning*, Boston, MA: Springer US, 2012, pp. 1137–1139. doi: 10.1007/978-1-4419-1428-6_18.

[24] UWA, “The Science Of Emotion: Exploring The Basics Of Emotional Psychology,” UWA Online.

[25] Batta Mahesh, “Machine Learning Algorithms -A Review,” *ResearchGate*, vol. 1, pp. 1–6, Jan. 2019.

[26] MatLab, “¿Qué es Machine Learning?,” MathWorks. Accessed: Jan. 17, 2024. [Online]. Available: https://la.mathworks.com/discovery/machine-learning.html

[27] Papers with code, “Emotion Classification,” Papers with code.

[28] Saranga Mahanta and Arvindpdmn, “Audio Feature Extraction,” devopedia.

[29] deepchecks, “One-hot encoding,” deepchecks. Accessed: Jan. 17, 2024. [Online]. Available: https://deepchecks.com/glossary/one-hot-encoding/#:~:text=One%2Dhot%20encoding%20in%20machine,categorical%20data%20in%20machine%20learning

[30] Sakshi Khanna, “A Comprehensive Guide to Train-Test-Validation Split in 2023,” Analytics Vidhya. Accessed: Jan. 17, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2023/11/train-test-validation-split/

[31] Ayush Gupta, “A Comprehensive Guide on Optimizers in Deep Learning,” Analytics Vidhya.

[32] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, “Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network,” *Applied Sciences*, vol. 13, no. 8, p. 4750, Apr. 2023, doi: 10.3390/app13084750.

[33] Tarang Shah, “About Train, Validation and Test Sets in Machine Learning,” Towards Data Science. Accessed: Jan. 17, 2024. [Online]. Available: https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

[34] George Lawton, “data preprocessing,” Data Management. Accessed: Jan. 17, 2024. [Online]. Available: https://www.techtarget.com/searchdatamanagement/definition/raw-data

[35] Gavin Wright, “raw data (source data or atomic data),” Data Management. Accessed: Jan. 17, 2024. [Online]. Available: https://www.techtarget.com/searchdatamanagement/definition/raw-data

[36] Google, “Clasificación: Verdadero o falso y positivo o negativo,” Google Machine Learning.

[37] SoundBridge, “Time Stretching,” SoundBridge.

[38] Pouya Hallaj, “Data Augmentation: Benefits and Disadvantages,” *Medium*, vol. 1, pp. 1–1, Sep. 2023, Accessed: Feb. 13, 2024. [Online]. Available: https://medium.com/@pouyahallaj/data-augmentation-benefits-and-disadvantages-38d8201aead

[39] Pier Paolo Ippolito, “Supercharging Your Machine Learning Models with Data Augmentation,” *Kili*, vol. 1, pp. 1–1, Dec. 2022, Accessed: Feb. 13, 2024. [Online]. Available: https://kili-technology.com/data-labeling/machine-learning/data-augmentation-guide#103

[40] R. Sarkar, S. Choudhury, S. Dutta, A. Roy, and S. K. Saha, “Recognition of emotion in music based on deep convolutional neural network,” *Multimed Tools Appl*, vol. 79, no. 1–2, pp. 765–783, Jan. 2020, doi: 10.1007/s11042-019-08192-x.

[41] P.-T. Yang, S.-M. Kuang, C.-C. Wu, and J.-L. Hsu, “Predicting Music Emotion by Using Convolutional Neural Network,” 2020, pp. 266–275. doi: 10.1007/978-3-030-50341-3_21.

[42] Edward Ma, “Data Augmentation for Audio,” *Medium*, vol. 1, pp. 1–1, Jun. 2019, Accessed: Feb. 13, 2024. [Online]. Available: https://medium.com/@makcedward/data-augmentation-for-audio-76912b01fdf6

[43] Ana Parra Galindo, “Reconocimiento automático de emociones en la música utilizando aprendizaje máquina,” Escuela Técnica Superior de Ingeniería Universidad de Sevilla, Sevilla, 2021. Accessed: Feb. 13, 2024. [Online]. Available: https://idus.us.es/bitstream/handle/11441/129092/TFG-3877-PARRA%20GALINDO.pdf?sequence=1&isAllowed=y

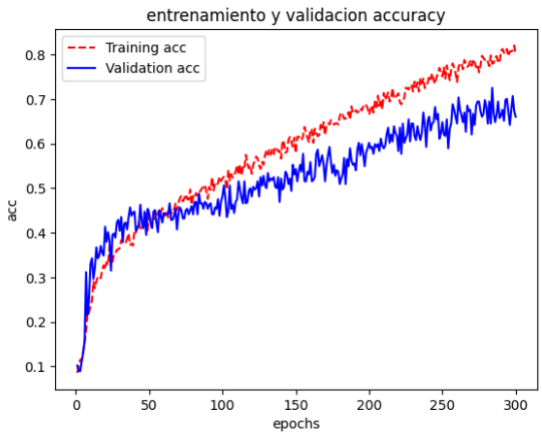
[44] R. Panda, R. Malheiro, and R. P. Paiva, “Novel Audio Features for Music Emotion Recognition,” *IEEE Trans Affect Comput*, vol. 11, no. 4, pp. 614–626, Oct. 2020, doi: 10.1109/TAFFC.2018.2820691.

[45] LabWindows/CVI, “Zero Padding,” ni. Accessed: Feb. 13, 2024. [Online]. Available: https://www.ni.com/docs/en-US/bundle/labwindows-cvi/page/advancedanalysisconcepts/lvac_zero_padding.html#:~:text=Zero%20padding%20is%20a%20technique,next%20higher%20power%20of%20two.

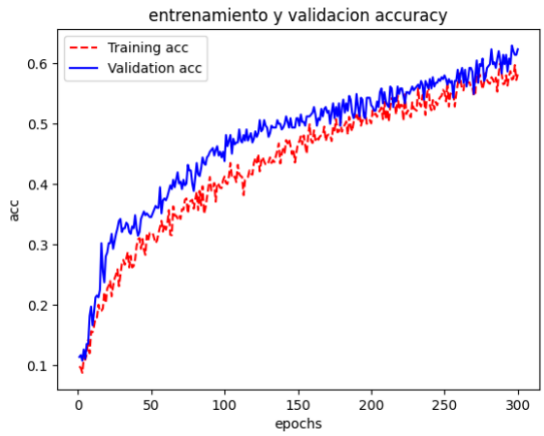
[46] K. Gurney, *An Introduction to Neural Networks*, 1st ed., vol. Gurney, Kevin. London: CRC Press, 2018. doi: 10.1201/9781315273570.

7. **ANEXOS**

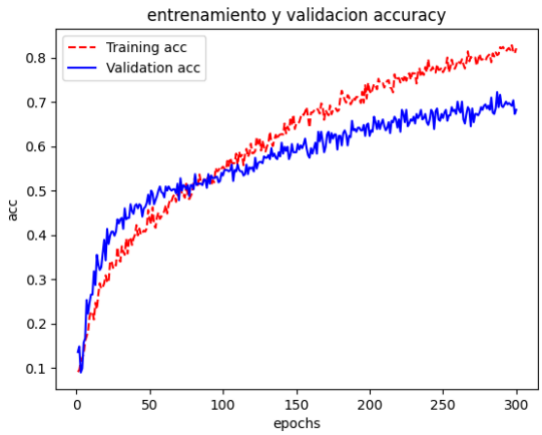
Anexo No. 1 Métricas de las tres configuraciones del modelo 1



Modelo 1 CFG 1

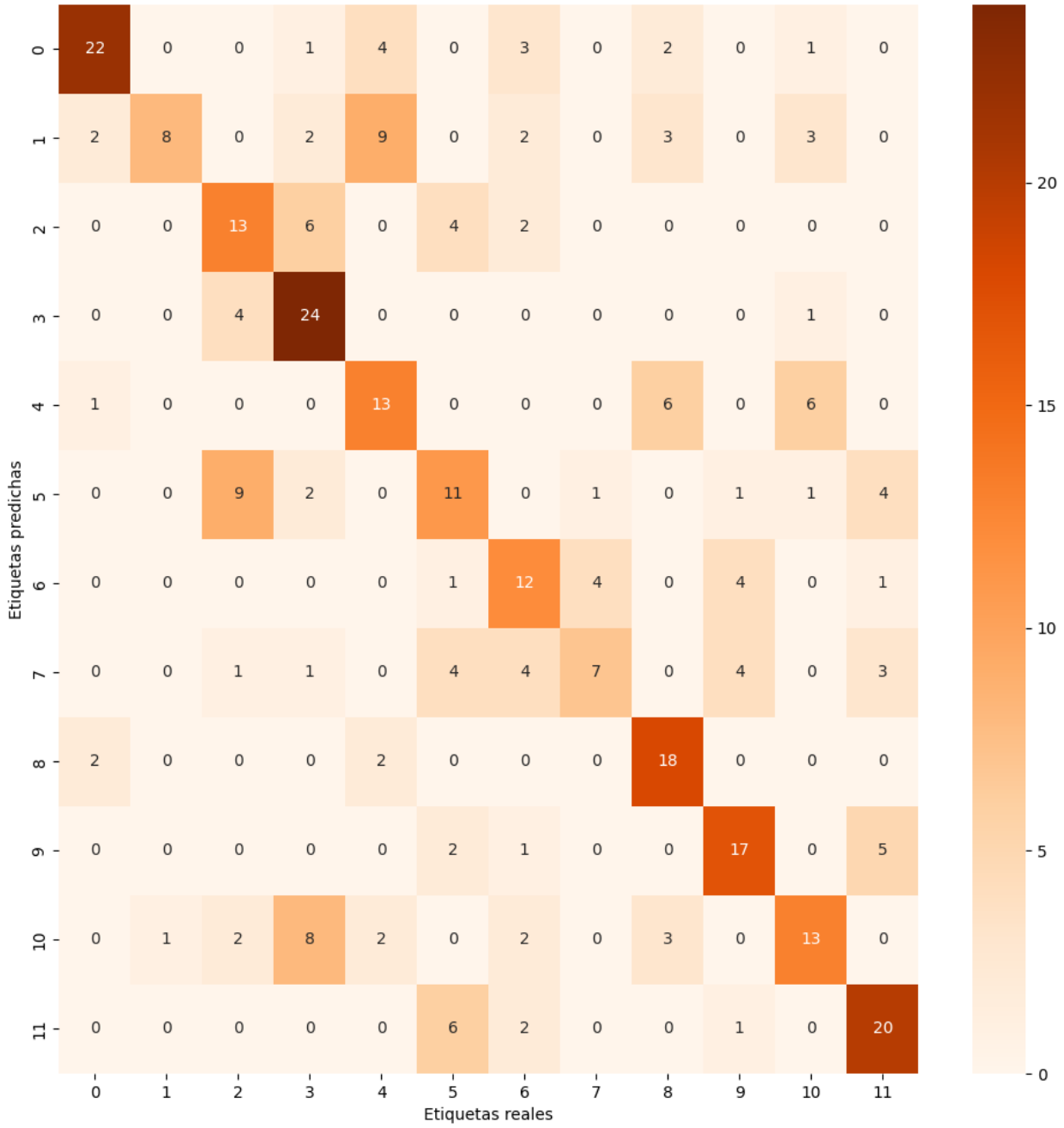


Modelo 1 CFG 2

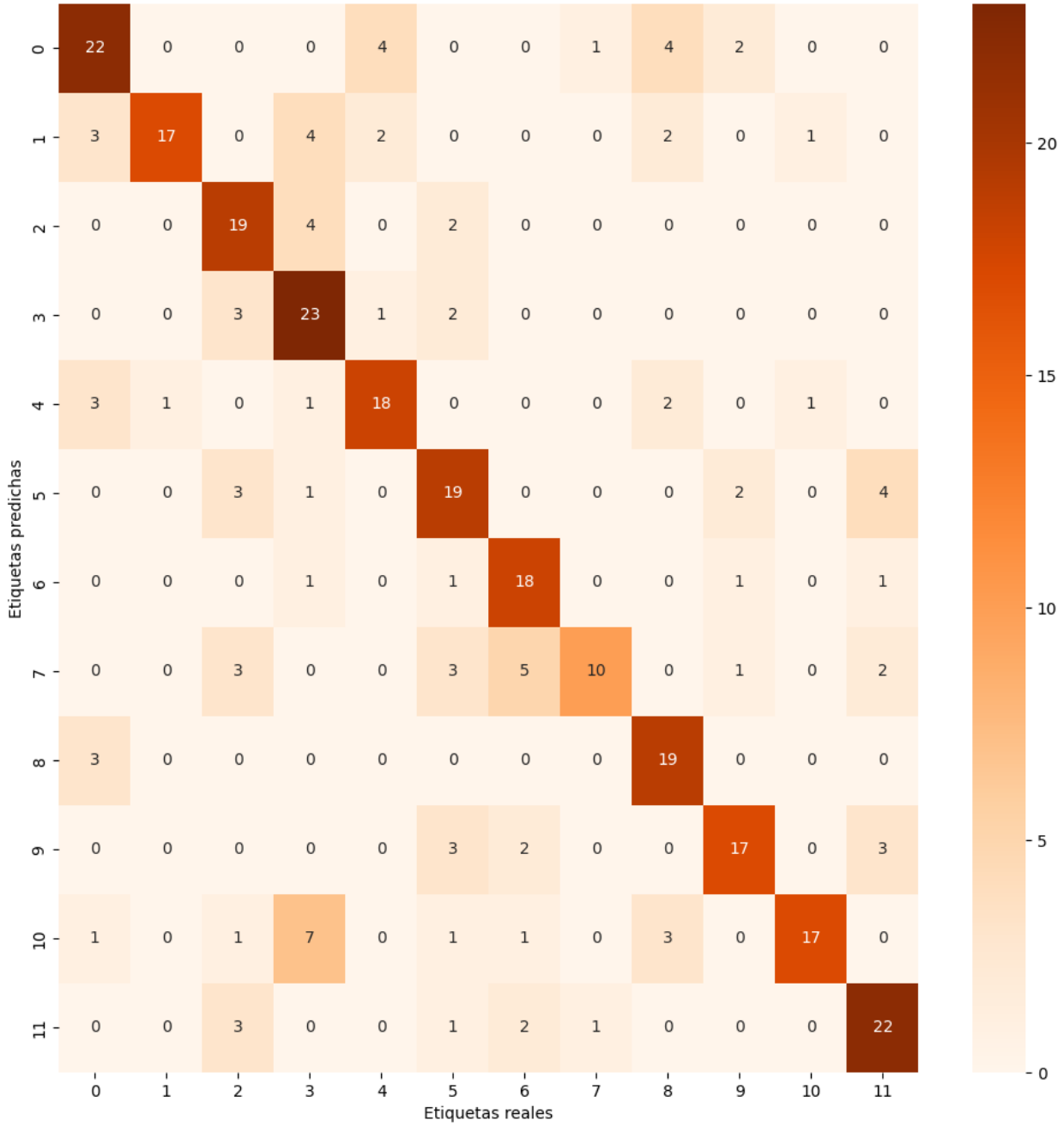


Modelo 1 CFG 3

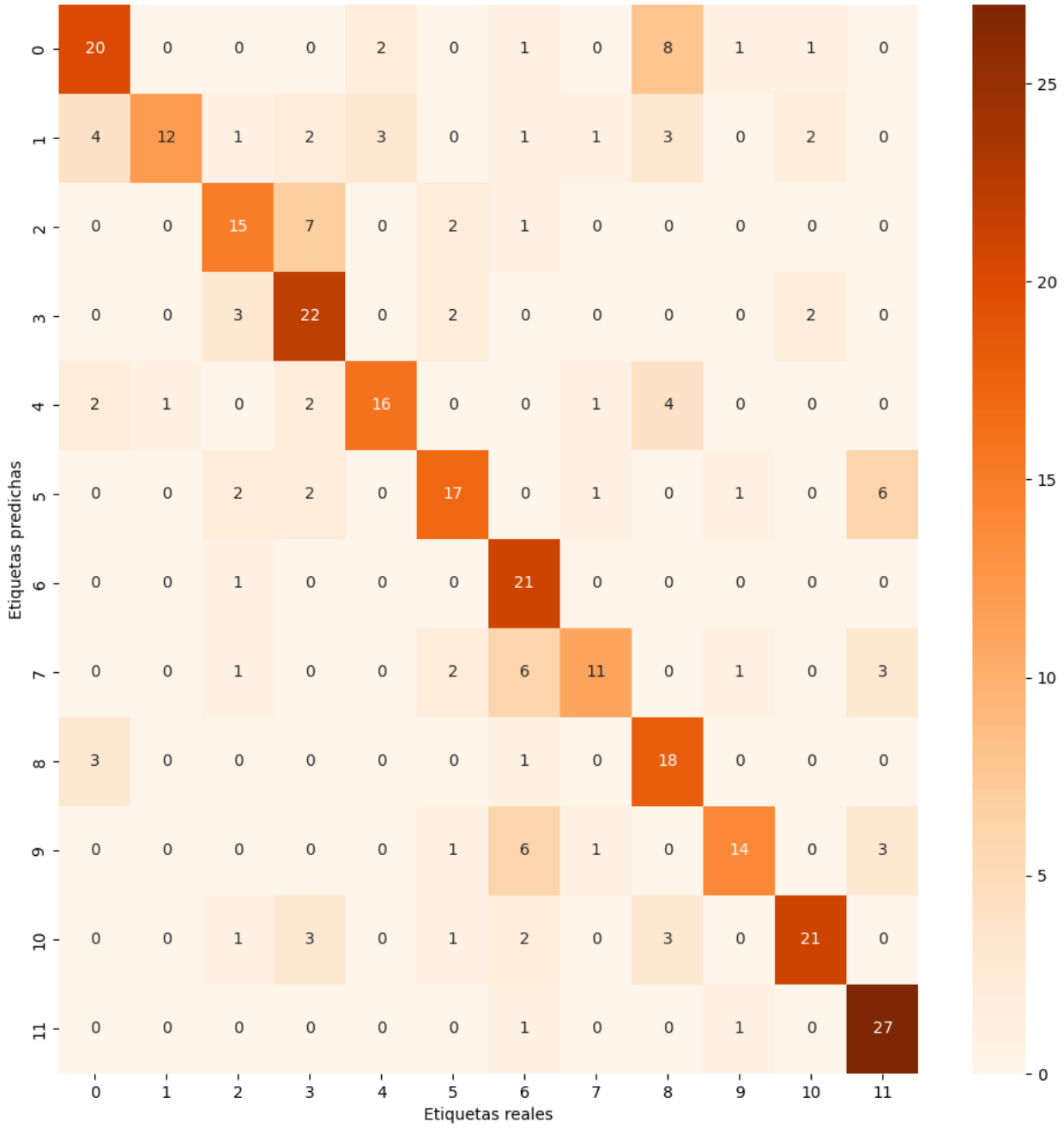
Anexo No. 2 Matriz de confusion CFG 1 modelo 1


Etiquetas predichas
Etiquetas reales
0 1 2 3 4 5 6 7 8 9 10 11
22 0 0 1 4 0 3 0 2 0 1 0
2 8 0 2 9 0 2 0 3 0 3 0
0 0 13 6 0 4 2 0 0 0 0 0
0 0 4 24 0 0 0 0 0 0 1 0
1 0 0 0 13 0 0 0 6 0 6 0
0 0 9 2 0 11 0 1 0 1 1 4
0 0 0 0 0 1 12 4 0 4 0 1
0 0 1 1 0 4 4 7 0 4 0 3
2 0 0 0 2 0 0 0 18 0 0 0
0 0 0 0 0 2 1 0 0 17 0 5
0 1 2 8 2 0 2 0 3 0 13 0
0 0 0 0 0 6 2 0 0 1 0 20
0 5 10 15 20

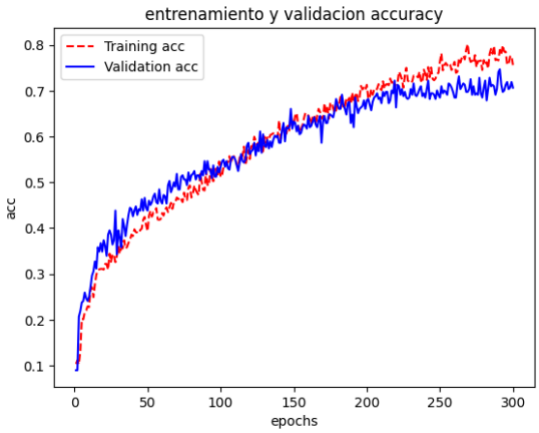
Anexo No. 3 Matriz de confusion modelo 1 CFG 2



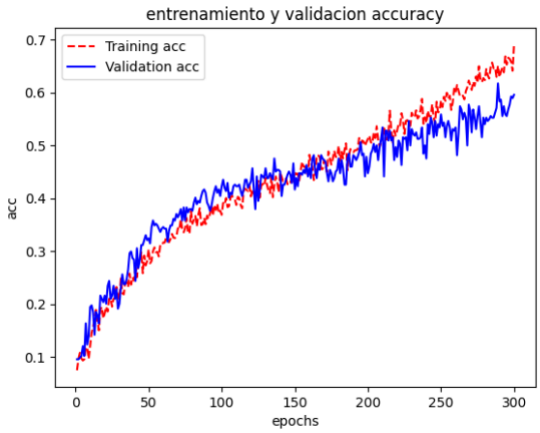
Anexo No. 4 Matriz de confusion modelo 1 CFG 3


Etiquetas predichas
Etiquetas reales
0 1 2 3 4 5 6 7 8 9 10 11
20 0 0 0 2 0 1 0 8 1 1 0
4 12 1 2 3 0 1 1 3 0 2 0
0 0 15 7 0 2 1 0 0 0 0 0
0 0 3 22 0 2 0 0 0 0 2 0
2 1 0 2 16 0 0 1 4 0 0 0
0 0 2 2 0 17 0 1 0 1 0 6
0 0 1 0 0 0 21 0 0 0 0 0
0 0 1 0 0 2 6 11 0 1 0 3
3 0 0 0 0 0 1 0 18 0 0 0
0 0 0 0 0 1 6 1 0 14 0 3
0 0 1 3 0 1 2 0 3 0 21 0
0 0 0 0 0 0 1 0 0 1 0 27
0 5 10 15 20 25

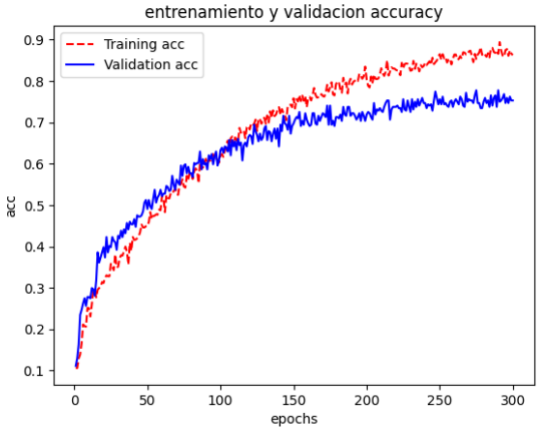
Anexo No. 5 Metricas de entrenamiento y validacion modelo 2



Modelo 2 CFG 1

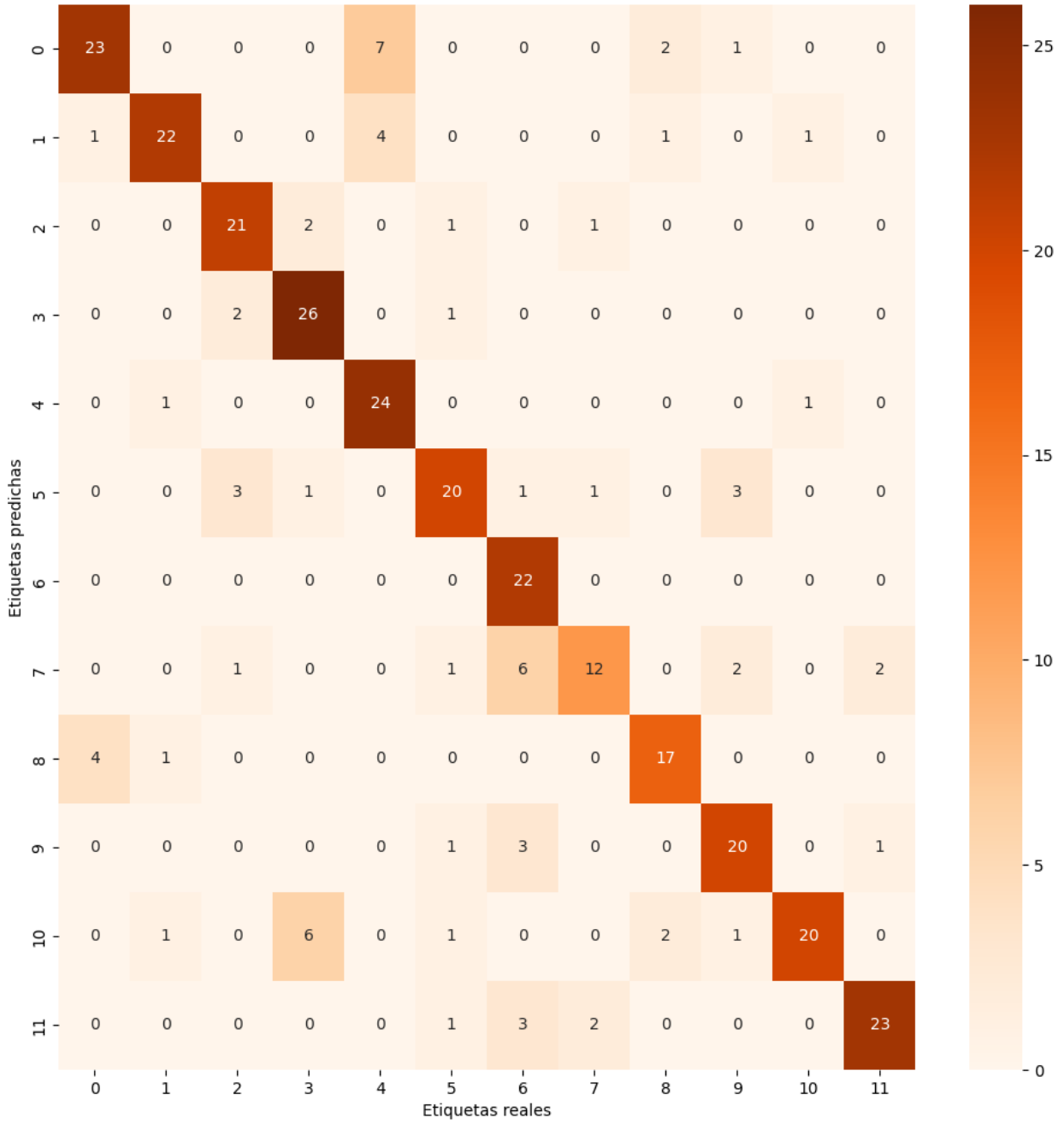


Modelo 2 CFG 2

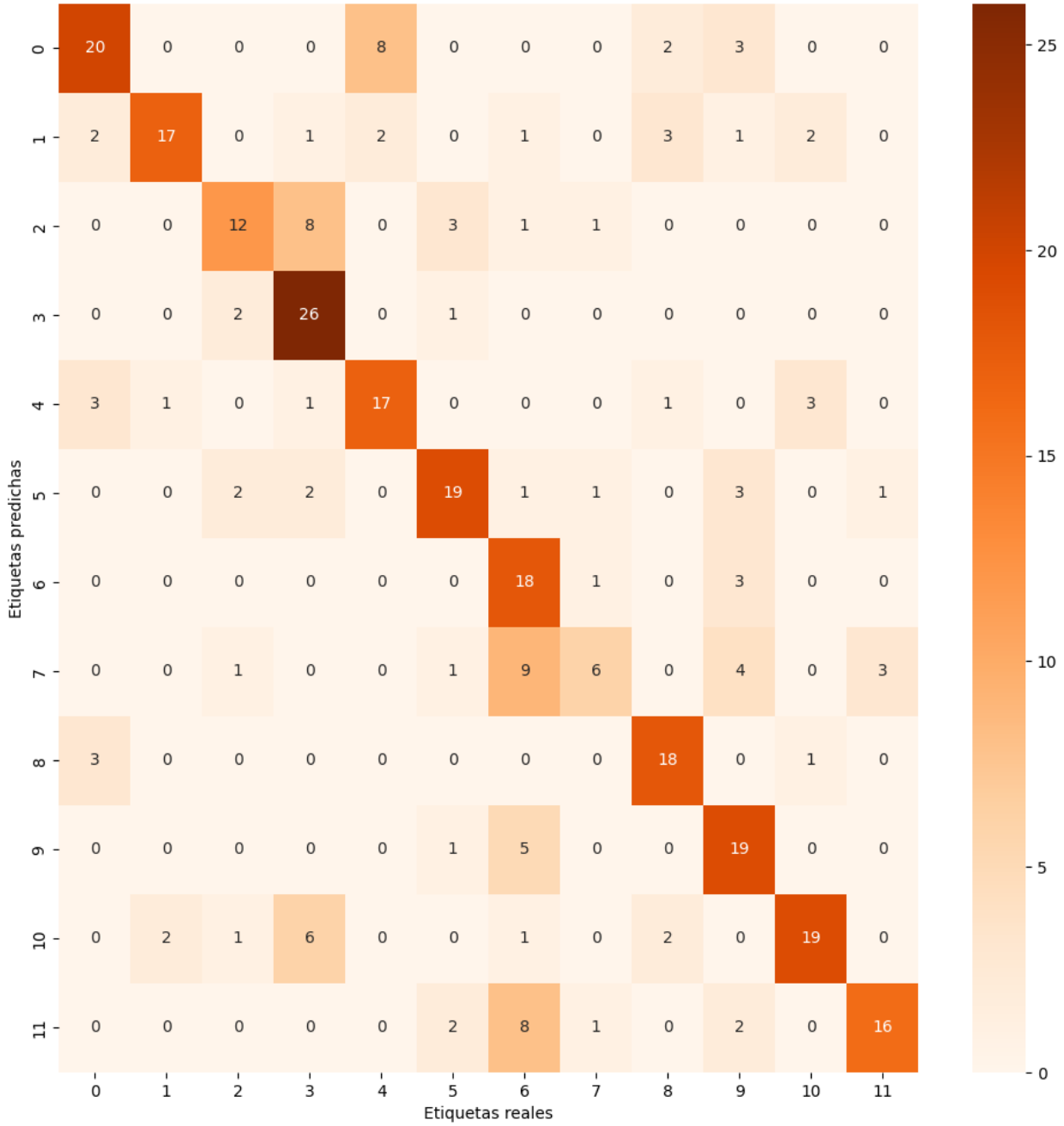


Modelo 2 CFG 3

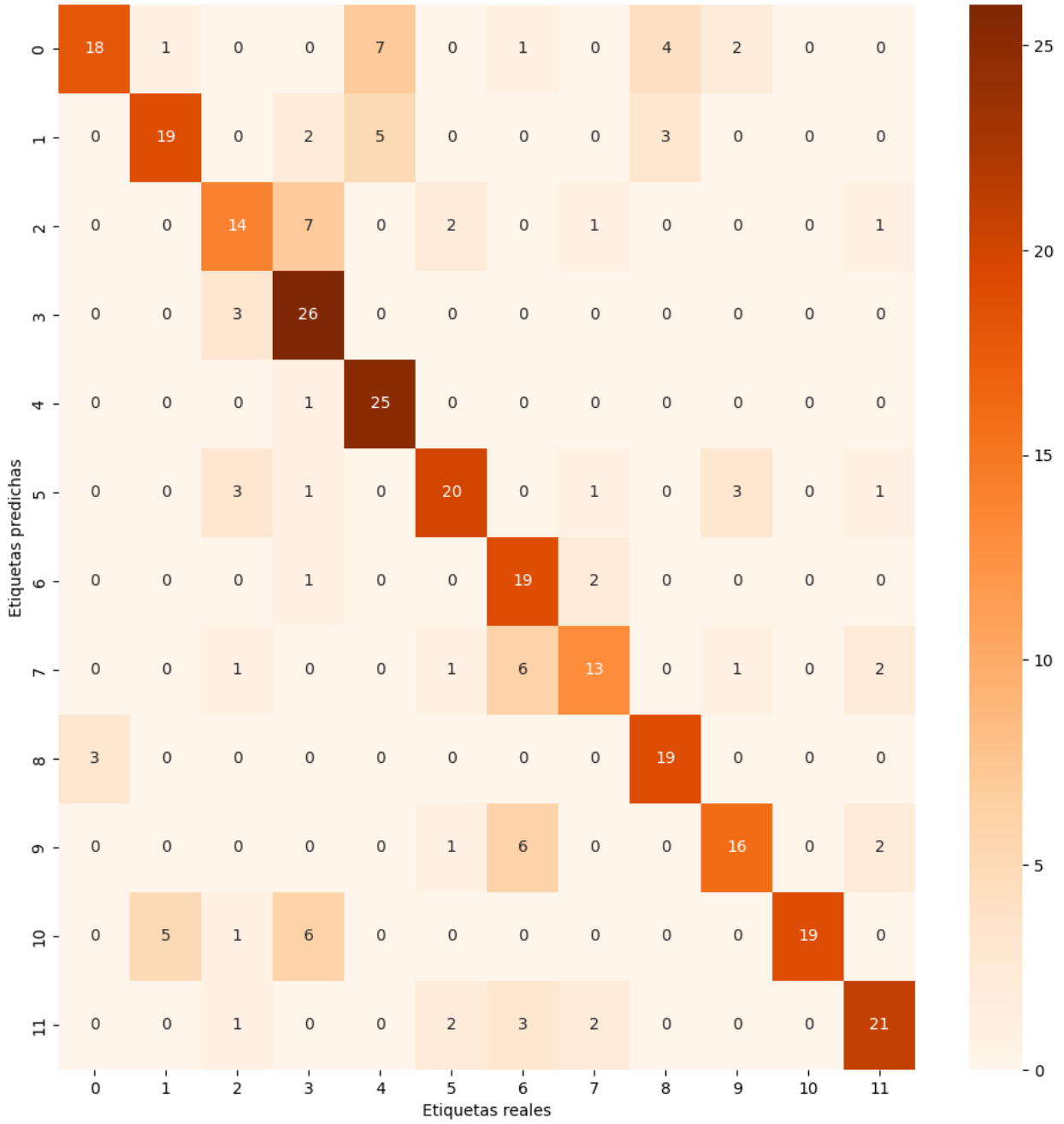
Anexo No. 6 Matriz de confusion modelo 2 CFG 1


Etiquetas predichas
Etiquetas reales
0 1 2 3 4 5 6 7 8 9 10 11
0 23 0 0 0 7 0 0 0 2 1 0 0
1 1 22 0 0 4 0 0 0 1 0 1 0
2 0 0 21 2 0 1 0 1 0 0 0 0
3 0 0 2 26 0 1 0 0 0 0 0 0
4 0 1 0 0 24 0 0 0 0 0 1 0
5 0 0 3 1 0 20 1 1 0 3 0 0
6 0 0 0 0 0 0 22 0 0 0 0 0
7 0 0 1 0 0 1 6 12 0 2 0 2
8 4 1 0 0 0 0 0 0 17 0 0 0
9 0 0 0 0 0 1 3 0 0 20 0 1
10 0 1 0 6 0 1 0 0 2 1 20 0
11 0 0 0 0 0 1 3 2 0 0 0 23
25
20
15
10
5
0

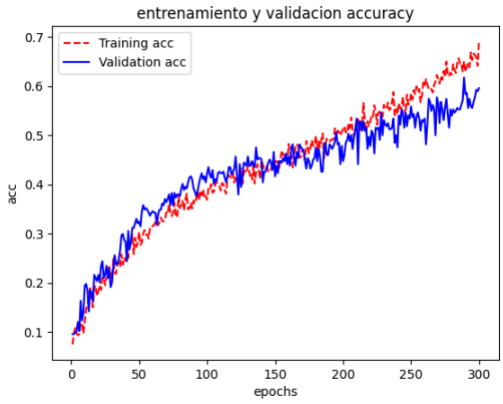
Anexo No. 7 Matriz de confusión modelo 2 CFG 2



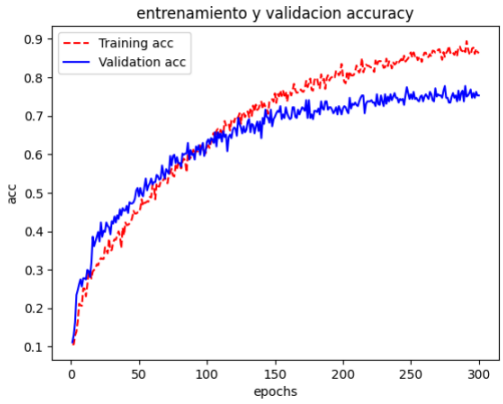
Anexo No. 8 Matriz de confusion modelo 2 CFG 3


Etiquetas predichas
Etiquetas reales
0 1 2 3 4 5 6 7 8 9 10 11
0 18 1 0 0 7 0 1 0 4 2 0 0
1 0 19 0 2 5 0 0 0 3 0 0 0
2 0 0 14 7 0 2 0 1 0 0 0 1
3 0 0 3 26 0 0 0 0 0 0 0 0
4 0 0 0 1 25 0 0 0 0 0 0 0
5 0 0 3 1 0 20 0 1 0 3 0 1
6 0 0 0 1 0 0 19 2 0 0 0 0
7 0 0 1 0 0 1 6 13 0 1 0 2
8 3 0 0 0 0 0 0 0 19 0 0 0
9 0 0 0 0 0 1 6 0 0 16 0 2
10 0 5 1 6 0 0 0 0 0 0 19 0
11 0 0 1 0 0 2 3 2 0 0 0 21
0 5 10 15 20 25

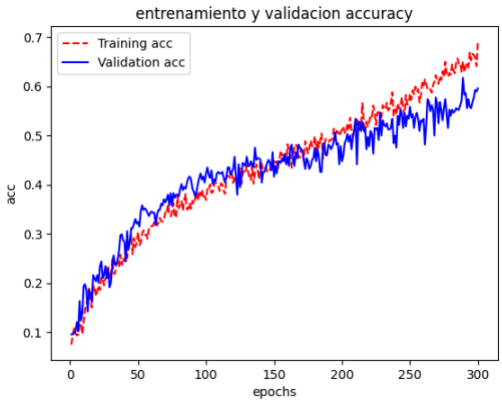
Anexo No. 9 Metricas de entrenamiento y validacion modelo 3



Modelo 3 CFG 1

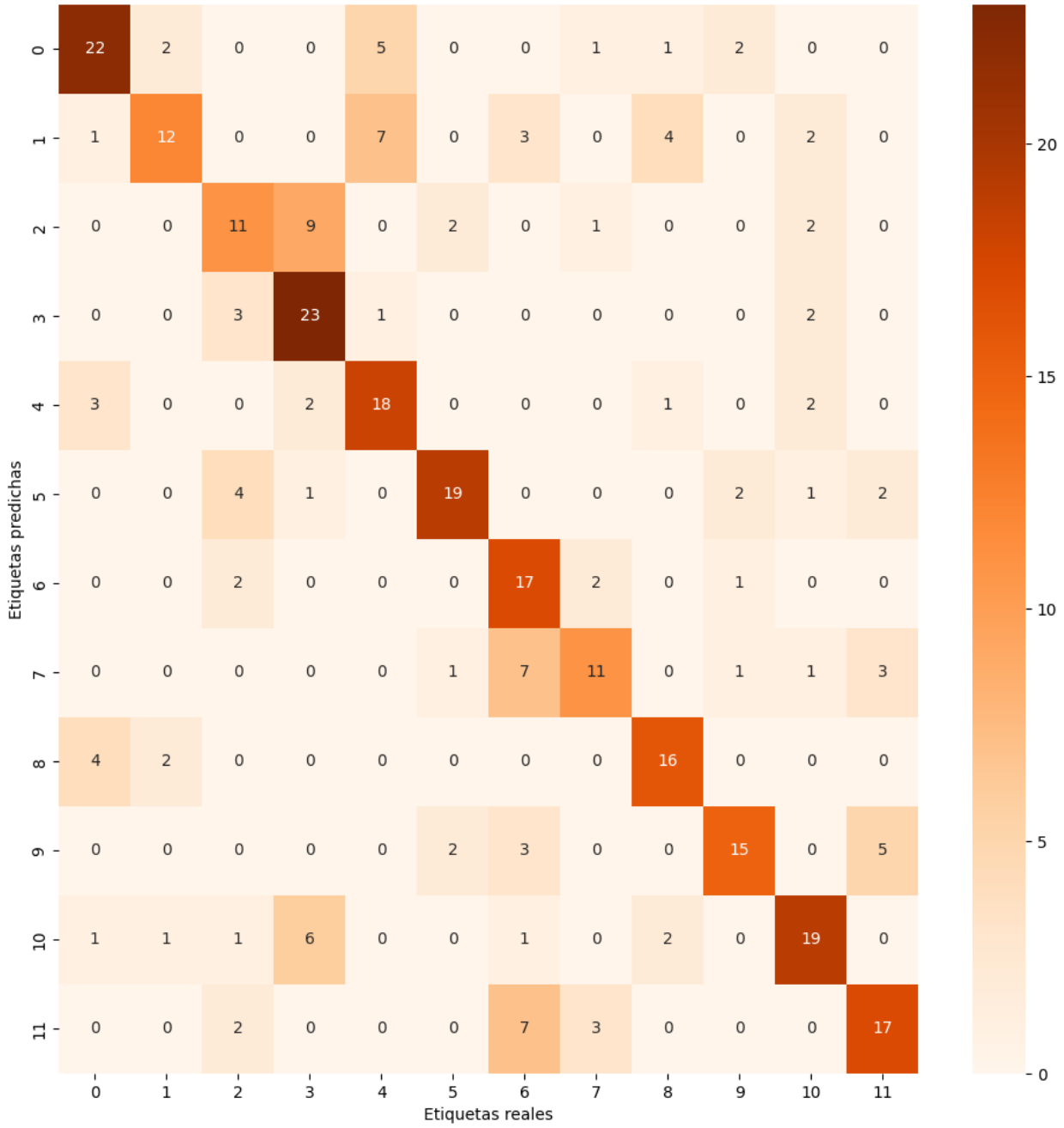


Modelo 3 CFG 2

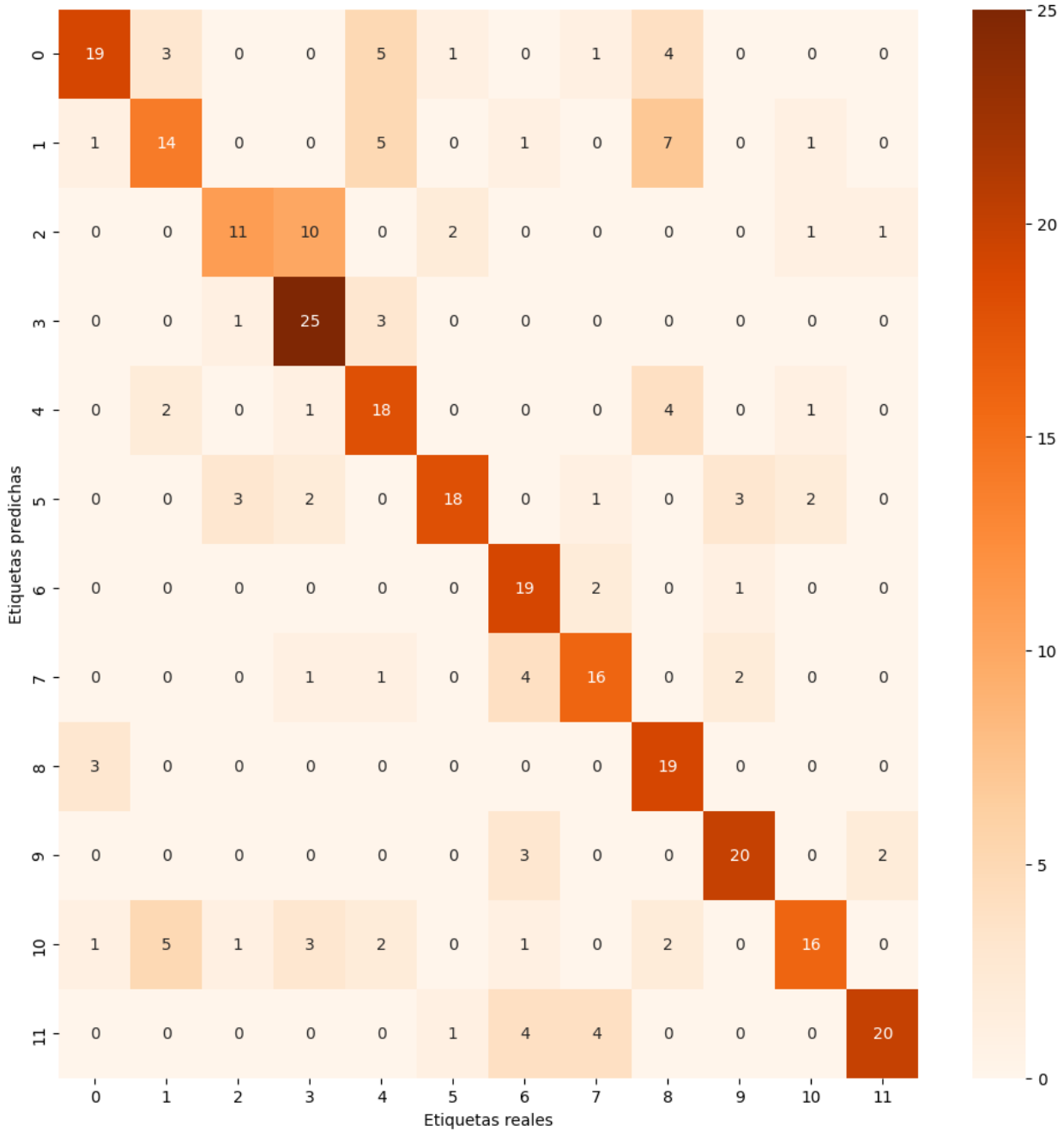


Modelo 3 CFG 3

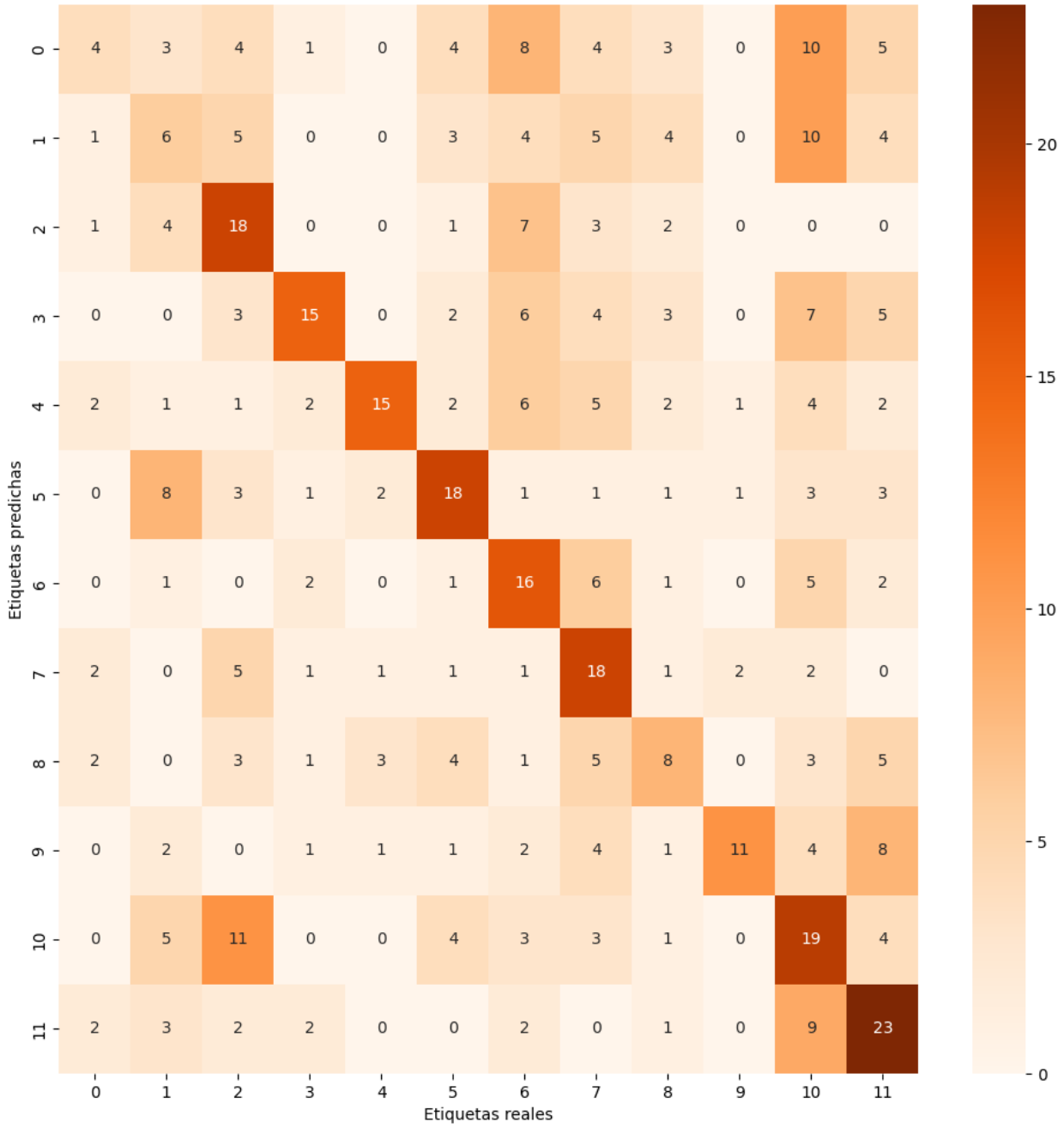
Anexo No. 10 Matriz de confusion modelo 3 CFG 1


Etiquetas predichas
Etiquetas reales
0 1 2 3 4 5 6 7 8 9 10 11
22 2 0 0 5 0 0 1 1 2 0 0
1 12 0 0 7 0 3 0 4 0 2 0
0 0 11 9 0 2 0 1 0 0 2 0
0 0 3 23 1 0 0 0 0 0 2 0
3 0 0 2 18 0 0 0 1 0 2 0
0 0 4 1 0 19 0 0 0 2 1 2
0 0 2 0 0 0 17 2 0 1 0 0
0 0 0 0 0 1 7 11 0 1 1 3
4 2 0 0 0 0 0 0 16 0 0 0
0 0 0 0 0 2 3 0 0 15 0 5
1 1 1 6 0 0 1 0 2 0 19 0
0 0 2 0 0 0 7 3 0 0 0 17
0 5 10 15 20

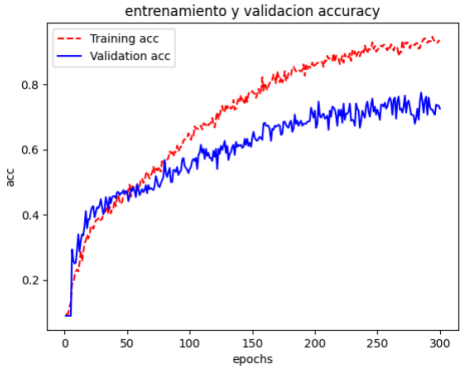
Anexo No. 11 Matriz de confusion modelo 3 CFG 2



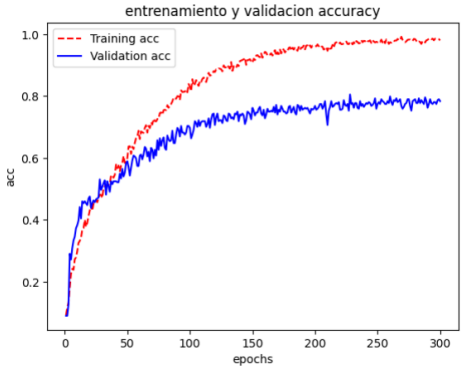
Anexo No. 12 Matriz de confusion modelo 3 CFG 3


Etiquetas predichas
Etiquetas reales
0 1 2 3 4 5 6 7 8 9 10 11
0 4 3 4 1 0 4 8 4 3 0 10 5
1 1 6 5 0 0 3 4 5 4 0 10 4
2 1 4 18 0 0 1 7 3 2 0 0 0
3 0 0 3 15 0 2 6 4 3 0 7 5
4 2 1 1 2 15 2 6 5 2 1 4 2
5 0 8 3 1 2 18 1 1 1 1 3 3
6 0 1 0 2 0 1 16 6 1 0 5 2
7 2 0 5 1 1 1 1 18 1 2 2 0
8 2 0 3 1 3 4 1 5 8 0 3 5
9 0 2 0 1 1 1 2 4 1 11 4 8
10 0 5 11 0 0 4 3 3 1 0 19 4
11 2 3 2 2 0 0 2 0 1 0 9 23
0
5
10
15
20

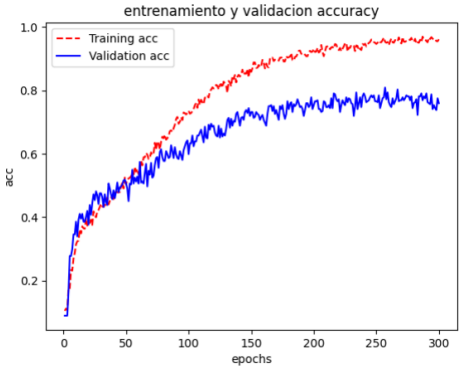
Anexo No. 13 Metricas de entrenamiento de validacion modelo 4



Modelo 4 CFG 1

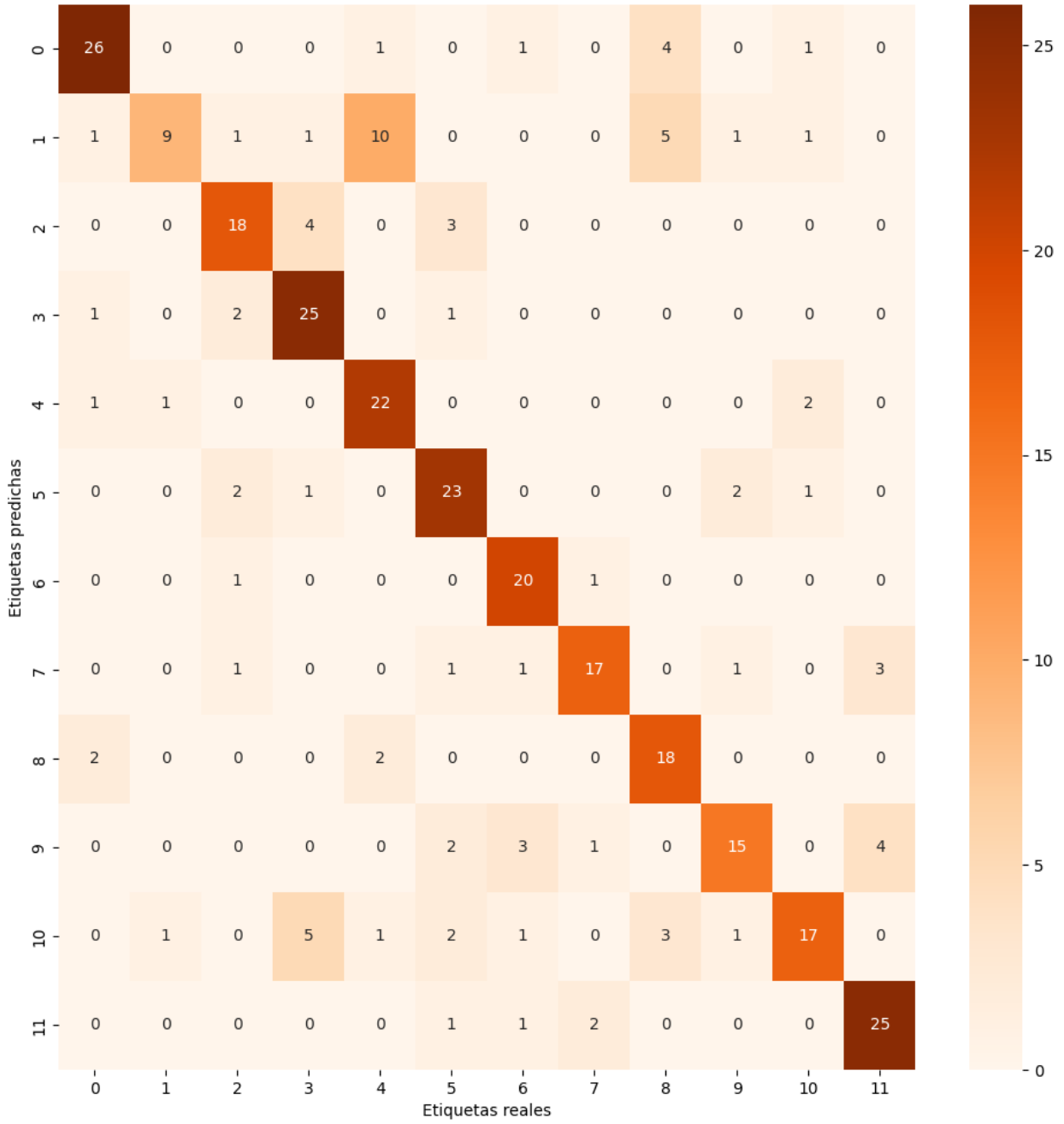


Modelo 4 CFG 2

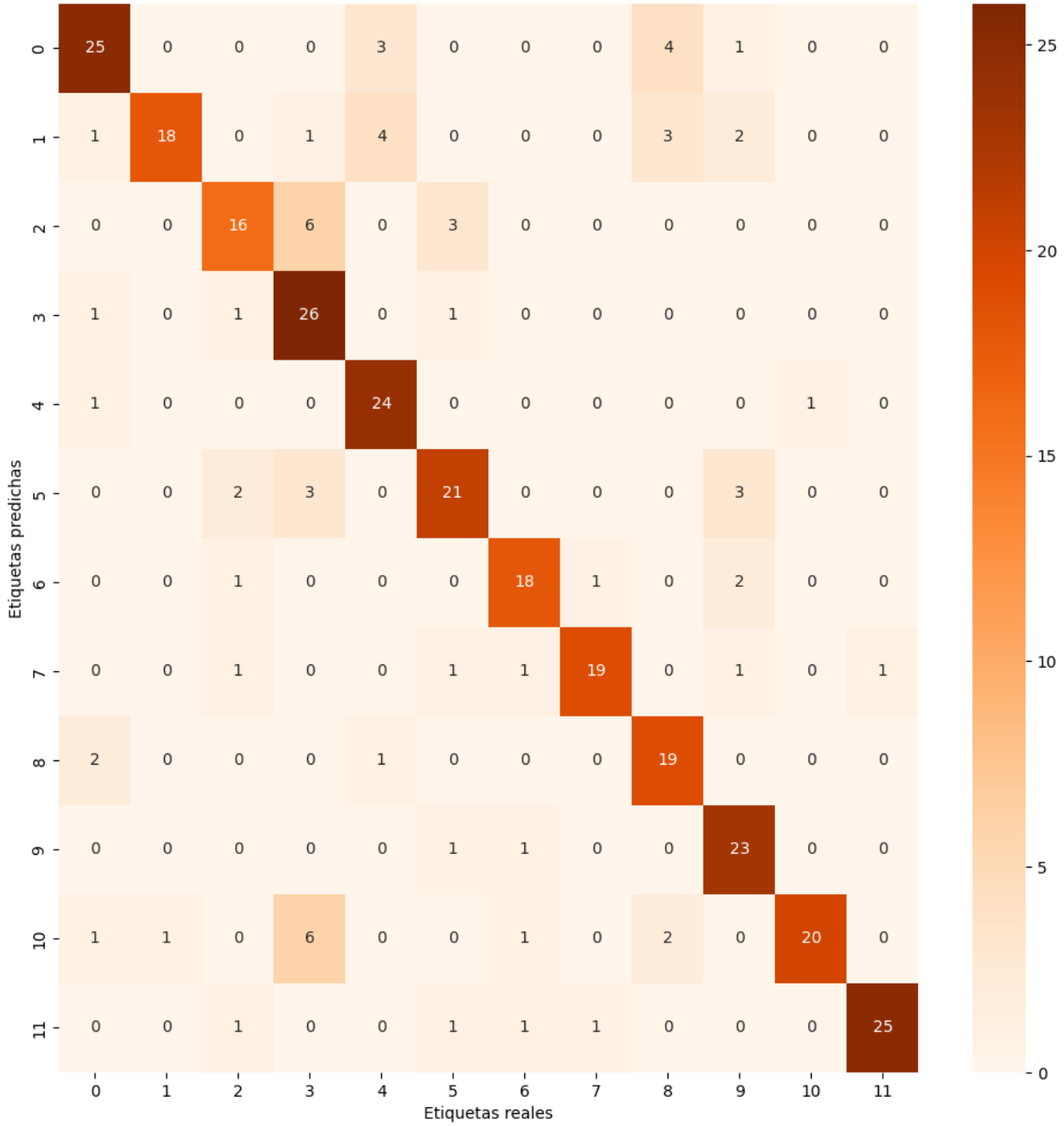


Modelo 4 CFG 3

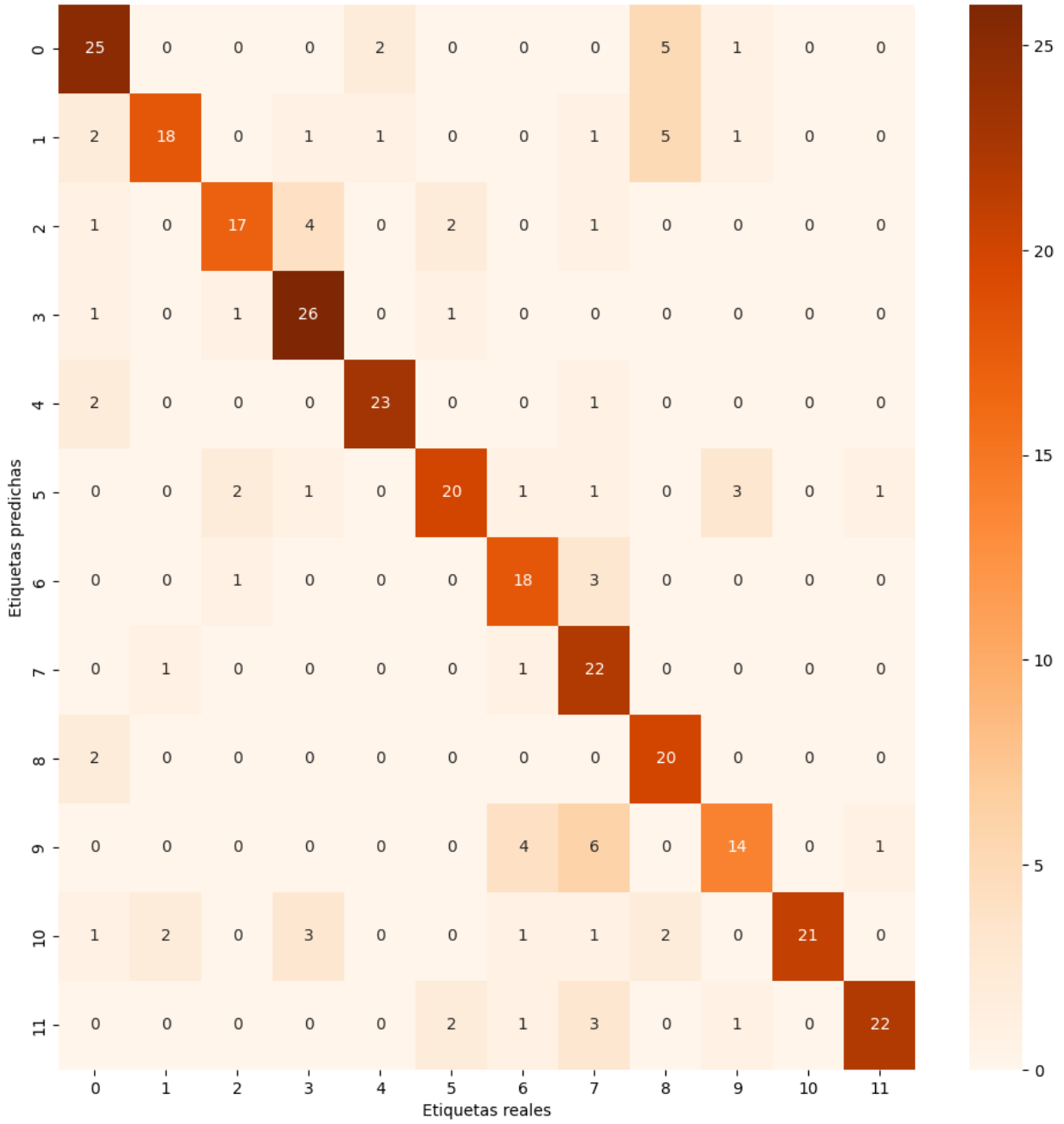
Anexo No. 14 Matriz de confusion modelo 4 CFG 1


Etiquetas predichas
Etiquetas reales
0 1 2 3 4 5 6 7 8 9 10 11
26 0 0 0 1 0 1 0 4 0 1 0
1 9 1 1 10 0 0 0 5 1 1 0
0 0 18 4 0 3 0 0 0 0 0 0
1 0 2 25 0 1 0 0 0 0 0 0
1 1 0 0 22 0 0 0 0 0 2 0
0 0 2 1 0 23 0 0 0 2 1 0
0 0 1 0 0 0 20 1 0 0 0 0
0 0 1 0 0 1 1 17 0 1 0 3
2 0 0 0 2 0 0 0 18 0 0 0
0 0 0 0 0 2 3 1 0 15 0 4
0 1 0 5 1 2 1 0 3 1 17 0
0 0 0 0 0 1 1 2 0 0 0 25
0 5 10 15 20 25

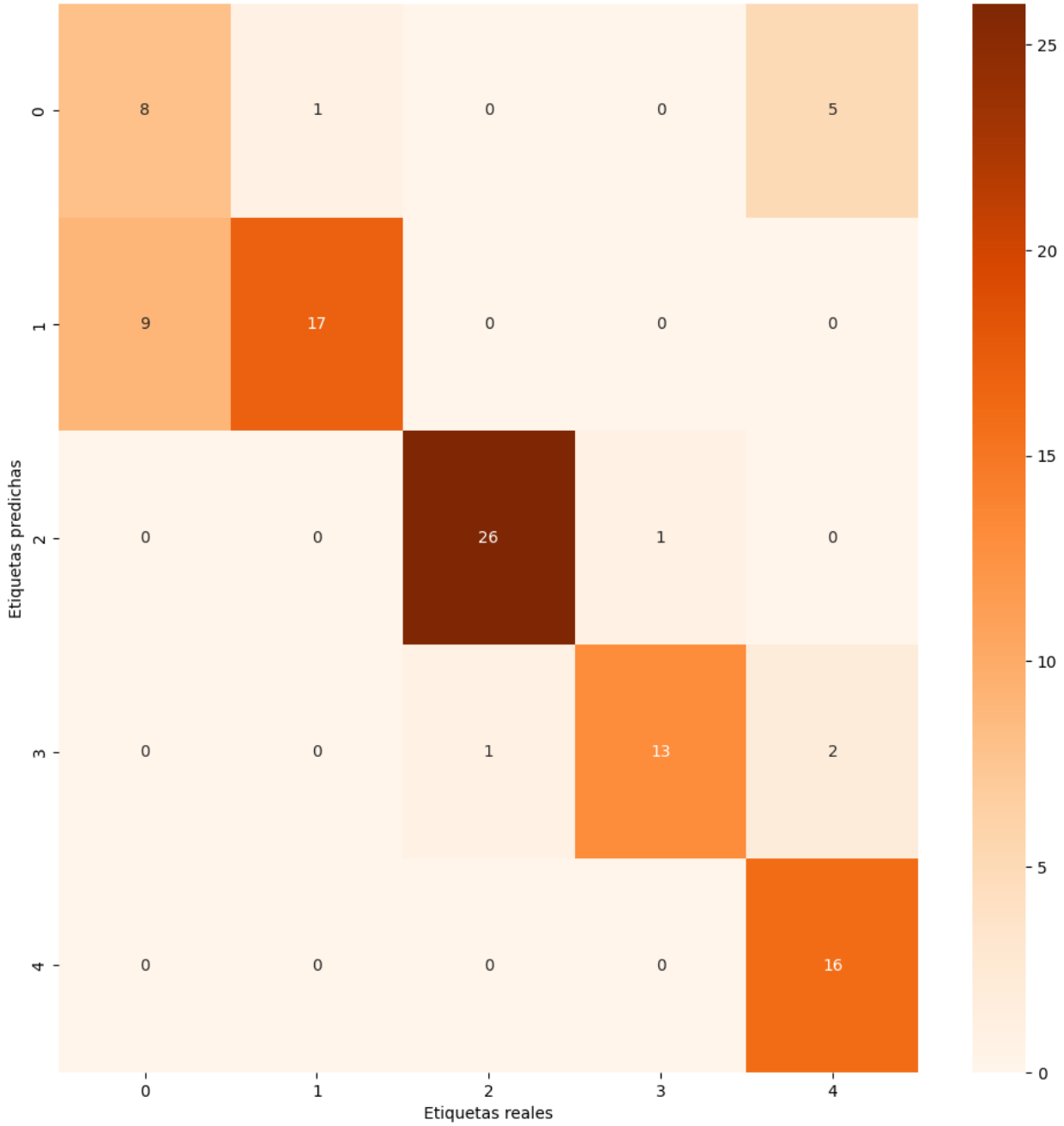
Anexo No. 15 Matriz de confusion modelo 4 CFG 2



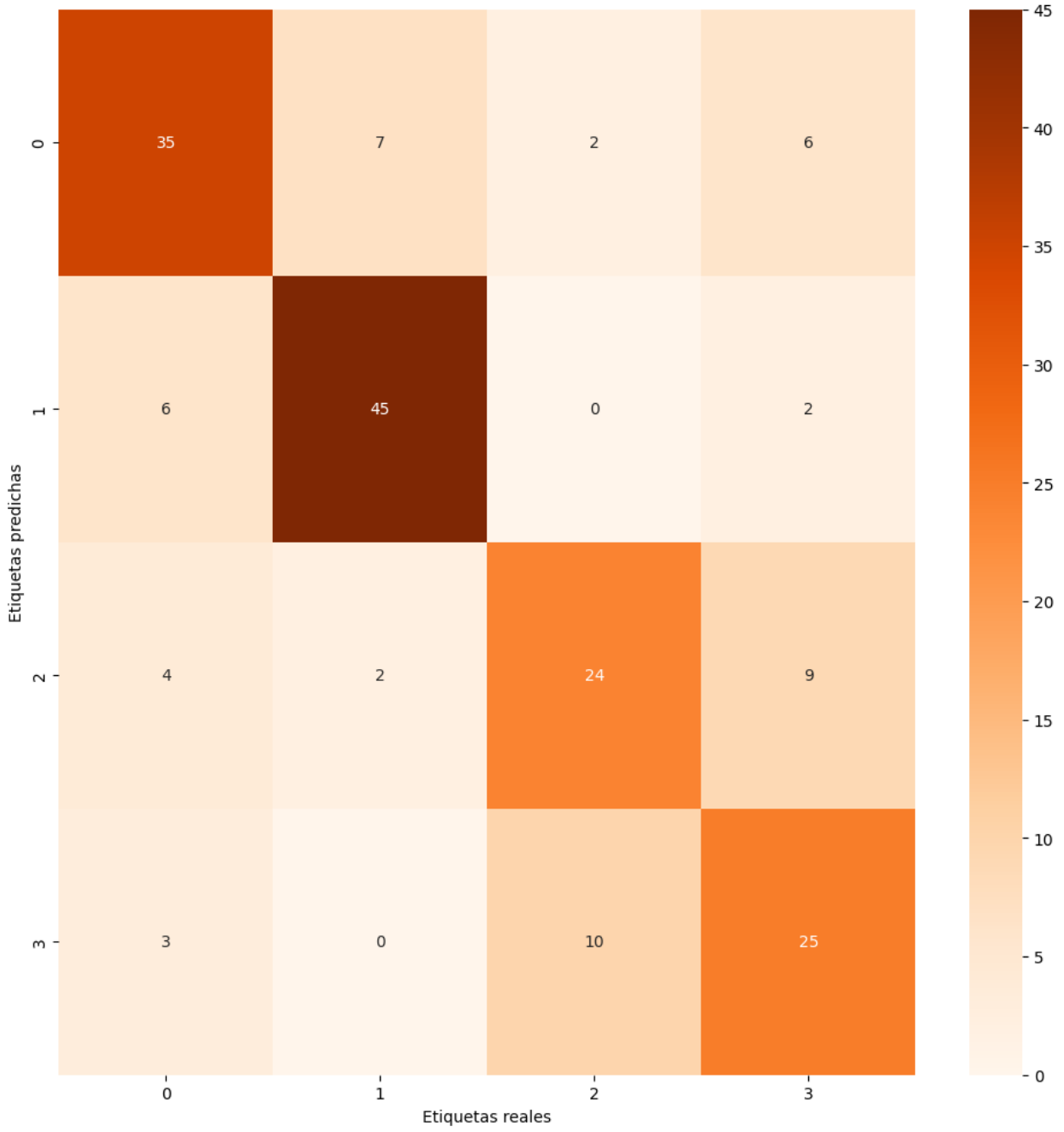
Anexo No. 16 Matriz de confusion modelo 4 CFG 3


Etiquetas predichas
Etiquetas reales
0 1 2 3 4 5 6 7 8 9 10 11
25 0 0 0 2 0 0 0 5 1 0 0
2 18 0 1 1 0 0 1 5 1 0 0
1 0 17 4 0 2 0 1 0 0 0 0
1 0 1 26 0 1 0 0 0 0 0 0
2 0 0 0 23 0 0 1 0 0 0 0
0 0 2 1 0 20 1 1 0 3 0 1
0 0 1 0 0 0 18 3 0 0 0 0
0 1 0 0 0 0 1 22 0 0 0 0
2 0 0 0 0 0 0 0 20 0 0 0
0 0 0 0 0 0 4 6 0 14 0 1
1 2 0 3 0 0 1 1 2 0 21 0
0 0 0 0 0 2 1 3 0 1 0 22
0 5 10 15 20 25

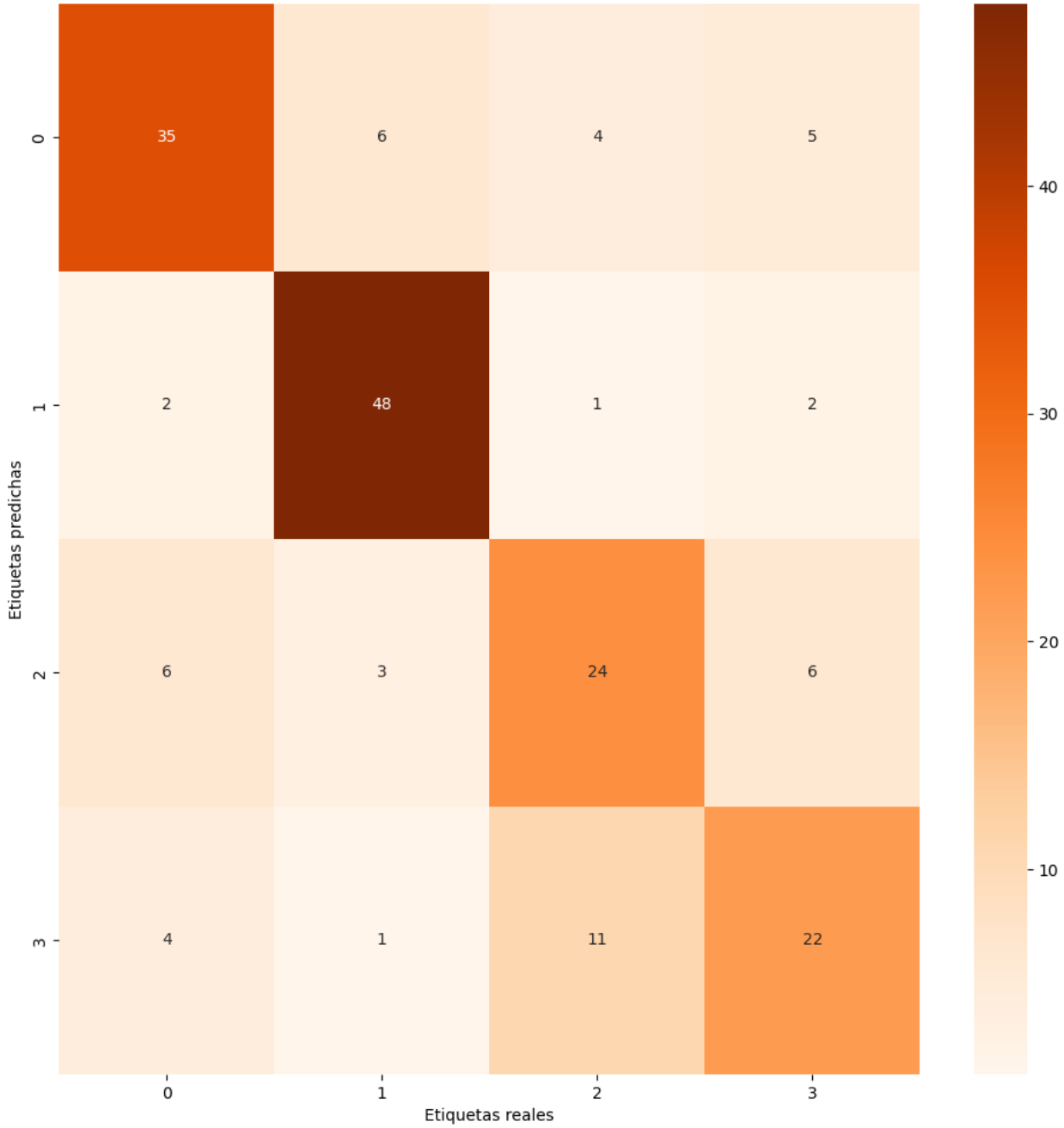
Anexo No. 17 Matriz de confusion hindi Dataset prueba



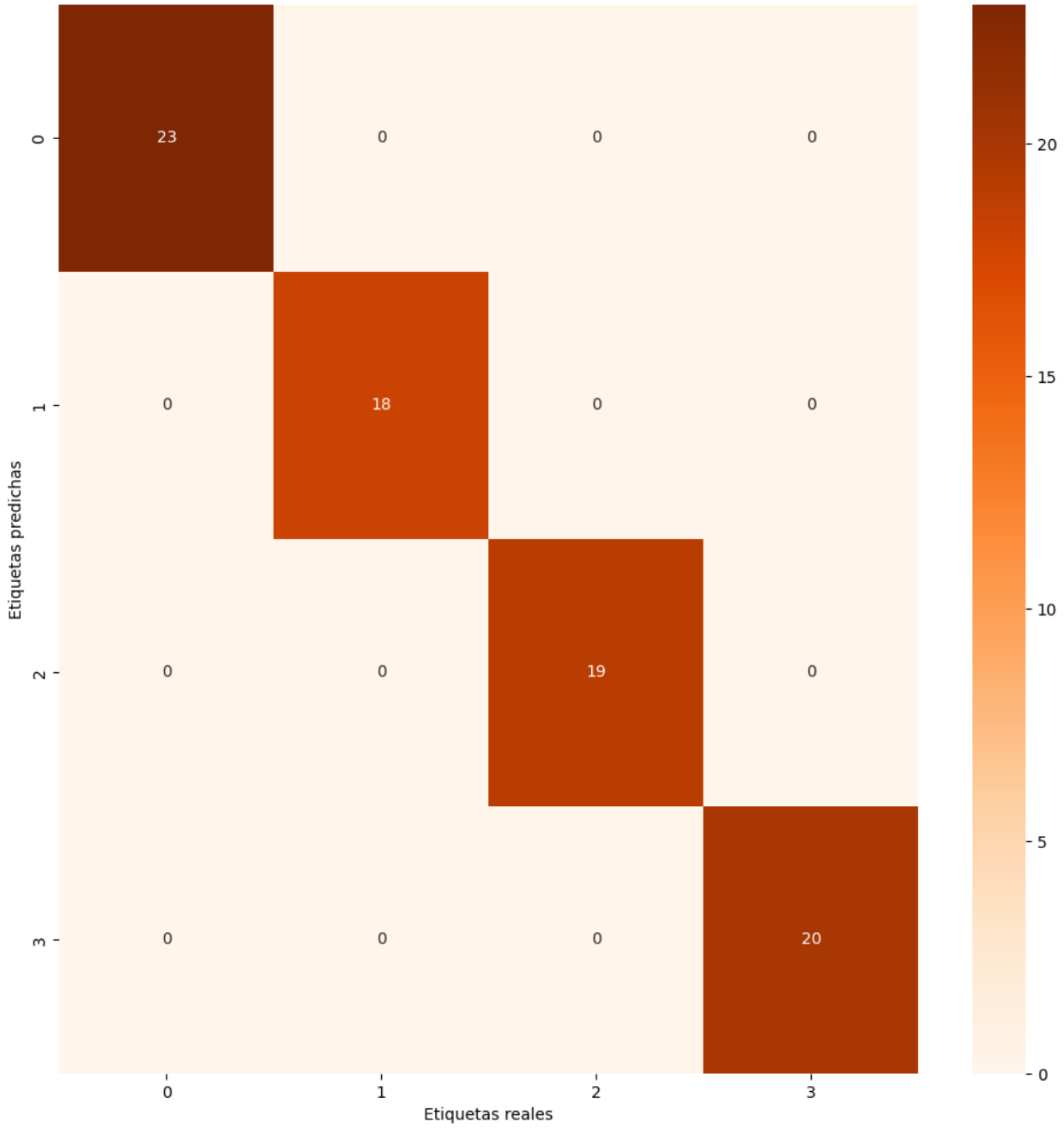
Anexo No. 18 Matriz de confusion Panda dataset prueba


35
7
2
6
6
45
0
2
4
2
24
9
3
0
10
25
0
1
2
3
Etiquetas predichas
Etiquetas reales
45
40
35
30
25
20
15
10
5
0

Anexo No. 19 Matriz de confusion Bi-modal dataset prueba



Anexo No. 20 Matriz de confusion Turkish Dataset prueba


23
0
0
0
0
18
0
0
0
0
19
0
0
0
0
20
0
1
2
3
Etiquetas predichas
Etiquetas reales
20
15
10
5
0