Modelos de clasificación de emociones basados en CNN y ViT

Santiago Ruiz Ramírez

Trabajo de grado

Asesor, docente

Edwin Nelson Montoya Múnera

Universidad EAFIT
Maestría en ciencias de los datos y analítica
Medellín, Colombia
2024

Resumen

El presente proyecto se centra en comparar el rendimiento de modelos de redes neuronales convolucionales (CNN) y transformadores de visión (ViT) para clasificar emociones en imágenes faciales. El problema radica en la precisión de las CNN, que aún enfrenta desafíos, mientras que los ViT han surgido como una alternativa prometedora, destacando la importancia de abordar las emociones en el contexto de la salud mental, ya que estas pueden influir en la capacidad de trabajo creativo y están vinculadas a diferentes afecciones de estudio clínico.
Palabras clave: clasificación de emociones; redes neuronales convolucionales; transformadores de visión; imágenes de rostros; modelos de clasificación; aprendizaje automático.

Abstract

The present project focuses on comparing the performance of convolutional neural network (CNN) and vision transformer (ViT) models to classify emotions in facial images. The problem lies in the accuracy of CNNs, which still faces challenges, while ViTs have emerged as a promising alternative, highlighting the importance of addressing emotions in the context of mental health, as these can influence the ability to creative work and are linked to different clinical study conditions.
Keywords: classification of emotions; convolutional neural networks; vision transformers; images of faces; classification models; machine learning.

Contenido

Índice de figuras

Índice de tablas

1 Introducción

1.1. Planteamiento del problema

El reto respecto al desempeño de las redes neuronales convolucionales (Convolutional Neural Network, CNN por sus siglas) en las tareas de clasificación de imágenes se centra en la precisión, que se relaciona con la percepción de confiabilidad en estos sistemas. Ya que aún con los importantes avances realizados por las redes neuronales convolucionales, todavía enfrentan desafíos para lograr una alta precisión, especialmente en conjuntos de datos complejos o bajo condiciones específicas. Mientras los transformadores de visión (Vision Transformers, ViT por sus siglas) han surgido como una alternativa con un desempeño prometedor para tareas de clasificación de imágenes.

La OMS define el término salud como: “un estado de completo bienestar físico, emocional y social” [Organización Mundial de la Salud, 2023]. Estudios realizados por la Organización Mundial de la Salud, se demuestran que las personas emocionalmente saludables, pueden desarrollar todas sus habilidades, aprender y trabajar adecuadamente, en consecuencia; trabajar por el bienestar emocional de las personas genera un impacto positivo en el desarrollo de sus actividades diarias.

Como se puede observar la salud no solo está relacionada a la ausencia de enfermedades o afecciones físicas; sino que también contempla las emociones como parte del resultado saludable de un individuo.

Estudiar las emociones humanas es un reto cada vez mayor, puesto que las dinámicas sociales tienen cada día más influencia en la percepción del mundo y en los sentimientos de las personas, generando así diferencias en las tomas de decisiones y posibles afectaciones a la salud.

En “Emotion Regulation and Mental Health” [Gross and Muñoz, 1995] se plantea como las emociones y su regulación son factores esenciales en la salud mental. Pudiendo verse la salud mental de los adultos como “la capacidad de trabajar de manera creativa”. Las emociones pueden desencadenar diferentes estados de ánimo entre los que se encuentra incluso la depresión y otras afecciones de estudio clínico.

1.2. Justificación

La mayor parte de los seres humanos son capaces de identificar las emociones que transmiten otros seres a través de sus expresiones faciales. El cerebro humano a lo largo de la historia ha sido la herramienta empleada para dicho procesamiento. Esta capacidad humana es tan evolutiva hasta llegar al punto de parecer una habilidad muy sencilla de realizar por las personas. La restricción de procesamiento inicia cuando el ser humano quiere procesar varias emociones a la vez.

La habilidad humana de procesamiento está relacionada a las técnicas computacionales de reconocimiento de patrones, lo que permite que la máquina aprenda un patrón, lo procese a partir de un conjunto de datos y genere un resultado en corto tiempo. Todo esto se hace posible gracias a técnicas como máquinas de soporte vectorial, Naive Bayes, árboles de decisión y redes neuronales artificiales entre otras. Si bien los humanos son capaces de hacer estos análisis sus tiempos de procesamiento son más lentos, de estabilidad y capacidad limitada, que es donde nace la necesidad de generar herramientas que permitan un procesamiento masivo, confiable y en corto tiempo.

Una herramienta de análisis de imágenes se hace necesaria para muchos campos, entre ellos la medicina y la psicología dado que permite hacer análisis masivos y así generar soluciones médicas de manera más rápida y certera.

Es aquí donde el presente proyecto se convierte en una oportunidad clara para ayudar en la aplicación de técnicas y modelos de clasificación de emociones.

1.3. Objetivos

Objetivo general

Realizar evaluación del desempeño del modelo de redes neuronales convolucionales en contraste con el modelo de transformadores de visión para la clasificación de emociones en imágenes de rostros.

Objetivos específicos

- Realizar análisis exploratorio y modelo basado en redes neuronales convolucionales CNN).
- Realizar análisis exploratorio y modelo basado en transformadores de visión (ViT).
- Comparar ambos modelos, redes neuronales convolucionales y transformadores de visión.

1.4. Alcance

- **Comparación:** La comparación entre los modelos entrenados para medir la clasificación de las emociones, y no se tendrán en cuenta otros aspectos como tiempo de respuesta o consumo computacional.
- **Emociones:** Dada la dificultad y subjetividad que se puede presentar al momento de clasificar una emoción, se establece que los modelos podrán reconocer las siguientes emociones: felicidad, enojo, tristeza, miedo, sorpresa, asco, neutralidad.
- **Conjunto de datos de entrenamiento:** Se utilizará para el entrenamiento de los modelos el conjunto de fer2013 (del Kaggle Facial Expression Recognition Challenge) que contiene 28,709 imágenes de personas, categorizadas en siete clases de emociones.
- **Prototipo funcional:** Se desarrollará una API de prototipo funcional para la carga de las imágenes y su clasificación mediante una interfaz web con ambos modelos.
- **Despliegue:** La validación cualitativa será realizada mediante el prototipo funcional con imágenes obtenidas de un ambiente controlado donde solo se tenga un rostro en la imagen.

1.5. Metodología

Este trabajo sigue la metodología CRISP-DM, que ofrece un enfoque sistemático para abordar problemas utilizando modelos de datos, aplicables en diversas industrias y casos de uso. En este capítulo se describen las actividades realizadas durante las diferentes fases del ciclo CRISP-DM.

I. Comprensión del negocio	**II. Comprensión de los datos**	**III. Preparación de los datos**
■ Definir el contexto y la importancia de la clasificación de emociones en imágenes de rostros. ■ Identificar los stakeholders y sus necesidades. ■ Establecer los criterios de éxito para la evaluación del desempeño de los modelos. ■ Comprender las limitaciones y restricciones del proyecto.	■ Recopilar conjuntos de datos de imágenes de rostros etiquetados con emociones. ■ Explorar la estructura y las características de los conjuntos de datos. ■ Identificar posibles problemas de calidad de datos, como desequilibrios de clases o ruido. ■ Realizar visualizaciones y estadísticas descriptivas para comprender mejor los datos.	■ Llevar a cabo técnicas de preprocesamiento de imágenes, como redimensionamiento, normalización y aumento de datos. ■ Dividir los datos en conjuntos de entrenamiento, validación y prueba. ■ Manejar desequilibrios de clases utilizando técnicas como sobremuestreo, submuestreo o pesos de clase. ■ Codificar las etiquetas de emociones en un formato adecuado para su uso en los modelos.

IV. Modelado	V. Evaluación
■ Implementar un modelo de Redes Neuronales Convolucionales (CNN) para la clasificación de emociones en imágenes de rostros. ■ Implementar un modelo de transformadores de visión (ViT) para la misma tarea. ■ Entrenar ambos modelos utilizando los conjuntos de datos preparados. ■ Ajustar hiperparámetros y realizar validación cruzada si es necesario para optimizar el rendimiento de los modelos.	■ Evaluar el desempeño de los modelos utilizando métricas relevantes como precisión, recall, F1-score y matriz de confusión. ■ Comparar el desempeño de los modelos de CNN y ViT en términos de métricas de evaluación. ■ Comparar el desempeño de los modelos de CNN y ViT. ■ Desplegar un prototipo funcional con el cual interpretar los resultados de los modelos y realizar análisis para comprender las fortalezas y debilidades de cada modelo.

La metodología CRISP-DM proporciona un marco estructurado para llevar a cabo la evaluación del desempeño de los modelos de inteligencia artificial propuestos en el contexto de clasificación de emociones en imágenes de rostros. Esta metodología se verá a detalle en el apartado “Metodología CRISP-DM” del marco teórico.

2 Marco teórico y estado del arte

2.1. Marco teórico

El objetivo de este capítulo es establecer los conceptos relevantes en el marco del proyecto. En primera instancia, se establecen conceptos sobre aprendizaje supervisado, haciendo énfasis en el funcionamiento de las redes neuronales convolucionales. En segunda instancia, se definen aspectos sobre la clasificación de estados emocionales por medio de imágenes. Por último, se establecen los conceptos asociados a las temáticas de programación de aplicaciones, las emociones y metodología CRISP-DM.

2.1.1. Aprendizaje supervisado

El Aprendizaje Supervisado es un enfoque en el campo del Aprendizaje Automático en donde un algoritmo se entrena utilizando un conjunto de datos, los cuales se representan como valores de entrada, al igual que las etiquetas correspondientes que representan las salidas. El objetivo principal es definir un comportamiento que pueda generalizarse para predecir con precisión las salidas de nuevas instancias no etiquetadas. [Cardoso et al., 2021]

El Aprendizaje Supervisado conlleva la elaboración de un modelo que tiene la capacidad de efectuar pronósticos fundamentados en las relaciones entre las variables de entrada y las respuestas etiquetadas, las cuales se obtienen durante la etapa de entrenamiento. Estos pronósticos se pueden aplicar en un amplio espectro de contextos, abarcando desde la identificación de contenido no deseado en comunicaciones electrónicas, como el correo electrónico, hasta la clasificación de emociones por medio de imágenes de rostros de personas [Wachter et al., 2021].

Ténicas de clasificación en el aprendizaje supervisado:
En el Aprendizaje Supervisado es posible encontrar diferentes algoritmos que permiten definir una clasificación por medio de cálculos matemáticos [Wang et al., 2021]:

- Vecinos más cercanos (K-Nearest Neighbors - KNN): KNN es un algoritmo de clasificación que agrupa las instancias basado en la similitud con sus vecinos más cercanos en

el espacio de características. No requiere una fase de entrenamiento formal y se utiliza para problemas de clasificación y regresión [Wang et al., 2021].

- Máquinas de soporte vectorial (SVM): Los SVM son algoritmos de clasificación que buscan encontrar el hiperplano de separación óptimo entre clases. Son eficaces en problemas de clasificación binaria y pueden manejar problemas de alta dimensionalidad. Los SVM también se pueden utilizar en problemas de regresión [Wang et al., 2021].

- Naive Bayes: El método Naive Bayes es un algoritmo de clasificación supervisado basado en el teorema de Bayes, que establece una relación probabilística entre una hipótesis y la evidencia relevante. Este método es "naive" (ingenuo) debido a la suposición simplificadora de independencia condicional entre las características o atributos utilizados para la clasificación, dada la clase a la que pertenece el objeto en cuestión [Sathyamoorthy et al., 2023].

- Regresión logística: La regresión logística es un modelo estadístico utilizado para la modelización y predicción de variables dependientes binarias. A diferencia de la regresión lineal, que se emplea para variables continuas, la regresión logística se adapta a situaciones en las que la variable de interés es categórica y dicotómica, con dos posibles resultados, como sí/no, éxito/fracaso, o positivo/negativo [Meena et al., 2023].

- Árboles de decisión: Los árboles de decisión son modelos de aprendizaje automático que se utilizan para abordar problemas de clasificación y regresión. Estos árboles estructuran el proceso de toma de decisiones en forma de un diagrama de flujo jerárquico, donde cada nodo representa una condición o característica, y cada borde representa el resultado de esa condición. La construcción de un árbol de decisión implica dividir repetidamente el conjunto de datos en subconjuntos más homogéneos en términos de la variable objetivo [Zheng et al., 2023].

Redes neuronales artificiales (ANN):
Las redes neuronales son modelos de aprendizaje profundo que constan de capas de unidades interconectadas que pueden aprender representaciones complejas de datos. Una ANN está compuesta por múltiples capas de neuronas interconectadas. La capa de entrada recibe los datos, las capas ocultas realizan cálculos intermedios y la capa de salida produce las predicciones finales. En el proceso de entrenamiento, los datos de entrada se propagan a través de la red capa por capa, y se ajustan los pesos de las conexiones para minimizar el error entre las predicciones y las salidas deseadas. Se utilizan en una variedad de aplicaciones, como visión por computadora, procesamiento de lenguaje natural y clasificación de imágenes [Wang et al., 2021].

Hiperparámetros en una red neuronal artificial:
Los parámetros representan la configuración que se puede administrar para una red neuronal

artificial. El principal objetivo es alcanzar el rendimiento máximo en términos de clasificación. Generalmente, se mencionan como elementos administrables dentro de los hiperparámetros a:

- Tasa de aprendizaje: La actualización de los pesos (W) durante el entrenamiento se realiza mediante la regla de aprendizaje, que puede expresarse como [Pennington et al., 2017]:

$$W^{(l)} \leftarrow W^{(l)} - \alpha \frac{\partial J}{\partial W^{(l)}} \tag{2-1}$$

  Donde:

- J es la función de costo.
- α es la tasa de aprendizaje.

Funciones de activación: La salida $z^{(l)}$ de una neurona en la capa l antes de aplicar la función de activación se calcula como [Pennington et al., 2017]:

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)} \tag{2-2}$$

La función de activación $a^{(l)}$ se obtiene aplicando la función $g^{(l)}(z^{(l)})$, donde $g^{(l)}$ puede ser la función sigmoide, tangente hiperbólica (tanh) o unidad lineal rectificada (ReLU). Estas funciones se expresan como [Pennington et al., 2017]:

$$\begin{aligned} \text{Sigmoide:} \& \quad & g(z) = \frac{1}{1+e^{-z}} \\ \text{Tangente Hiperbólica (tanh):} \& \quad & g(z) = \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \\ \text{Unidad Lineal Rectificada (ReLU):} \& \quad & g(z) = \max(0, z) \end{aligned} \tag{2-3}$$

Capas ocultas: En una red neuronal con L capas, donde $h^{(l)}$ representa el número de neuronas en la capa oculta l, las operaciones realizadas en la capa oculta l pueden expresarse matemáticamente como [Pennington et al., 2017]:

$$a^{(l)} \& = g^{(l)}(W^{(l)} a^{(l-1)} + b^{(l)}) \tag{2-4}$$

Donde:

- $a^{(l-1)}$ es el vector de activación de la capa anterior.
- $W^{(l)}$ es la matriz de pesos para la capa l.
- $b^{(l)}$ es el vector de sesgo para la capa l.
- $g^{(l)}$ es la función de activación aplicada elemento por elemento.

La selección adecuada de hiperparámetros influye directamente en la capacidad del modelo para aprender patrones complejos en datos. La tasa de aprendizaje, por ejemplo, regula la magnitud de los ajustes de peso durante el proceso de entrenamiento, siendo su elección determinante para evitar convergencia prematura o estancamiento [Smith, 2017].

En la estratificación de capas y neuronas, los hiperparámetros inciden directamente en la capacidad del modelo para representar y generalizar información. Una red con demasiadas capas puede sufrir de sobreajuste, capturando ruido en lugar de patrones significativos, mientras que una cantidad insuficiente podría limitar su capacidad para modelar relaciones complejas. La elección de funciones de activación, por otro lado, determina la no linealidad del modelo y su capacidad para aprender representaciones más complejas. La interacción sutil entre estos hiperparámetros requiere un enfoque cuidadoso para optimizar el rendimiento del modelo [Smys et al., 2020].
Los hiperparámetros, al ser ajustados durante el proceso de entrenamiento, impactan significativamente en la eficiencia y convergencia del modelo. El ajuste manual de estos parámetros puede ser un proceso arduo y propenso a errores, lo que ha llevado al desarrollo de técnicas automáticas de optimización de hiperparámetros. Estas estrategias buscan encontrar configuraciones óptimas mediante algoritmos de búsqueda, reduciendo así la carga en el investigador y mejorando la eficacia del modelo. La optimización de hiperparámetros se ha convertido, por ende, en un área activa de investigación en el ámbito de las redes neuronales artificiales [Smys et al., 2020].

Métricas para la evaluación de los modelos:
La evaluación de modelos de aprendizaje supervisado es crucial para determinar su rendimiento y su capacidad para generalizar a datos no vistos. Existen varias métricas que se utilizan comúnmente para evaluar la eficacia de estos modelos. A continuación, se presentan algunas de las métricas de evaluación más comunes:

- Exactitud (Accuracy): La precisión mide la proporción de predicciones correctas en relación con el total de predicciones. Se recomienda usar cuando las clases están balanceadas [Japkowicz, 2013].

$$\text{Exactitud (Accuracy)} = \frac{\text{Verdaderos Positivos} + \text{Verdaderos Negativos}}{\text{Total de Predicciones}} \tag{2-5}$$

- Sensibilidad (Recall): El recall mide la proporción de ejemplos positivos que se han clasificado correctamente. Es útil cuando es importante identificar todos los casos positivos [Japkowicz, 2013].

$$\text{Recall (Sensibilidad)} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}} \tag{2-6}$$

- Precisión (Precision) La precisión mide la proporción de predicciones correctas con respecto a las verdaderas [Japkowicz, 2013].

$$\text{Precisión (Precision)} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Positivos}} \tag{2-7}$$

- F1-Score: El F1-score es una medida que combina precisión y recall en una ténica métrica. Es útil cuando se desea un equilibrio entre ambas [Japkowicz, 2013].

$$\text{F1-Score} = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}} \tag{2-8}$$

- Matriz de confusión: La matriz de confusión es una herramienta que ofrece una evaluación detallada del rendimiento de un modelo al comparar las predicciones del modelo con las clases reales de los datos. Se destaca por su utilidad en contextos de clasificación binaria, donde se requiere la discriminación precisa entre dos clases distintas [Japkowicz, 2013].

	Clase Positiva	Clase Negativa
Predicción Positiva	Verdaderos Positivos (TP)	Falsos Positivos (FP)
Predicción Negativa	Falsos Negativos (FN)	Verdaderos Negativos (TN)

Redes neuronales convolucionales:
Son un tipo especializado de redes neuronales artificiales inicialmente implementadas para el procesamiento y análisis de datos con una estructura de cuadrícula, siendo más prominentes en tareas de procesamiento de imágenes. La principal diferencia entre las (CNN) y otras redes neuronales radica en su capacidad para procesar y analizar datos con estructura de cuadrícula, como imágenes, de manera altamente efectiva. Las CNN se han implementado para abordar específicamente esta necesidad, lo que las hace únicas en comparación con otros tipos de redes neuronales, como las redes neuronales recurrentes [Wäldchen and Mäder, 2018].

La arquitectura básica de una red neuronal convolucional (CNN) consta de varias capas interconectadas implementadas para procesar y analizar datos con estructura de cuadrícula, como imágenes. A continuación, se detallan las diferentes capas de una CNN [Carrera et al., 2021]:

- Capa de entrada: La capa de entrada recibe la imagen o el dato con estructura de cuadrícula en su resolución original. Cada celda de la cuadrícula representa un valor de píxel o característica [Carrera et al., 2021].

- Capas de convolución: Las capas de convolución son fundamentales en una CNN. Utilizan filtros (kernels) que se deslizan sobre la imagen para extraer características locales. Cada filtro detecta patrones específicos, como bordes, texturas o formas, generando mapas de características [Carrera et al., 2021].

- Funciones de activación: Después de cada operación de convolución, se aplica una función de activación no lineal, como ReLU (Rectified Linear Unit), para introducir no linealidad en el modelo y permitir la extracción de características más complejas [Carrera et al., 2021].
- Capas de submuestreo (Pooling): Las capas de submuestreo, como Max-Pooling, reducen la resolución espacial de los mapas de características, manteniendo las características más importantes y disminuyendo la cantidad de datos [Carrera et al., 2021].
- Capas totalmente conectadas (fully connected layers): Las capas totalmente conectadas conducen a la clasificación realizada con el modelo [Carrera et al., 2021].

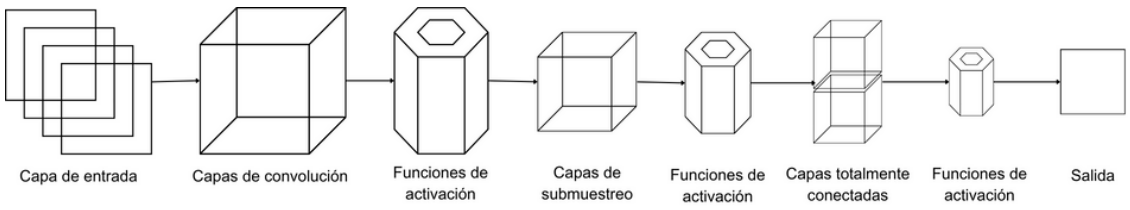

Figura 2-1: Gráfico de arquitectura de una red neuronal convolucional (CNN).

Transformadores de visión:
Los transformadores de visión (ViT) son un tipo de modelo de aprendizaje automático diseñado para tareas de visión por computadora, que se basan en la arquitectura de los transformadores, originalmente desarrollada para procesamiento de lenguaje natural, y adaptados para trabajar con datos visuales. Los ViT logran excelentes resultados en tareas de clasificación de imágenes, superando en algunos casos a las redes neuronales convolucionales (CNN) y requiriendo menos recursos computacionales para ser entrenadas [Dosovitskiy et al., 2020]. Estos dividen una imagen en parches aplanados (pequeños bloques), sin tomar la imagen como un todo, y tratan cada parche como una “palabra” y generando una secuencia de vectores que es entregada a un codificador de transformador, similar a cómo los transformadores procesan palabras en un texto. Posteriormente, aplican el modelo de clasificación resultante para capturar relaciones entre estos parches [Dosovitskiy et al., 2020].

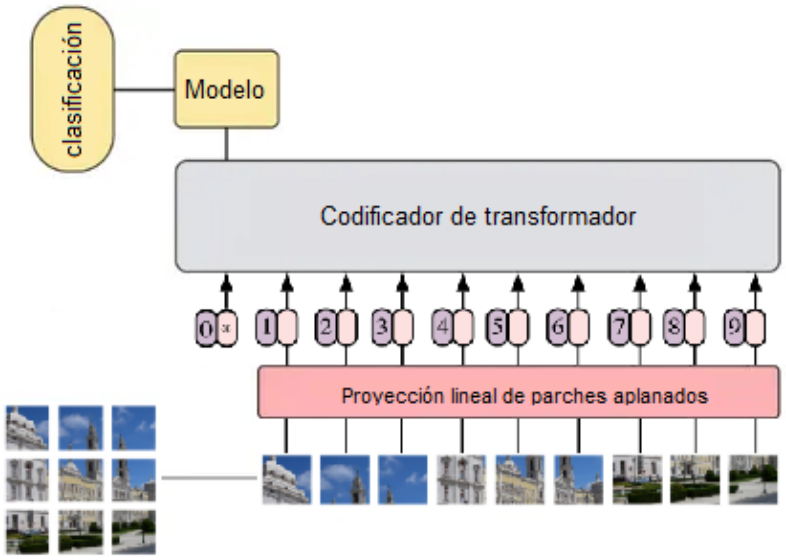

Figura 2-2: Gráfico de arquitectura de transformadores de visión (ViT). [Dosovitskiy et al., 2020]

2.1.2. Preparación de los datos

La preparación de datos en el análisis de imágenes mediante técnicas de aprendizaje automático reviste una importancia crítica para garantizar la eficacia y generalización del modelo. La primera etapa es la normalización de píxeles, donde se ajustan los valores para asegurar que se encuentren en un rango específico, generalmente [0, 1] o [-1, 1], facilitando así la convergencia durante el entrenamiento. La uniformidad en la escala de píxeles es esencial para evitar sesgos y mejorar la estabilidad del modelo [Li et al., 2023].

La segunda etapa es el redimensionamiento de las imágenes, donde se homogenizan sus dimensiones para facilitar la entrada coherente al modelo. Este proceso garantiza que las imágenes presenten la misma resolución, lo que simplifica significativamente el procesamiento y permite un entrenamiento más adecuado de las redes neuronales convolucionales (CNN). La consistencia en las dimensiones de entrada es importante para mantener la coherencia en la representación de las características aprendidas por el modelo [Wessels et al., 2022].

Además, la aplicación de técnicas de aumento de datos se erige como una estrategia esencial en la preparación de datos. La generación de variantes ligeramente modificadas de las imágenes de entrenamiento, mediante rotaciones, zooms y espejeos, amplía la diversidad del conjunto de datos y favorece la capacidad del modelo para generalizar a patrones no vistos durante el entrenamiento. Este enfoque de enriquecimiento de datos fortalece la robustez del modelo, mitigando así el riesgo de sobreajuste y mejorando su capacidad predictiva en con-

diciones del mundo real [Li et al., 2023].

- Entrenamiento: Durante esta etapa, el modelo se ajusta a un conjunto de datos de entrenamiento etiquetado. Las imágenes se introducen en la red, y los pesos y sesgos del modelo se ajustan iterativamente para minimizar una función de pérdida que mide la discrepancia entre las predicciones y las etiquetas reales. Se utiliza un optimizador, como el descenso del gradiente estocástico (SGD) o algoritmos más avanzados como Adam, para ajustar los parámetros del modelo [Yang et al., 2022].

- Pruebas: Después del entrenamiento, el modelo se evalúa en un conjunto de datos de pruebas independiente que no se utilizó durante el entrenamiento. Esto ayuda a medir el rendimiento del modelo en datos no vistos y a detectar posibles problemas de sobreajuste. Se ajustan los hiperparámetros, como la tasa de aprendizaje o la dimensionalidad del conjunto de datos, según sea necesario, y se repite el proceso de entrenamiento y pruebas si es necesario [Yang et al., 2022].

- Validación: Una vez que el modelo ha sido entrenado y probado, se evalúa en un conjunto completamente nuevo y no visto previamente, denominado conjunto de validación. Este conjunto proporciona una evaluación final del rendimiento del modelo en datos no utilizados anteriormente. Se calculan métricas de rendimiento, como precisión, recall, F1-score u otras métricas específicas del problema, para evaluar la capacidad del modelo para generalizar a datos no vistos [Yang et al., 2022].

Dependencias para etapas de análisis

A continuación, se presentan las bibliotecas y los módulos utilizados:

- OpenCV (Open Source Computer Vision Library):
  Esta biblioteca de visión por computadora se destaca como una biblioteca fundamental en el ámbito del procesamiento de imágenes. Su arquitectura modular y su capacidad para funcionar en una variedad de plataformas hacen que la convierten en una opción versátil para aplicaciones que abarcan desde el análisis de imágenes médicas hasta la visión por computadora industrial. [Galety et al., 2022].

- Keras ImageDataGenerator:
  El módulo ImageDataGenerator de Keras representa una herramienta fundamental en el ámbito del aprendizaje profundo aplicado a imágenes. Este componente ofrece una interfaz versátil para la generación dinámica de datos durante el entrenamiento de modelos de redes neuronales convolucionales (CNN). Su capacidad para aplicar transformaciones en tiempo real a las imágenes de entrada, como rotaciones y zooms, no solo enriquece la diversidad del conjunto de datos, sino que también mitiga el riesgo de sobreajuste al introducir variabilidad durante el proceso de entrenamiento
  [Godishala et al., 2022].

- NumPy (Numerical Python):
  Es una biblioteca fundamental en el ámbito del computación científica en Python. Implementada para facilitar operaciones numéricas eficientes, NumPy proporciona una estructura de datos central: el array multidimensional. Este objeto permite la representación y manipulación eficiente de datos en forma de matrices, siendo esencial en el ámbito de la álgebra lineal, estadísticas y manipulación de datos para aplicaciones científicas y de aprendizaje automático. La capacidad de NumPy para realizar operaciones vectorizadas acelera significativamente los cálculos numéricos, convirtiéndola en una herramienta esencial en la implementación de algoritmos y en la manipulación de datos científicos de gran escala [Khan et al., 2022].

- Scikit-image:
  Se conoce como una biblioteca especializada en el procesamiento de imágenes en Python. Su enfoque se centra en proporcionar una colección de algoritmos y herramientas para el análisis y manipulación eficiente de imágenes. Con una interfaz amigable, scikit-image se ha utilizado para tareas que abarcan desde la segmentación y filtrado de imágenes hasta la extracción de características [Kuwahara et al., 2023].

2.1.3. Interfaz de programación de aplicaciones

En este apartado se profundiza en la Interfaz de Programación de Aplicaciones (API), con el propósito de tener las bases para el desarrollo de un prototipo funcional que permita utilizar y evaluar los modelos de clasificación.
La Interfaz de Programación de Aplicaciones (API) se define como un conjunto estandarizado de reglas y protocolos destinados a facilitar la interacción y la comunicación entre diversas aplicaciones informáticas. Su función principal consiste en establecer un marco normativo que rige las solicitudes y respuestas entre entidades digitales, actuando como un intermediario esencial en la transferencia eficiente y segura de datos y funcionalidades [Ehsan et al., 2022].
La evolución de las API ha sido motivada por la necesidad imperante de modularidad en el desarrollo de sistemas informáticos, permitiendo así la construcción de aplicaciones más adaptables y flexibles. Este enfoque modular no solo favorece la escalabilidad, sino también la actualización independiente de componentes y la colaboración efectiva entre equipos de desarrollo [Ehsan et al., 2022].
Dentro de las diversas categorías de API, se destacan las API web, las cuales posibilitan la comunicación entre servicios a través de internet, y las API de bibliotecas, que ofrecen funciones específicas integrables en aplicaciones. Asimismo, las API RESTful, basadas en el protocolo HTTP, han emergido como un estándar eficiente para el desarrollo de servicios web [van der Vlist et al., 2022]. En el contexto de la economía de las plataformas, las API soportan un papel crucial al permitir que grandes empresas tecnológicas abran sus servicios mediante estas interfaces, facilitando a terceros la construcción sobre su infraestructura

[Martínez-Plumed et al., 2019]. Este enfoque ha propiciado la formación de ecosistemas digitales, donde aplicaciones y servicios interactúan de manera sinérgica, creando así una red interconectada de funcionalidades [van der Vlist et al., 2022].

Comunicación entre API's:
La comunicación entre APIs, en el ámbito de la tecnología, se rige por principios fundamentales que permiten la interacción eficiente y segura entre distintas aplicaciones informáticas. Este proceso, esencial para la colaboración digital, sigue un conjunto de pasos claves que aseguran la transferencia efectiva de datos y funcionalidades. Primordialmente, la comunicación entre APIs se basa en el intercambio de solicitudes y respuestas, siguiendo un modelo cliente-servidor. Cuando una aplicación, denominada cliente, requiere acceder a los recursos de otra aplicación, conocida como servidor, envía una solicitud a través de la API. Esta solicitud incluye información específica sobre la acción deseada y puede ir en conjunto de parámetros necesarios para la operación.
A continuación, se define el comportamiento general de una API por medio de pseudocódigo.

- Petición - Obtener datos:
- Establecer endpoint URL = "https://api.servidor.com/datos"
- Respuesta = realizar solicitud (endpointURL)
- Datos = procesar respuesta (respuesta) - Hacer algo con los datos (datos)

Comportamiento general de una API:
Primordialmente, la comunicación entre APIs se basa en el intercambio de solicitudes y respuestas, siguiendo un modelo cliente-servidor [Ehsan et al., 2022]. Cuando una aplicación, denominada cliente, requiere acceder a los recursos de otra aplicación, conocida como servidor, envía una solicitud a través de la API. Esta solicitud incluye información específica sobre la acción deseada y puede ir en conjunto de parámetros necesarios para la operación. La API del servidor, a su vez, procesa la solicitud, lleva a cabo la acción solicitada y genera una respuesta. Esta respuesta contiene la información solicitada o indica el resultado de la operación realizada. La comunicación suele estar basada en protocolos estándar, como HTTP o HTTPS, que garantizan la uniformidad y seguridad en la transferencia de datos.
El formato de intercambio de datos es un aspecto crítico en la comunicación entre APIs. En la mayoría de los casos, se utiliza el formato JSON (JavaScript Object Notation) debido a su legibilidad, ligereza y facilidad de interpretación por parte de las aplicaciones [Ehsan et al., 2022]. Sin embargo, otros formatos, como XML, también se han utilizado en el pasado, dependiendo de los requisitos y estándares de la industria. La autenticación y la autorización son aspectos esenciales en la comunicación entre APIs para garantizar la seguridad. Los mecanismos como claves de API, tokens de acceso y protocolos como OAuth se emplean para verificar la identidad del cliente y permitir o restringir el acceso a los recursos del servidor. En la arquitectura RESTful, una variante común en la implementación de APIs, los endpoints definen

los puntos de acceso a los servicios ofrecidos por la API. Cada endpoint representa una operación específica y se accede mediante métodos HTTP estándar como GET, POST, PUT o DELETE [Ehsan et al., 2022].

Transmisión de imágenes por medio API's:

- Codificación de la imagen:
  Antes de enviar la imagen, es necesario codificarla en un formato que sea eficiente para la transmisión binaria. Los formatos comunes para imágenes incluyen JPEG, PNG o GIF. Cada píxel de la imagen se representa y codifica según el formato elegido [Savva and Stylianou, 2023].

- Solicitud o respuesta HTTP:
  La imagen codificada se incorpora a la solicitud o respuesta HTTP. En una solicitud, esto podría implicar la inclusión de la imagen como parte del cuerpo (payload) de la solicitud, mientras que en una respuesta, la imagen se adjuntaría de manera similar. El tipo de contenido (Content-Type) en los encabezados HTTP indica el formato de la imagen [Savva and Stylianou, 2023].

- Envío de la solicitud o respuesta:
  La solicitud HTTP, con la imagen codificada, se envía al servidor a través del método apropiado (por ejemplo, POST o PUT). En el caso de la respuesta, la imagen se envía desde el servidor al cliente como parte de la respuesta HTTP [Savva and Stylianou, 2023].

- Decodificación en el extremo receptor: En el extremo receptor, la imagen recibida se decodifica según el formato especificado en el tipo de contenido. Esto implica convertir los datos binarios nuevamente en una representación de imagen que pueda ser comprendida y procesada [Savva and Stylianou, 2023].

- Manejo de errores y seguridad:
  Durante este proceso, se deben considerar aspectos de manejo de errores y seguridad. Mecanismos como códigos de estado HTTP, como el código 200 para éxito o códigos 4xx y 5xx para errores, son cruciales. Además, en situaciones de seguridad, se pueden aplicar técnicas como la autenticación y el cifrado para proteger la integridad y confidencialidad de la imagen transmitida [Savva and Stylianou, 2023].

2.1.4. Introducción a las emociones

Historia y aspecto psicológico de las emociones:
Las emociones, esas reacciones psicofisiológicas que todos experimentamos, son fundamentales en la vida humana. Estas respuestas complejas, que se manifiestan ante estímulos

específicos, juegan un papel crucial en el comportamiento humano, toma de decisiones, y relaciones interpersonales [Ekman, 1992]. Las emociones no son meras reacciones pasajeras; son el reflejo de la interpretación y respuesta al entorno, y tienen profundas implicaciones en la salud mental y física.

Desde una perspectiva psicológica, las emociones han sido objeto de estudio y debate durante décadas. Tradicionalmente, se han clasificado en categorías básicas como “felicidad”, “tristeza”, “enojo”, “sorpresa”, “miedo” y “asco” [Ekman, 1999]. Estas categorías, propuestas inicialmente por Paul Ekman, se basan en la idea de que hay emociones universales que todos los seres humanos experimentan y expresan de manera similar, independientemente de su cultura o educación. Sin embargo, esta perspectiva ha sido objeto de críticas y revisiones. Investigaciones más recientes sugieren que las emociones no se limitan a estas categorías y pueden ser vistas más como un espectro o una combinación de experiencias básicas [Russell, 2003]. Esta idea propone que las emociones no son entidades fijas, sino construcciones que emergen de la interacción entre la biología, la cognición y el entorno.

Históricamente, las emociones han sido estudiadas desde diversas disciplinas. Charles Darwin, en su obra “La expresión de las emociones en el hombre y los animales”, fue uno de los pioneros en explorar la universalidad de las expresiones faciales y sus orígenes evolutivos [Darwin, 1872]. Darwin propuso que ciertas expresiones emocionales tienen un carácter universal y son el resultado de la evolución. Estas expresiones, argumentó, tienen un valor adaptativo y han sido seleccionadas a lo largo de la evolución porque facilitan la supervivencia y reproducción. A lo largo del tiempo, la comprensión de las emociones ha evolucionado, incorporando perspectivas neurocientíficas, cognitivas y sociales. Investigadores como Joseph LeDoux han explorado las bases neuronales de las emociones, identificando circuitos cerebrales específicos involucrados en la experiencia y expresión emocional [LeDoux, 2000]. Por otro lado, Lisa Feldman Barrett ha propuesto que las emociones son construcciones psicológicas que emergen de la interacción entre procesos cognitivos y fisiológicos [Barrett, 2006].

También, se ha identificado que estructuras cerebrales como la amígdala y el córtex prefrontal juegan roles cruciales en la percepción, regulación y expresión de las emociones [LeDoux, 2000]. Estos descubrimientos han llevado a avances en la comprensión de trastornos emocionales y han informado tratamientos terapéuticos.

Con los avances tecnológicos, la clasificación y el análisis de emociones a través de imágenes, especialmente expresiones faciales, se han convertido en un área de investigación prominente. Las técnicas avanzadas de procesamiento de imágenes han permitido a los investigadores analizar expresiones faciales y otras señales no verbales para identificar y clasificar emociones con precisión [Calvo and D’Mello, 2010a]. Estos avances no solo han proporcionado herramientas valiosas para la investigación básica, sino que también tienen aplicaciones prácticas en áreas como la psicoterapia, la medicina y la inteligencia artificial.

La autoconciencia y autorregulación emocional son aspectos fundamentales en la adaptación y bienestar de los individuos. La capacidad de reconocer y entender las propias emociones, así como de regularlas adecuadamente, tiene implicaciones profundas en la salud mental,

relaciones interpersonales y éxito en diversas áreas de la vida [Gross, 2002].
Las emociones son una parte integral de la experiencia humana. Aunque tradicionalmente se han categorizado en emociones básicas, las investigaciones actuales sugieren una visión más matizada y compleja. Las emociones emergen de la interacción entre la biología, la cognición y el entorno, y tienen profundas implicaciones en la vida diaria. A medida que la tecnología avanza, las herramientas para estudiar y comprender las emociones se vuelven más sofisticadas, abriendo nuevas posibilidades y desafíos para el futuro de la investigación emocional.

La neurociencia afectiva:
Lisa Feldman Barrett es una figura central en el estudio contemporáneo de las emociones. Su trabajo ha desafiado y redefinido las concepciones tradicionales sobre cómo el cerebro crea y experimenta las emociones [Barrett, 2017a]. A lo largo de su carrera, Barrett ha combinado investigaciones empíricas con teorías innovadoras para ofrecer una visión más matizada y contextualizada de la naturaleza de las emociones.
Desde sus primeros trabajos, Barrett ha cuestionado la idea de que las emociones son respuestas innatas y universales a estímulos específicos. En lugar de ver las emociones como categorías fijas y predefinidas, Barrett propone que son construcciones del cerebro que emergen de la interacción entre el cuerpo, un cerebro flexible y el ambiente cultural y social [Barrett, 2012]. Esta perspectiva, conocida como teoría constructivista de las emociones, sugiere que las emociones son el resultado de un proceso activo de interpretación y construcción, en lugar de ser meras reacciones pasivas.
En su libro “How Emotions are Made: The Secret Life of the Brain”, Barrett profundiza en esta idea, argumentando que las emociones son el resultado de predicciones que el cerebro hace constantemente [Barrett, 2017b]. Estas predicciones se basan en experiencias pasadas y en la información presente, y son esenciales para la supervivencia, ya que permiten al cerebro anticipar y prepararse para eventos futuros. Esta idea de las emociones como predicciones es revolucionaria y ha llevado a repensar cómo se estudian y comprenden las emociones en la psicología y la neurociencia.
Otro aspecto crucial del trabajo de Barrett es su enfoque en la variabilidad de la experiencia emocional. A través de estudios de neuroimagen y experimentos psicológicos, Barrett y su equipo han demostrado que no hay una “huella digital” única para cada emoción en el cerebro [Barrett, 2016]. En cambio, diferentes personas pueden tener patrones neuronales distintos para la misma emoción, y una emoción específica puede manifestarse de diferentes maneras en el cerebro según el contexto. Esta variabilidad, según Barrett, es el resultado de la naturaleza constructiva de las emociones y de la influencia del contexto y la cultura en su experiencia y expresión.
Barrett también ha explorado cómo las diferencias culturales y lingüísticas pueden influir en la forma en que las personas experimentan y expresan emociones. En culturas diferentes, las personas pueden tener conceptos emocionales únicos que no tienen un equivalente directo en otras culturas [Barrett, 2019]. Esto sugiere que el cerebro no solo se basa en la biología para

construir emociones, sino que también se apoya en el lenguaje y la cultura. Esta interacción entre biología, cultura y lenguaje es central en la teoría constructivista de Barrett y ha llevado a una mayor apreciación de la diversidad y complejidad de la experiencia emocional.
Además de su trabajo teórico, Barrett ha sido una defensora activa de la investigación interdisciplinaria y colaborativa. Ha trabajado con neurocientíficos, psicólogos, lingüistas y antropólogos para abordar preguntas sobre la naturaleza y función de las emociones desde múltiples perspectivas. Esta aproximación interdisciplinaria ha enriquecido el campo de la neurociencia afectiva y ha llevado a nuevos descubrimientos y comprensiones.
El trabajo de Lisa Feldman Barrett ha transformado la forma en que se entienden las emociones. Su enfoque constructivista ha desafiado las concepciones tradicionales y ha ofrecido una visión más rica y matizada de la naturaleza de las emociones. A medida que se continúa explorando la complejidad de la experiencia emocional, el trabajo de Barrett seguirá siendo una referencia esencial y una fuente de inspiración.

Clasificación de emociones mediante imágenes:
La clasificación de emociones a través de imágenes ha experimentado un auge significativo con los avances tecnológicos recientes. Las técnicas de procesamiento de imágenes y aprendizaje automático han permitido a los investigadores analizar expresiones faciales y otras señales visuales para identificar y clasificar emociones con una precisión sin precedentes. Sin embargo, este progreso no está exento de desafíos [Calvo and D'Mello, 2010b].
Uno de los principales desafíos en la clasificación de emociones mediante imágenes es la influencia de la deseabilidad social. Las personas, consciente o inconscientemente, pueden modificar la forma en que expresan sus emociones cuando saben que están siendo observadas o evaluadas [Krumm and Davies, 2017]. Esto puede llevar a resultados sesgados, especialmente en situaciones donde los participantes pueden sentir la necesidad de presentarse de una manera particular.
Además, informar a los participantes sobre la clasificación de emociones puede alterar la forma en que manifiestan sus emociones. El simple hecho de saber que sus emociones están siendo monitoreadas puede influir en la autenticidad de las expresiones emocionales, lo que puede complicar la interpretación de los resultados [Pantic and Rothkrantz, 2008].
Desde una perspectiva ética, es esencial garantizar el bienestar y la privacidad de los participantes en estudios de clasificación emocional. Las imágenes, especialmente las del rostro, contienen información personal y sensible. Por lo tanto, es crucial asegurarse de que estas imágenes se manejen con cuidado, se almacenen de forma segura y se utilicen solo con el consentimiento informado de los participantes [Harvey, 2016].
Además, es importante considerar el impacto psicológico de la clasificación de emociones en los participantes. En algunos casos, ser informado sobre sus emociones o ser consciente de que están siendo evaluados puede generar ansiedad o malestar. Es esencial que los investigadores estén atentos a estos riesgos y proporcionen el apoyo necesario a los participantes [McStay, 2018].

La clasificación de emociones mediante imágenes también plantea desafíos técnicos. La calidad de la imagen, la iluminación, el ángulo y otros factores pueden influir en la precisión de la clasificación. Además, las emociones son complejas y multifacéticas, y no siempre se manifiestan de la misma manera en diferentes individuos o culturas. Esto requiere algoritmos sofisticados capaces de adaptarse y aprender de una amplia variedad de datos [**?**].
A pesar de estos desafíos, la clasificación de emociones mediante imágenes ofrece oportunidades emocionantes para avanzar en la comprensión de la experiencia emocional. Las técnicas avanzadas de procesamiento de imágenes, combinadas con algoritmos de aprendizaje automático, pueden proporcionar insights valiosos sobre cómo las emociones se manifiestan y se perciben. Estos avances tienen el potencial de informar intervenciones terapéuticas, mejorar la interacción humano-máquina y enriquecer la comprensión de la naturaleza humana.
Con la creciente integración de la inteligencia artificial en la vida diaria, la clasificación de emociones mediante imágenes también tiene aplicaciones prácticas. Desde sistemas de reconocimiento facial en aeropuertos hasta aplicaciones de salud mental que monitorean el bienestar emocional, la capacidad de clasificar y entender las emociones a través de imágenes tiene un impacto directo en la sociedad.
En conclusión, mientras que la clasificación de emociones mediante imágenes es un campo en rápido desarrollo que ofrece oportunidades y desafíos, es esencial abordar los desafíos éticos y metodológicos para garantizar que la investigación se realice de manera responsable y que los resultados sean válidos y significativos.

2.1.5. Metodología CRISP-DM

La minería de datos ha emergido como una disciplina esencial en la era de la información, donde grandes volúmenes de datos están disponibles para las organizaciones en diversas formas y formatos. En este contexto, la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) se erige como un marco teórico robusto y ampliamente utilizado para guiar proyectos de minería de datos de manera sistemática y efectiva. Desarrollada por un consorcio de empresas líderes en la industria de minería de datos en 1996 [Schröer et al., 2021]. CRISP-DM proporciona una estructura bien definida que abarca desde la comprensión inicial del problema de negocio hasta la implementación práctica de soluciones basadas en datos. Este marco teórico se basa en principios científicos y prácticos, y ha demostrado su utilidad en una amplia gama de aplicaciones y sectores industriales [Shafique and Qaiser, 2014] [Wirth and Hipp, 2000].

Introducción a CRISP-DM:
La metodología CRISP-DM se ha consolidado como un estándar en la industria de minería de datos debido a su enfoque estructurado y su capacidad para adaptarse a diferentes contextos y problemas. CRISP-DM establece un conjunto de fases interconectadas que guían el proceso desde la comprensión inicial del negocio y los datos hasta la implementación de

soluciones y el monitoreo continuo de su desempeño. Estas fases son iterativas y permiten que los equipos de proyecto revisen y ajusten su enfoque a medida que adquieren un mayor entendimiento del problema y los datos disponibles [Wirth and Hipp, 2000].

Fases del proceso:
CRISP-DM consta de seis fases principales, cada una con objetivos específicos y tareas asociadas que se describen a continuación [Shafique and Qaiser, 2014]:

- Comprensión del negocio (Business Understanding): Esta fase implica establecer un entendimiento claro de los objetivos del negocio y cómo pueden abordarse mediante técnicas de minería de datos. Se identifican las necesidades, requerimientos y objetivos del proyecto, y se traducen en un problema de minería de datos bien definido [Shafique and Qaiser, 2014].

- Comprensión de los datos (Data Understanding): En esta etapa, se recopilan los datos relevantes para el proyecto y se realiza una exploración inicial para comprender su estructura, calidad y distribución. Se identifican posibles problemas y se evalúa la idoneidad de los datos disponibles para abordar el problema planteado
[Shafique and Qaiser, 2014].

- Preparación de los datos (Data Preparation): Aquí se llevan a cabo las tareas necesarias para preparar los datos para el modelado posterior. Esto puede incluir la limpieza de datos, la integración de diferentes fuentes de datos, la selección de variables relevantes y la transformación de los datos en un formato adecuado para el análisis [Shafique and Qaiser, 2014].

- Modelado (Modeling): En esta fase, se seleccionan y aplican diversas técnicas de modelado de datos para construir modelos que ayuden a resolver el problema de minería de datos identificado. Se exploran diferentes enfoques y se evalúan sus resultados para determinar el modelo más adecuado [Shafique and Qaiser, 2014].

- Evaluación (Evaluation): Se evalúan y validan los modelos construidos para asegurarse de que satisfagan los objetivos del negocio. Esto implica probar los modelos en datos independientes y comparar su rendimiento con criterios predefinidos de éxito [Shafique and Qaiser, 2014].

- Despliegue (Deployment): Finalmente, se implementan los modelos en el entorno operativo y se monitorean para garantizar que generen valor continuo para el negocio. Se

desarrollan planes de acción para la integración de los modelos en los procesos existentes y se establecen mecanismos de seguimiento para medir su desempeño a lo largo del tiempo [Shafique and Qaiser, 2014].

Iteratividad en CRISP-DM

Una característica fundamental de CRISP-DM es su enfoque iterativo, que reconoce la naturaleza dinámica y compleja de los proyectos de minería de datos. Las fases del proceso no son necesariamente secuenciales y pueden requerir revisión y repetición a medida que se adquiere un mayor entendimiento del problema y los datos disponibles. Esta iteratividad permite a los equipos de proyecto ajustar su enfoque y estrategia en función de los hallazgos y resultados obtenidos en etapas anteriores [Wirth and Hipp, 2000].

Roles y responsabilidades:

CRISP-DM define varios roles clave en un proyecto de minería de datos, cada uno con responsabilidades específicas [Martínez-Plumed et al., 2019]:

- Patrocinador del proyecto: Es la persona responsable de garantizar que el proyecto de minería de datos esté alineado con los objetivos estratégicos del negocio y que cuente con los recursos necesarios para su éxito [Martínez-Plumed et al., 2019].

- Analista de negocios: Se encarga de comprender los objetivos del negocio y traducirlos en problemas de minería de datos bien definidos. Es responsable de garantizar que los resultados del proyecto sean relevantes y accionables para la organización [Martínez-Plumed et al., 2019].

- Analista de datos: Es el encargado de recopilar, explorar y preparar los datos para su análisis. Utiliza diversas técnicas y herramientas para limpiar, integrar y transformar los datos en un formato adecuado para el modelado [Martínez-Plumed et al., 2019].

- Experto en minería de datos: Es el responsable de seleccionar y aplicar las técnicas de modelado de datos más adecuadas para resolver el problema planteado. Tiene experiencia en el uso de algoritmos de aprendizaje automático y estadísticas avanzadas [Martínez-Plumed et al., 2019].
  Administrador del proyecto: Es el encargado de coordinar y gestionar todas las actividades relacionadas con el proyecto de minería de datos. Se asegura de que se cumplan los plazos y presupuestos establecidos y de que se mantenga una comunicación efectiva entre todos los miembros del equipo [Martínez-Plumed et al., 2019].

Aplicaciones y beneficios de CRISP-DM:

CRISP-DM se ha utilizado con éxito en una amplia gama de aplicaciones y sectores industriales, incluyendo marketing, finanzas, salud, manufactura, entre otros. Algunos de los beneficios clave de utilizar CRISP-DM incluyen [Huber et al., 2019].

- Estructura y claridad: Proporciona una estructura clara y bien definida para guiar el proceso de minería de datos desde la comprensión inicial del problema hasta la implementación de soluciones prácticas [Huber et al., 2019].

- Flexibilidad y adaptabilidad: Es lo suficientemente flexible como para adaptarse a diferentes contextos y problemas, lo que permite a los equipos de proyecto ajustar su enfoque según sea necesario [Huber et al., 2019][Wiemer et al., 2019].

- Enfoque iterativo: Reconoce la naturaleza iterativa de los proyectos de minería de datos y permite a los equipos revisar y ajustar su enfoque a medida que adquieren un mayor entendimiento del problema y los datos disponibles
  [Huber et al., 2019] [Wiemer et al., 2019].

La metodología CRISP-DM proporciona un marco teórico sólido y estructurado para guiar proyectos de minería de datos de manera efectiva. Basado en principios científicos y prácticos, CRISP-DM abarca desde la comprensión inicial del negocio y los datos hasta la implementación práctica de soluciones y el monitoreo continuo de su desempeño. Su enfoque iterativo y su capacidad para adaptarse a diferentes contextos y problemas lo convierten en un estándar ampliamente utilizado en la industria de minería de datos, ayudando a las organizaciones a aprovechar el poder de los datos para tomar decisiones informadas y generar valor para el negocio.

2.2. Estado del arte

Este capítulo aborda las dos categorías centrales de esta investigación: el reconocimiento y clasificación de imágenes, y las redes neuronales convolucionales y transformadores de visión. A continuación, se presentan investigaciones previas desarrolladas en estos dos temas y que están directamente relacionadas con el tema de estudio: “Clasificación de estados emocionales mediante la aplicación de redes neuronales convolucionales, transformadores de visión y modelos de generación de imágenes en el análisis de rostros”. Los antecedentes permitieron crear un panorama de las diferentes posturas que presentan los autores frente a estas dos categorías y cómo dichos puntos de vista aportan a la construcción del conocimiento.

2.2.1. Reconocimiento y clasificación de imágenes

El reconocimiento de imágenes está dado por el proceso de identificar un objeto o una característica en una imagen o un vídeo. Utilizado en numerosas aplicaciones, como detección de defectos, gestión de imágenes médicas y vigilancia de la seguridad. En este sentido la ingeniería adquiere un rol importante, pues de su mano es posible gestionar el reconocimiento de imágenes y acelerar tareas tediosas en cuanto al procesamiento y análisis manual de las mismas. A continuación, se presentan los estudios de diversos autores en materia de reconocimiento de imágenes y detección facial.

En [Guo et al., 2013] estudiaron cómo las expresiones faciales transmiten señales de comunicación no verbal y proporcionan información acerca de las emociones humanas que juega un papel importante en las interacciones cara a cara, así como interacciones persona-máquina. Los Investigadores analizaron datos de píxeles sin procesar, extrayendo características de las imágenes, lo que permitió hacer una reducción de características y aplicación de la clasificación SVM (máquina de vectores de soporte). Es así como los autores lograron una mejor precisión de 0,89. Como resultado de su trabajo generaron una clasificación binaria para imágenes de “felicidad” o “enojo”.

Por su parte [Wang et al., 2019] investigaron sobre el reconocimiento de expresiones faciales y la riqueza de expresiones que existen, para ello desarrollaron un nuevo Marco de reconocimiento de expresión facial (FER) mediante la participación incorporar de las poses faciales en una imagen sintetizando y clasificando proceso de fijación. Este trabajo generó dos grandes novedades, en primera instancia, los autores crearon un nuevo conjunto de datos de expresiones faciales de más de 200.000 imágenes con 119 personas, 4 poses y 54 expresiones etiquetando cara a cara los cambios sutiles de emociones para fines de reconocimiento. En segunda instancia, gracias al volumen de datos analizados fue posible validar la teoría FER sobre poses, expresiones y tiró cero desequilibradas lo que permite una identificación de sujetos, proponiendo así una pose facial generativa.

Con [Anas et al., 2020] los autores realizaron un reconocimiento facial basado en redes neuronales convolucionales profundas, implementando un aprendizaje profundo basado en CNN. El modelo que usaron se ajustó utilizando el conjunto de datos FER2013. Actualmente, Aff-Wild2 es el conjunto de datos más grande y más reciente que proporciona la data necesaria (es decir, imagen, video, audio) para la expresión facial. Durante esta investigación basada en reconocimiento (FER) sobre estimación de excitación de valencia. Se estudió la detección de unidades de acción facial y siete expresiones básicas. Logrando como resultado potencial la adición de redes neuronales profundas (DNN) que toma información de puntos de referencia faciales y puede ser concatenado al modelo CNN para formar un híbrido CNN-DNN modelo para una clasificación de expresiones faciales más robusta. Durante este año [Chavan and Kulkarni, 2020] desarrollaron una optimización de la red neuronal convolucional profunda para el reconocimiento de expresiones faciales. Creando una red neuronal convolucional grande y profunda para clasificar 40.000 imágenes del conjunto de datos con siete

categorías: “asco”, “miedo”, “felicidad”, “enojo”, “tristeza”, “neutralidad”, “sorpresa”. En este proyecto aplicaron redes neuronales convolucionales (CNN) para el reconocimiento de expresiones faciales y desarrollaron el modelo en Theano y Caffe para el proceso de formación, logrando así un 61 % de precisión. Este trabajo presenta resultados de Implementación acelerada de las GPU CNN. Optimización profunda CNN que permite reducir el tiempo de entrenamiento del sistema.

En [Rasheed et al., 2022] estudiaron el reconocimiento facial de clasificación de emociones usando Haar en cascada y red neuronal convolucional. Los investigadores encontraron que en las últimas dos décadas, muchas aplicaciones relacionadas con las emociones faciales se han desarrollado y diferentes métodos se utilizan para clasificar emociones faciales humanas. El objetivo de su investigación fue clasificar emociones faciales humanas, es decir, “felicidad”, “tristeza”, “asco”, “enojo”, “miedo”, “neutralidad” o “sorpresa”. Trabajando con la clasificación Haar Cascade para detección de rostros y detección de emociones. Las características tipo Haar tienen particularidad de detectar bordes alrededor del objeto mientras que las características de línea y rectángulo se utilizan para detectar la línea inclinada del objeto. La técnica propuesta reconoce en el rostro emociones a través de redes neuronales convolucionales de aprendizaje profundo basadas en FER2013. El conjunto de datos FER2013 contiene 35887 entradas de siete emociones faciales y se clasifica como 28709 imágenes de entrenamiento, 3589 valores imágenes de datos y 3589 imágenes de prueba. El método propuesto consiste en seis capas convolucionales, tres capas de agrupación máxima y cuatro capas totalmente convolucionales. A través de la experimentación como resultado, el método propuesto logró una precisión de validación del 65,59 %.

Por su parte, [Abdulhussein and Saud, 2023] diseñaron un algoritmo híbrido de clasificación y reconocimiento de emociones en un rostro. Los autores argumentan que detectar y analizar el rostro humano representa una necesidad para muchas áreas como la medicina, la seguridad, la educación, entre otras. El reconocimiento facial de emociones (FER) es uno de los aspectos importantes del análisis de imágenes faciales que se necesita para muchas aplicaciones y tecnologías. En general, los trabajos de investigación en este ámbito se concentran en mejorar la precisión de la clasificación y disminuir el tiempo necesario para su detección. Los autores en su trabajo proponen un algoritmo híbrido FER basado en los algoritmos V-J (Viola-Jones) y CNN el cual se prueba, analiza, y genera como resultado una mejora en la eficiencia, la perspectiva de mejora y la posibilidad de intercambiar el rol entre los algoritmos V-J y CNN.

El grupo de autores en [Khare et al., 2024] realizó una revisión sistemática entre 2014 y 2023 sobre el reconocimiento de emociones e inteligencia artificial. Los autores definen el reconocimiento de emociones como la capacidad de inferir con precisión las emociones humanas a partir de numerosas fuentes y modalidades. Utilizando cuestionarios, señales físicas y señales fisiológicas, el reconocimiento de emociones ha ganado atención debido a sus diversas áreas de aplicación, como la informática afectiva, la atención sanitaria, la interacción entre humanos y robots. Con su trabajo de investigación quieren proporcionar una revisión exhaustiva y sistemática de la emoción. las técnicas de reconocimiento de la década actual y la aplicación

del reconocimiento de emociones mediante el uso físico y señales fisiológicas. Las señales físicas implican el habla y la expresión facial, mientras que las señales fisiológicas incluyen electroencefalograma, electrocardiograma, respuesta galvánica de la piel y seguimiento ocular. Los resultados proporcionan una introducción a varios modelos de emociones, estímulos utilizados para provocar emociones y antecedentes de las experiencias existentes.
Por su parte los investigadores en [Ahammed et al., 2023] desarrollaron un aprendizaje por meta transferencia para la emoción contextual. A partir de la detección en afirmación facial. La comunicación entre personas depende más de la expresión humana que de cualquier otro factor. La identificación de emociones tales como “felicidad”, “enojo”, “sorpresa”, “tristeza”, “asco”, “miedo” y “neutralidad” permiten a los seres humanos comunicar sus sentimientos a través de palabras, lenguaje corporal y expresiones faciales. La capacidad de los humanos para revelar procesos de pensamiento internos a través de expresiones faciales es crucial para estudiar su comportamiento. Muchos campos utilizan análisis de expresiones faciales, incluidos aquellos relacionados con una mayor seguridad. La mayoría de los estudios sobre clasificación de emociones sólo han utilizado modelos sencillos de CNN y RNN. Sin embargo, entrenar un conjunto de datos tan grande puede tomar mucho tiempo porque los modelos necesitan un vasto conjunto de datos. Para ello los autores sugieren un modelo que combina el modelo Mobile Net-V2 con la estrategia de aprendizaje por transferencia para acelerar el tiempo de generación y mejorar la precisión de la clasificación de emociones. Los investigadores compilaron un gran CIFE de datos de la literatura relevante. La cual fue llevada a detección mediante la arquitectura del modelo sugerido para ser probada en los datos mencionados anteriormente. El resultado muestra que el nuevo sistema debería poder identificar con mayor precisión que el anterior.
El grupo de investigadores en [Lasri et al., 2019] realizaron un análisis para clasificar las emociones faciales de los estudiantes mediante técnicas de Red neuronal convolucional. Los autores crearon un sistema que reconoce las emociones de los estudiantes a partir de sus caras. El sistema consta de tres fases: detección de rostros mediante Haar Cascades, normalización y reconocimiento de emociones mediante CNN en base de datos FER2013. Este análisis lo aplicaron a siete tipos de expresiones que generan los estudiantes cuando están recibiendo información por parte de un docente. Los resultados obtenidos muestran que el reconocimiento de emociones faciales es factible en educación, por lo tanto, puede ayudar a los docentes a modificar su presentación según las emociones de los estudiantes.
La mayoría de los autores citados, coincidieron en estudiar el grupo de siete emociones determinado por: “felicidad”, “tristeza”, “asco”, “enojo”, “miedo”, “neutralidad” o “sorpresa”, dado que estas son las emociones más empleadas por los seres humanos en sus proceso de comunicación e iteración. De igual forma la mayoría busca optimizar el tiempo de respuesta y entremedio de acuerdo a la combinación de técnicas aplicada y así precisar cada vez más los modelos aplicados a clasificación de emociones.

2.2.2. Redes neuronales convolucionales y transformadores de visión

Las redes neuronales son otra forma de emular ciertas características propias de los seres humanos, tales como la capacidad de memorizar y de asociar actividades. Si se revisa con atención aquellos problemas que no pueden expresarse a través de un algoritmo, se observará que todos ellos tienen una característica en común: la experiencia. Mediante la experiencia se generan patrones repetitivos que crean redes y permiten llegar a un resultado. Con las redes neuronales es posible simular estas características y obtener un nuevo sistema para el tratamiento de la información. A continuación, se exponen diferentes estudios que relacionan la aplicación de las redes neuronales en campos como el agro y la salud.

En [Sánchez, 2018] el autor desarrolló una técnica para la recuperación de imágenes por contenido usando descriptores generados por redes neuronales convolucionales. El investigador desarrolló un método para la recuperación de imágenes indexadas en bases de datos a partir de su contenido visual, sin necesidad de realizar anotaciones textuales. Logrando vectores de rasgos a partir de los contenidos visuales mediante técnicas de redes neuronales artificiales con aprendizaje profundo. Sánchez propone el empleo de redes neuronales convolucionales pre entrenadas para crear los descriptores globales. Los resultados obtenidos por el método propuesto, sobre bases de datos disponibles públicamente, fueron superiores a los de los métodos tradicionales y comparables con otros basados en aprendizaje profundo, que constituyen el estado del arte en la recuperación de imágenes por contenido.

Por su parte [Centeno et al., 2023] diseñaron una herramienta de corte para optimizar parámetros de clasificación de especies maderables con redes neuronales convolucionales. La gran diversidad de especies maderables tropicales demanda el desarrollo de nuevas tecnologías de identificación con base en sus patrones o características anatómicas. La aplicación de redes neuronales convolucionales (CNN) para el reconocimiento de especies maderables tropicales se ha incrementado en los últimos años por sus resultados prometedores. Los autores evaluaron la calidad de las imágenes macroscópicas con tres herramientas de corte para mejorar la visualización y distinción de las características anatómicas en el entrenamiento del modelo CNN. Iniciaron con la recolección de muestras entre el 2020 y 2021 en áreas de explotación forestal y aserraderos de la selva en Perú, luego, cortaron las piezas en secciones transversales a fin de generar una base de datos de imágenes macroscópicas de la sección transversal de la madera. Como resultado obtuvieron 3 750 imágenes macroscópicas con un microscopio portátil que corresponden a 25 especies maderables. Logrando determinar que la calidad de las imágenes es decisiva en la clasificación de especies maderables, porque permite una mejor visualización y distinción de las características anatómicas en el entrenamiento con los modelos de red neuronal convolucional EfficientNet B0 y Custom Vision, lo cual se evidenció en las métricas de precisión.

El grupo de autores en [Juárez Trujillo et al., 2023] realizaron la calibración de una cámara multiespectral utilizando redes neuronales convolucionales. Su análisis presenta una metodología para la calibración de cámara multiespectral mediante redes neuronales convoluciona-

les. Donde capturaron imágenes en RGB de la cámara multiespectral para cada uno de los estándares, las muestras son tomadas con las mismas condiciones de iluminación y con el mismo ángulo de captura. Las imágenes son fragmentadas en pequeños tamaños de matrices agregados a una clase específica, y guardada con una etiqueta especial para distinguirla de toda la base de datos de la clase, el mismo proceso lleva los 7 cerámicos de color restantes Lucideon Std. Uno de los cerámicos corresponderá a una clase en particular con una dimensión igual para todas las clases. Finalmente, con base a la metodología presentada los autores logran calibrar la cámara con respecto a las referencias.

Los investigadores en [Plazas López et al., 2022] diseñaron un modelo predictivo computacional que facilita el reconocimiento de la lengua de señas colombiana (LSC) en un entorno hotelero y turístico. Aplicaron técnicas de inteligencia artificial y redes neuronales profundas en el aprendizaje y la predicción de gestos en tiempo real, los que les permitió construir una herramienta para disminuir la brecha y fortalecer la comunicación. Los autores implementaron algoritmos de redes neuronales convolucionales sobre captura de datos en tiempo real. Capturaron el movimiento mediante cámaras de vídeo de dispositivos móviles; así, obtuvieron las imágenes que forman el conjunto de datos. Las imágenes se utilizaron como datos de entrenamiento para un modelo computacional óptimo que puede predecir el significado de una imagen recién presentada. Como resultado se obtuvo la evaluación del rendimiento del modelo usando medidas categóricas y comparando diferentes configuraciones para la red neuronal. Adicional a esto, todo está soportado con el uso de herramientas como Tensorflow, OpenCV y MediaPipe. Logrando un modelo capaz de identificar y traducir 39 señas diferentes entre palabras, números y frases básicas enfocadas al sector hotelero, con una tasa de éxito del 97,6 % en un ambiente de uso controlado.

En [Bird and Lotfi, 2024] se empleó una combinación de técnicas avanzadas, incluyendo la generación de datos sintéticos con Latent Diffusion, el uso de redes neuronales convolucionales (CNN) para la clasificación de imágenes, y la interpretación de predicciones a través de Gradient Class Activation Mapping (Grad-CAM). Estas metodologías permitieron alcanzar una precisión del 92.98 % en la clasificación binaria entre imágenes reales y generadas por IA, proporcionando una sólida base para la mejora de la capacidad de reconocimiento en un contexto donde las imágenes generadas por IA desafían los límites de la percepción humana. Además, se lanzó el conjunto de datos CIFAKE, que comprende 120,000 imágenes, para facilitar investigaciones futuras en el campo de la visión por computadora.

Según lo establecido en [Ahmadi, 2023], las actividades fraudulentas dirigidas a tarjetas resultaron en una pérdida global de $32.34 mil millones en 2021, con un aumento del 14 % con respecto al año anterior. Las instituciones financieras están adoptando tecnologías de aprendizaje automático y automatización como OpenAI para combatir este problema. Sin embargo, los defraudadores digitales están utilizando tácticas sofisticadas, como el phishing y el aprovechamiento de OpenAI para desarrollar información engañosa. Por lo tanto, las técnicas de detección de fraude deben evolucionar constantemente, integrando diferentes herramientas de aprendizaje automático para construir sistemas de seguridad más robustos. La utilización

responsable de OpenAI en el ámbito financiero requiere una mayor conciencia y competencia para maximizar sus beneficios y mitigar los riesgos de explotación por parte de actores malintencionados.
En [Bhat and Jain, 2023] se establece que los modelos CLIP (Contrastive Language-Image Pre-training) desarrollados por OpenAI han logrado resultados sobresalientes en diversas tareas de reconocimiento y recuperación de imágenes, exhibiendo una fuerte capacidad de rendimiento en cero-shot. Se describe la creación del conjunto de datos LAION5B, que condujo al desarrollo de los modelos open ViT-H/14, ViT-G/14 que superan al modelo L/14 de OpenAI. Además, se evaluó el rendimiento de varios modelos CLIP como reconocedores de rostros en cero-shot, encontrando que los modelos CLIP se desempeñan bien en tareas de reconocimiento facial, pero que aumentar el tamaño del modelo CLIP no necesariamente conduce a una mayor precisión. También se investigó la robustez de los modelos CLIP contra ataques de envenenamiento de datos, concluyendo que son robustos contra estos ataques. Se subraya la importancia de considerar las posibles consecuencias y el mal uso de los motores de búsqueda construidos utilizando modelos CLIP, que podrían funcionar inadvertidamente como motores de reconocimiento facial no intencionales, planteando importantes cuestiones legales y éticas que deben abordarse.
En [Pomazan et al., 2023] se explora el uso de redes neuronales convolucionales (CNN) para el reconocimiento de emociones en el marketing publicitario. Se discute cómo las CNN pueden entrenarse para reconocer las emociones en anuncios y cómo esto puede mejorar la efectividad de las campañas publicitarias. Los resultados muestran que las CNN pueden identificar correctamente las emociones en el 88.25 % de los casos. Se planea mejorar los datos y abordar la sensibilidad al ruido y la precisión en diferentes grupos demográficos. La aplicación desarrollada tiene ventajas en precisión y costo, pero requiere mejoras en la sensibilidad y la precisión en la clasificación de emociones.

2.3. Marco contextual

Los estados emocionales juegan un papel importante en la interacción y el bienestar humanos. Comprender y clasificar con precisión estos estados emocionales puede tener profundas implicaciones para diversos campos, incluida la psicología, la salud y la interacción con el entorno. Se clasifican estados emocionales mediante la aplicación de redes neuronales convolucionales (CNN) y transformadores de visión (ViT).
La captura de imágenes de rostros de personas con su consentimiento permite analizar las expresiones faciales y extraer datos valiosos sobre sus estados emocionales. Donde es esencial considerar las implicaciones éticas y las preocupaciones de privacidad asociadas con la captura y el análisis de imágenes faciales de las personas. Al abordar este contexto y aprovechar técnicas avanzadas de aprendizaje automático, se pretende estudiar un sistema fiable y eficaz para clasificar los estados emocionales basado en el análisis facial, como medio para

mejorar la precisión y confianza en la clasificación para promover el bienestar emocional y mejorar la calidad de vida.
Se utilizará la base de datos FER2013. Esta “se generó utilizando la API de búsqueda de imágenes de Google y se presentó durante los Desafíos ICML 2013 (International Conference on Machine Learning)” [Lasri et al., 2019].
Las imágenes de la base de datos se encuentras normalizadas a un tamaño de 48x48 píxeles. La base de datos FER2013 contiene 35887 imágenes con 7 clasificaciones de emoción donde 28709 de imágenes son de entrenamiento, 3589 de imágenes son de validación y 3589 de imágenes son de prueba. Allí, se presenta el reto de un desbalance en la distribución de las clases, algo habitual en la visión por computadora.
Este desequilibrio ocurre cuando ciertas clases están subrepresentadas en comparación con otras, lo que puede llevar al modelo de aprendizaje automático a desarrollar un sesgo hacia las clases más frecuentes y afectar negativamente su capacidad para generalizar bien a nuevas instancias de datos, especialmente aquellas pertenecientes a las clases minoritarias.

3 Desarrollo de los modelos y resultados

Este capítulo se enfoca en el desarrollo y la evaluación de modelos de clasificación de emociones basados en redes neuronales convolucionales (CNN) y transformadores de visión (ViT). En él se describen la preparación del conjunto de datos para ambos modelos, el entrenamiento y evaluación de los mismos y se presentan los resultados obtenidos.

La exploración en el campo del reconocimiento de emociones faciales ha evolucionado significativamente con el desarrollo de arquitecturas de red avanzadas. Los modelos basados en redes neuronales convolucionales (CNN) han demostrado ser poderosos para tareas de visión por computadora, incluyendo la interpretación de las emociones humanas a partir de expresiones faciales. Sin embargo, los estudios han revelado ciertas limitaciones en el enfoque de CNN, particularmente en cuanto a la necesidad de un preprocesamiento extenso y la captura de dependencias a largo plazo en los datos. En contraste, los transformadores de visión (ViT) ofrecen un nuevo paradigma mediante el cual la atención global puede aplicarse a las representaciones de imágenes, potencialmente superando algunas de las barreras inherentes a las CNN.

En este trabajo se propuso examinar el potencial de los ViT en el contexto del reconocimiento emocional, motivado por su exitosa aplicación en tareas de procesamiento de lenguaje natural y su emergente prominencia en el análisis de imágenes. El uso de ViT se alinea con la dirección actual de la investigación, que apunta hacia modelos que pueden aprender características jerárquicas y dependencias complejas más allá de las capacidades de las CNN. Con el objetivo de superar los desafíos identificados en el desempeño del modelo CNN, este trabajo explora cómo la arquitectura distintiva de los ViT puede capturar matices y sutilezas en las emociones faciales con una precisión mejorada y una generalización más robusta. Al hacerlo, se busca contribuir al cuerpo de conocimiento existente con una metodología refinada y perspectivas mejoradas en la clasificación precisa de emociones.

Se utilizan matrices de confusión para una evaluación cuantitativa del desempeño, revelando cómo los modelos confunden ciertas emociones, y se despliega y utiliza el prototipo funcional para una evaluación cualitativa, proporcionando pistas para futuras mejoras. La visualización de estos datos es clave para comprender el desempeño de los modelos, ofreciendo una ventana al aprendizaje de máquina.

Con una comprensión detallada de la estructura y eficacia del modelo, se adentra en la discusión de sus implicaciones. Este segmento no solo documenta el trabajo, sino que también sienta las bases para un análisis más profundo, abriendo camino hacia la reflexión sobre estos hallazgos en el contexto más amplio de la inteligencia artificial y el análisis computacional

de emociones humanas.

3.1. Preparación del conjunto de datos

El sobremuestreo aleatorio, o “Random Over Sampling”, es una técnica utilizada para abordar el problema del desequilibrio de clases en los conjuntos de datos de aprendizaje automático [Lemaître et al., 2017].
En este caso, el conjunto de datos presentaba un marcado desequilibrio. La emoción “felicidad” era predominante con 8989 instancias, en contraste con la escasez de representación de “asco”, que solo tenía 547 instancias. Sin un tratamiento adecuado, cualquier modelo entrenado en este conjunto de datos tendería a reconocer mejor la emoción “felicidad” y desempeñarse deficientemente en la detección de “asco”.

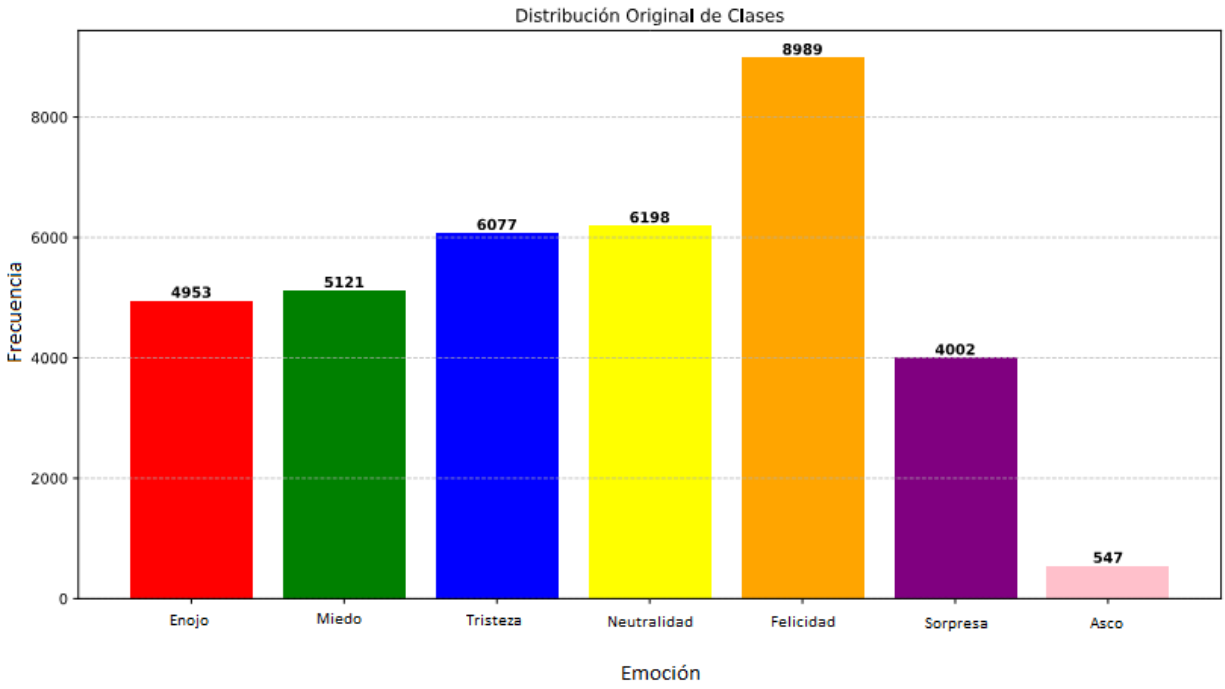

Figura 3-1: Distribución desequilibrada de las clases en el conjunto de datos FER2013.

Para rectificar este desbalance, se aplicó el sobremuestreo aleatorio, que incrementa el número de muestras en las clases minoritarias replicándolas aleatoriamente hasta alcanzar un número comparable de instancias en todas las clases. De esta manera, las clases minoritarias reciben un refuerzo en su representatividad, permitiendo que el modelo aprenda con una exposición más uniforme a cada clase.
Este procedimiento igualó la representación de clases en el conjunto de datos, con aproximadamente 6300 instancias por emoción, generando una distribución uniforme y facilitando

que el modelo desarrolle una capacidad de generalización más robusta. Al permitir que el modelo aprenda con la misma intensidad de cada clase, se mejora su capacidad de clasificar las emociones de forma más equitativa y precisa.

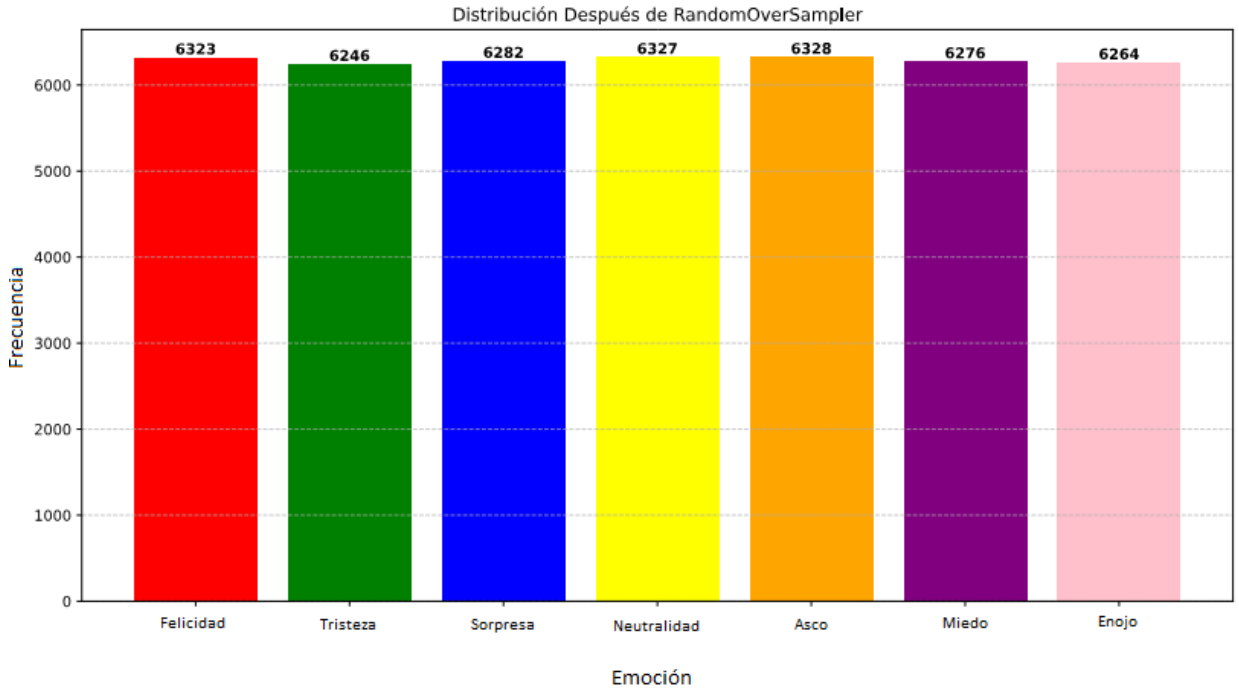

Figura 3-2: Distribución equilibrada de las clases para el conjunto de datos FER2013.

3.2. Desarrollo del modelo de clasificación de emociones basado en CNN

En el presente estudio, se adoptó una arquitectura de red neuronal convolucional (CNN) para el análisis y clasificación de imágenes faciales, con el objetivo de identificar siete emociones humanas distintivas. La arquitectura se compone de una secuencia jerarquizada de capas convolucionales y de agrupamiento (pooling), seguidas por capas totalmente conectadas (fully connected layers), optimizadas para la tarea de clasificación multiclase, como se ilustra en la figura “Gráfico de arquitectura de una red neuronal convolucional (CNN)” en el capítulo de marco teórico y estado del arte.

Las capas convolucionales, dotadas de filtros o núcleos de convolución, tienen la función primordial de detectar patrones espaciales locales en las imágenes, tales como bordes, texturas y otros elementos visuales básicos en sus capas iniciales, y características de mayor nivel de abstracción en las capas subsiguientes. Las operaciones de agrupamiento subsiguiente reducen la dimensionalidad espacial de las representaciones, incrementando la invarianza a

las traslaciones y distorsiones menores en la imagen.
Posteriormente, las capas totalmente conectadas interpretan las características abstractas extraídas, conduciendo a la clasificación final. La red emplea la función de activación ReLU (Rectified Linear Unit) por su eficiencia computacional. La última capa de la red, antes de la salida, no implementa ReLU, dado que sus resultados alimentan una función de activación softmax que distribuye la probabilidad a lo largo de las clases de emociones.

```python
class SimpleCNN(nn.Module):
    def __init__(self, num_classes=7):
        super(SimpleCNN, self).__init__()
        self.conv1 = nn.Conv2d(1, 32, kernel_size=3, padding=1)
        self.conv2 = nn.Conv2d(32, 64, kernel_size=3, padding=1)
        self.fc1 = nn.Linear(64 * 12 * 12, 128)
        self.fc2 = nn.Linear(128, num_classes)

    def forward(self, x):
        x = F.relu(F.max_pool2d(self.conv1(x), 2))
        x = F.relu(F.max_pool2d(self.conv2(x), 2))
        x = x.view(x.size(0), -1)
        x = F.relu(self.fc1(x))
        x = self.fc2(x)
        return x
```

Código 3.1: Implementación de SimpleCNN en Python

Capas Conv2d (self.conv1 y self.conv2)

- Capa self.conv1: Esta es la primera capa convolucional que toma imágenes en escala de grises como entrada (1 canal de entrada) y aplica 32 filtros (o núcleos). El kernel_size=3 significa que cada filtro tiene un tamaño de 3x3, y padding=1 asegura que la salida de la convolución tenga las mismas dimensiones espaciales que la entrada al agregar un borde de ceros alrededor de la entrada. Esta capa extrae características de bajo nivel como bordes, ángulos y texturas.

- Capa self.conv2: La segunda capa convolucional incrementa la profundidad a 64 filtros, permitiendo que la red capture una gama más rica de características. Al igual que la primera capa, utiliza un kernel_size=3 y padding=1. Al aumentar el número de filtros desde 32 a 64, la red puede construir representaciones más complejas y abstractas de los datos de entrada.

Capas de Pooling

- Pooling en self.conv1(x): Después de la primera convolución, se aplica una operación de "max pooling" con un tamaño de ventana de 2. Esto reduce las dimensiones espaciales de la salida a la mitad, lo que ayuda a hacer que la representación sea más compacta y más resistente a pequeñas variaciones en la posición de las características dentro de la imagen.
- Pooling en self.conv2(x): Similar a la capa anterior, se aplica "max pooling" para reducir aún más las dimensiones espaciales. Esta reducción ayuda a disminuir la cantidad de parámetros y la computación en las capas subsiguientes, contribuyendo a controlar el sobreajuste.

Capas Fully Connected (self.fc1 y self.fc2)

- Capa self.fc1: Antes de esta capa, se aplana la salida de la última capa de pooling para convertirla en un vector lineal. Esta capa densa transforma el vector aplanado de dimensiones (64 * 12 * 12) a un espacio de características de 128 dimensiones. La elección de 128 unidades es un balance entre la capacidad de la red y la complejidad computacional, permitiendo suficiente flexibilidad para aprender relaciones complejas sin ser excesivamente propensa al sobreajuste.
- Capa self.fc2: La última capa densa mapea el espacio de características de 128 dimensiones al número de clases objetivo (num_classes=7 en este caso). Proporciona la puntuación de cada una de las 7 emociones posibles en el conjunto de datos FER2013.

Función de activación ReLU

En ambas capas convolucionales y la primera capa densa, se utiliza la función de activación ReLU (F.relu). ReLU ayuda a introducir no linealidades en el modelo, permitiendo que la red aprenda relaciones complejas entre las características y las clases. Además, ReLU tiene el beneficio de ser computacionalmente eficiente y reducir el problema de desvanecimiento del gradiente en comparación con otras funciones de activación.

Flujo de datos en forward

La función forward define cómo fluyen los datos a través de la red. Comienza con la entrada x pasando sucesivamente por las capas convolucionales y de pooling, luego se aplana y finalmente pasa a través de las capas densas. El flujo termina con la salida de self.fc2(x), que son los logits (puntuaciones sin escalar) para cada clase.

Función de activación softmax

La función de activación softmax distribuye la probabilidad a lo largo de las clases de emociones y es seguida por la salida, que genera la clasificación.

3.3. Desarrollo del modelo de clasificación de emociones basado en ViT

En el presente proyecto se realizó la implementación y evaluación de un modelo de clasificación de emociones basado en transformadores de visión. Se utilizó el conjunto de datos 'FER2013', tras un preprocesamiento adecuado, que incluyó la normalización y aumento de datos para mejorar la generalización del modelo, se realizó la configuración de los transformadores para procesar secuencias de parches de imágenes.
Los transformadores de visión (ViT) representan una clase de modelos basados en la mecánica de atención, que se originó en el procesamiento del lenguaje natural y ha ganado tracción en el procesamiento de imágenes. A diferencia de las CNN, que procesan imágenes a través de la convolución local, ViT divide las imágenes en parches y utiliza mecanismos de atención para capturar las interacciones globales entre ellos, como se ilustra en la figura "Gráfico de arquitectura de transformadores de visión (ViT)" en el capítulo de marco teórico y estado del arte..
Se procedió a realizar un ajuste profundo de los hiperparámetros y se empleó validación cruzada para asegurar la robustez del modelo. El entrenamiento, la validación y las pruebas se llevaron a cabo utilizando la siguiente infraestructura de computación de alto rendimiento para facilitar una evaluación exhaustiva del modelo ViT.

Preprocesamiento de datos

Se estandarizan las imágenes del conjunto de datos "FER2013" a una escala de grises de 48x48 píxeles. Para aumentar la variabilidad y mejorar la generalización, se aplicaron técnicas de aumento de datos, como rotaciones aleatorias, volteos horizontales y ajustes de brillo.

Arquitectura del modelo

Para este estudio, se eligió el modelo 'vit-base-patch16-224-in21k' [Deng et al., 2009]; [Dosovitskiy et al., 2020]; [Wu et al., 2020] de Google, un transformador preentrenado que ha demostrado ser excepcionalmente capaz en tareas de visión por computadora. Este modelo se distingue por su arquitectura que procesa imágenes en parches de 16x16 y su entrenamiento previo en más de 21k categorías de imágenes, lo que proporciona una comprensión extensa de características visuales.

Afinamiento del modelo preentrenado

El proceso de transferencia de aprendizaje se implementó para afinar el ViT preentrenado en el conjunto de datos específico de emociones faciales. La etapa de ajuste fino implicó la reconfiguración de la última capa del modelo para alinearla con las siete categorías emocionales del conjunto de datos 'FER2013'. Además, se realizó un cuidadoso ajuste de hiperparámetros para maximizar el rendimiento del modelo en esta tarea de clasificación particular.

Entrenamiento del modelo

Para el entrenamiento, se utilizó el optimizador Adam con una tasa de aprendizaje inicial de 1e-4 y un decay exponencial. Se emplearon mecanismos de parada temprana basados en la pérdida de validación para prevenir el sobreajuste.

Regularización y ajuste de hiperparámetros

Se introducen técnicas de regularización, como dropout con una tasa de 0.1 y regularización L2, para mejorar la robustez del modelo. Los hiperparámetros se afinaron mediante un proceso iterativo, guiado por la optimización de la pérdida de validación.

Validación y pruebas

Se realizó la división del conjunto de datos en 80 % para entrenamiento, 10 % para validación y 10 % para pruebas. La eficacia del modelo se evaluó a través de la precisión y el puntaje F1 en el conjunto de pruebas, después de completar el entrenamiento con el conjunto de validación.

3.4. Resultados de la clasificación de emociones mediante CNN

La selección de hiperparámetros óptimos, un proceso crítico para la convergencia y generalización del modelo, se llevó a cabo mediante un enfoque de optimización bayesiana implementado por Optuna [Akiba et al., 2019] en el proceso de entrenamiento del modelo, con una tasa de aprendizaje (Learning Rate) de 0.0014 y un decaimiento de peso (Weight Decay) de 1.65e-06. Este enfoque sistemático permitió la exploración del espacio de hiperparámetros y la identificación de la tasa de aprendizaje y la regularización óptimas para el modelo, minimizando la función de pérdida y maximizando la validación cruzada del rendimiento.
La eficiencia del modelo se evaluó con métricas estandarizadas en aprendizaje automático, incluyendo precisión, sensibilidad (recall), y la medida F1, que equilibra la precisión y la sensibilidad. La matriz de confusión se utilizó como una herramienta visual para discernir la

capacidad del modelo para clasificar correctamente cada una de las emociones, así como para identificar patrones sistemáticos de error.

La evaluación del rendimiento del modelo entrenado en el conjunto de datos FER2013 se efectuó mediante la medición de la exactitud (accuracy), la sensibilidad (recall) y la puntuación F1, proporcionando una perspectiva holística de su desempeño general. La exactitud alcanzada fue del 63 %, lo que indica que el modelo pudo clasificar correctamente más de la mitad de las instancias presentadas, Lo cual es congruente con la precisión encontrada en la literatura donde [Chavan and Kulkarni, 2020] logran un 61 % de precisión y [Rasheed et al., 2022] un 65 % de precisión aplicando variaciones de CNN para clasificación de emociones. Esta métrica, aunque supera el umbral del azar para un problema de clasificación multiclase, sugiere que aún hay margen para mejorar la capacidad predictiva del modelo.

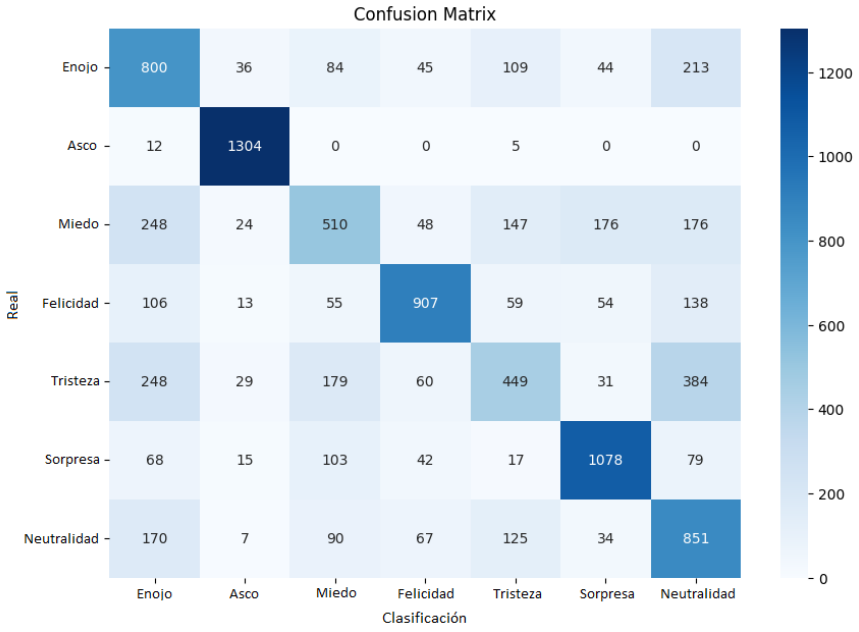

Figura 3-3: Matriz de confusión para el modelo CNN post-optimización de hiperparámetros.

Emoción	Precisión	Recall	F1-Score	Soporte
Enojo	0.48	0.60	0.54	1331
Asco	0.91	0.99	0.95	1321
Miedo	0.50	0.38	0.43	1329
Felicidad	0.78	0.68	0.73	1332
Tristeza	0.49	0.33	0.39	1380
Sorpresa	0.76	0.77	0.76	1402
Neutralidad	0.46	0.63	0.53	1344
Exactitud				0.62
Promedio macro				0.63
Promedio ponderado				0.62

Tabla 3-1: Informe de clasificación para el modelo CNN post-optimización de hiperparámetros.

La sensibilidad o recall, también del 62.5 %, revela que el modelo tiene una habilidad moderada para identificar correctamente todas las instancias positivas de cada clase. Esto es relevante en aplicaciones prácticas donde el costo de no detectar una emoción correcta puede ser significativo. La puntuación F1, que equilibra la precisión y la sensibilidad, fue de aproximadamente 61.9 %. Este es un indicador crítico en contextos donde un equilibrio entre la precisión de la detección y la capacidad para recuperar todas las instancias relevantes es esencial.

Al inspeccionar las matrices de confusión se evidencian tendencias claras en la clasificación. En el caso de las emociones más frecuentes como “felicidad”, se observa una alta tasa de acierto, corroborada por una precisión de 0.91 y una sensibilidad de 0.99 en el conjunto de validación. Esta alta eficacia puede atribuirse a la naturaleza distintiva de las características visuales asociadas con la “felicidad”, que son posiblemente más fáciles de aprender para el modelo.

Sin embargo, para emociones con expresiones faciales más sutiles o variadas, como “Tristeza” y “miedo”, el modelo exhibe una mayor confusión, como lo demuestra el número de falsos positivos y falsos negativos en las matrices de confusión. La clasificación precisa de estas emociones requiere que el modelo aprenda a identificar características visuales que a menudo son menos evidentes o más susceptibles a variaciones individuales.

Las métricas por clase destacan los desafíos asociados con cada emoción específica. Por ejemplo, la emoción “asco” muestra una excelente precisión y sensibilidad, lo cual es prometedor pero debe interpretarse con precaución dado el pequeño número de muestras de esta clase en el conjunto de datos. Por otro lado, la baja puntuación F1 para “miedo” refleja dificultades en la clasificación correcta, lo que sugiere una necesidad de enfoques adicionales como el aumento de datos o la ingeniería de características más avanzada para estas emociones específicas.

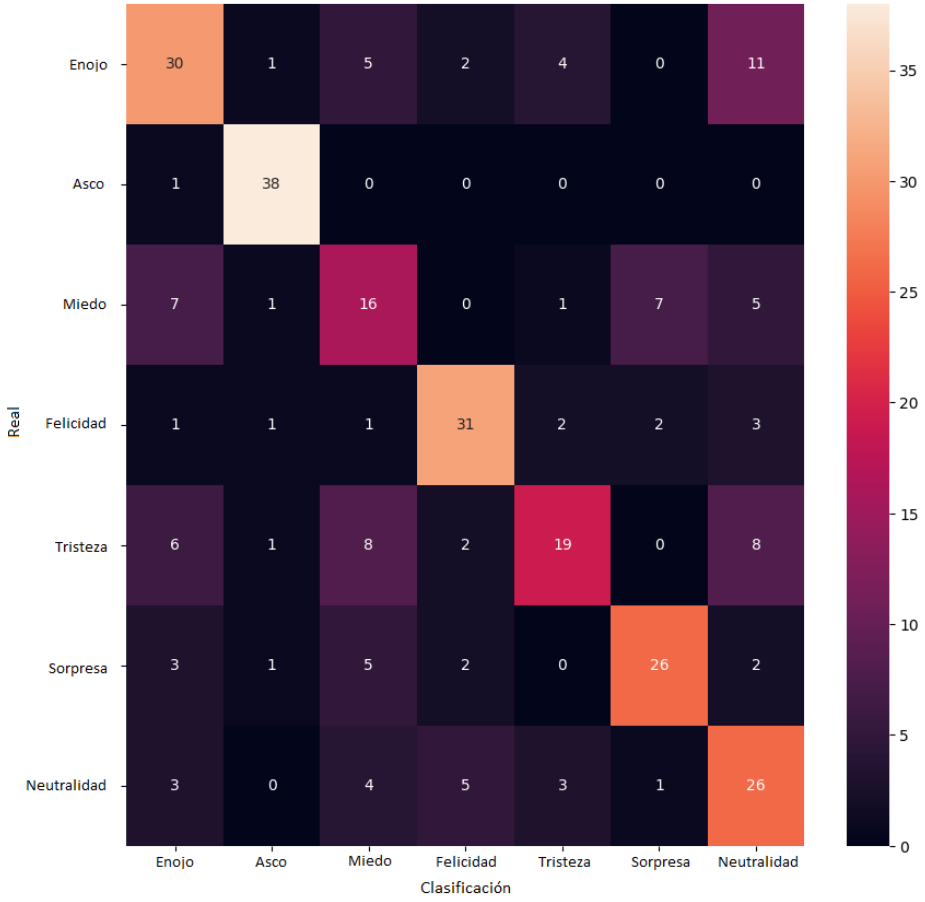

Figura 3-4: Matriz de confusión del conjunto de validación para el modelo CNN.

Emoción	Precisión	Recall	F1-Score	Soporte
Enojo	0.59	0.57	0.58	53
Asco	0.88	0.97	0.93	39
Miedo	0.41	0.43	0.42	37
Felicidad	0.74	0.76	0.75	41
Tristeza	0.66	0.43	0.52	44
Sorpresa	0.72	0.67	0.69	39
Neutralidad	0.47	0.62	0.54	42
Exactitud				0.63
Promedio macro				0.64
Promedio ponderado				0.63

Tabla 3-2: Informe de clasificación del conjunto de validación para el modelo CNN.

En resumen, los resultados obtenidos ofrecen una comprensión valiosa de la capacidad del

modelo “SimpleCNN” para el reconocimiento de emociones. Si bien se demuestra una competencia razonable en la clasificación general, los patrones identificados en las matrices de confusión y las métricas detalladas señalan oportunidades específicas para mejorar la arquitectura del modelo, el preprocesamiento de datos y las técnicas de entrenamiento. Estos hallazgos sientan las bases para futuras iteraciones y mejoras en la búsqueda de un modelo más preciso y equitativo en la clasificación de emociones humanas.
Comparativamente, el modelo actual se alinea con los hallazgos de investigaciones previas en términos de los desafíos planteados por emociones menos representadas o más ambiguas. Sin embargo, destaca por su enfoque de balanceo de clases, que es menos común en estudios similares y representa una contribución metodológica significativa a la disciplina.
La discusión se extiende para comparar el rendimiento del modelo con otros trabajos relacionados, tomando en cuenta las diferencias en arquitecturas de red, técnicas de preprocesamiento y métodos de optimización de hiperparámetros. Este análisis comparativo no solo valida las elecciones metodológicas sino que también identifica áreas de oportunidad para futuras iteraciones del modelo.

3.5. Resultados de la clasificación de emociones mediante ViT

Los resultados del entrenamiento del modelo transformador de visión muestran una mejora constante en la precisión y una disminución en la pérdida a lo largo de las épocas. Específicamente, el modelo alcanzó una precisión del 75.99 % en el conjunto de pruebas después de 10 épocas, con una tendencia general de reducción en la pérdida de validación, indicando un aprendizaje efectivo y una buena generalización. La matriz de confusión revela un rendimiento sólido en categorías como “felicidad” y “asco”, pero sugiere la necesidad de mejorar la identificación de categorías como “tristeza” y “miedo”. A continuación, se presenta el informe de clasificación detallado, junto con una representación gráfica de la matriz de confusión y la evolución del entrenamiento a lo largo de las épocas.

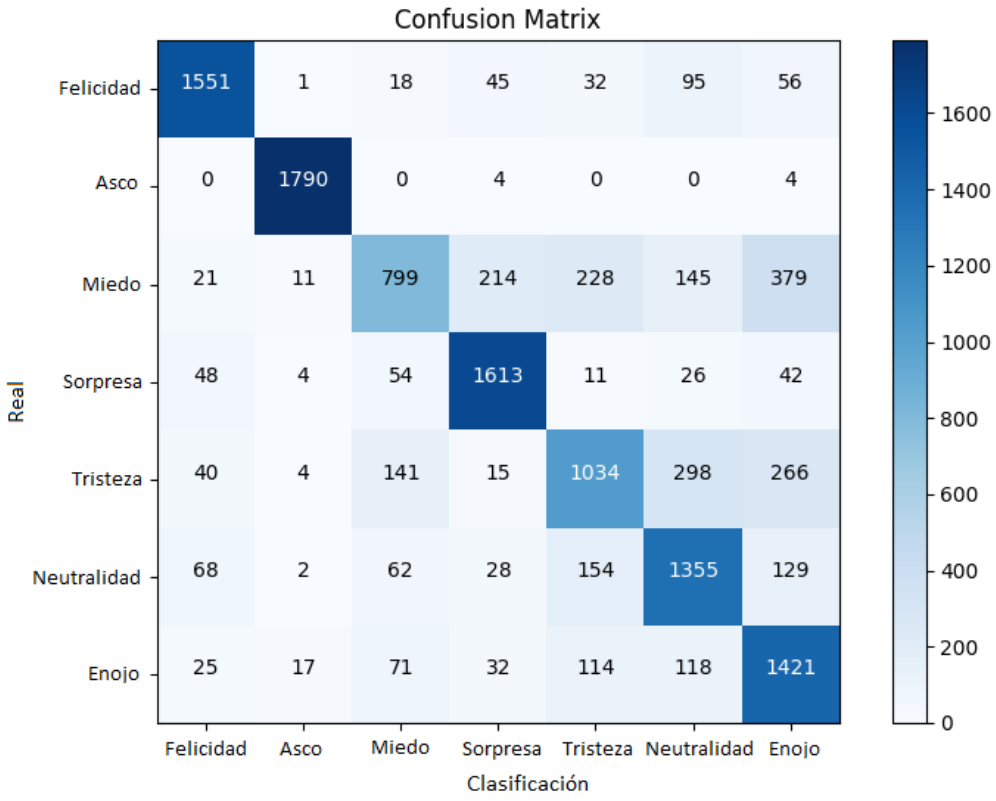

Figura 3-5: Matriz de confusión para el modelo ViT.

Emoción	Precisión	Recall	F1-Score	Soporte
Felicidad	0.8848	0.8626	0.8736	1798
Asco	0.9787	0.9956	0.9870	1798
Miedo	0.6978	0.4446	0.5432	1797
Sorpresa	0.8268	0.8971	0.8605	1798
Tristeza	0.6573	0.5751	0.6135	1798
Neutralidad	0.6652	0.7536	0.7066	1798
Enojo	0.6186	0.7903	0.6940	1798
Media/Total	0.7613	0.7599	0.7541	12585

Tabla 3-3: Informe de clasificación para el modelo ViT.

Los resultados reflejan un desempeño robusto en la clasificación de emociones, como se evidencia en la precisión general y las métricas F1. El modelo ViT, con su arquitectura única y mecanismo de atención global, ha mostrado una habilidad notable para discernir entre distintas emociones, al igual que se encontró en la literatura donde [Bhat and Jain, 2023] logran

un 89 % de precisión aplicando modelos ViT a tareas de reconocimiento y recuperación de imágenes.
La sensibilidad o recall del modelo, calculada de la misma manera que para el modelo 'SimpleCNN', resalta la competencia del ViT para identificar instancias positivas. Esto es especialmente significativo en aplicaciones donde el reconocimiento preciso de la emoción es crítico. El puntaje F1, que se mantiene como una métrica crucial para balancear la precisión y la sensibilidad, muestra una mejora con el uso de ViT comparado con el modelo basado en CNN, indicando que ViT es particularmente efectivo en contextos donde se busca un compromiso entre estas dos métricas.
Las matrices de confusión de ViT muestran una disminución en la confusión entre emociones complejas como "tristeza" y "miedo", lo que podría atribuirse a la capacidad del modelo de capturar mejor las dependencias globales en las imágenes. Las emociones con características visuales distintivas, como "felicidad", se benefician de la representación de atención detallada de ViT, lo que posiblemente resulta en tasas de acierto aún más altas que las obtenidas con las CNN.
Sin embargo, es importante notar que, a pesar de los avances, las emociones con variaciones sutiles en la expresión facial todavía representan un desafío. Las visiones transformadoras, aunque avanzadas, no están exentas de las dificultades presentadas por la variabilidad individual y las sutilezas en las expresiones faciales. Las métricas por clase revelan que, aunque algunas emociones como "asco" pueden tener altas puntuaciones de precisión y recall, esto podría estar influido por el desequilibrio de clases en el conjunto de datos.
En resumen, el rendimiento mejorado de los ViT sugiere que su enfoque de atención y aprendizaje de representaciones profundas es ventajoso para la clasificación de emociones faciales. No obstante, sigue existiendo la necesidad de mejorar la capacidad del modelo para lidiar con emociones de expresiones más sutiles, y esto podría explorarse a través de métodos de preprocesamiento innovadores, aumento de datos, o la incorporación de datos más diversos para entrenamiento.

3.6. Despliegue con prototipo funcional

El objetivo principal de este despliegue es el desarrollar una aplicación para la clasificación de emociones en imágenes, la aplicación permite a los usuarios subir una imagen y luego utiliza los modelos basados en CNN y ViT para clasificar las emociones presentes en la imagen y de esta manera realizar la evaluación cualitativa de estos modelos de clasificación.

A continuación, se detallan los principales componentes y funcionalidades del prototipo funcional:

- Estructura del proyecto: Se utilizó una estructura modular para organizar el código del proyecto, separando las diferentes responsabilidades en módulos y paquetes. El archi-

vo main.py es el punto de entrada de la aplicación FastAPI, donde se configura y se inicializa la aplicación, mientras que el directorio app contiene los módulos y paquetes relacionados con la aplicación, como modelos, servicios, controladores, etc.

- Modelos de detección de emociones: Se implementaron dos modelos de aprendizaje automático para la detección de emociones: un modelo CNN (Convolutional Neural Network) y un modelo ViT (Vision Transformer).Los modelos fueron previamente entrenados y se cargaron en la aplicación utilizando las bibliotecas torch y transformers. Se definieron clases (CNNModel y ViTModel) para encapsular la lógica de carga y predicción de cada modelo.

- Servicios y lógica de negocio: Se creó una clase EmotionDetectionService para manejar la lógica de negocio relacionada con la detección de emociones. Esta clase se encarga de recibir las imágenes (ya sea a través de un archivo subido o una URL), preprocesarlas y pasarlas a los modelos de detección de emociones para obtener las predicciones. Se implementó el manejo de errores y excepciones para proporcionar mensajes de error significativos al usuario en caso de problemas al procesar las imágenes.

- Controladores y rutas de la API: Se definieron controladores y rutas utilizando FastAPI para manejar las solicitudes HTTP y interactuar con la interfaz de usuario. El controlador principal (emotion_detection_controller.py) define las rutas para la página principal y para procesar las solicitudes de detección de emociones. Se utilizaron dependencias de FastAPI, como UploadFile y Form, para manejar la carga de archivos y los datos de formulario enviados por el usuario.

- Interfaz de usuario: Se creó una interfaz de usuario básica utilizando HTML y Jinja2 para permitir a los usuarios interactuar con la aplicación. La interfaz de usuario incluye un formulario para subir una imagen o proporcionar una URL de imagen, y muestra los resultados de la detección de emociones después de procesar la imagen.

- Manejo de errores y logging: Se implementó un manejo de errores adecuado en toda la aplicación para capturar y manejar excepciones de manera apropiada. Se utilizó el módulo logging para registrar mensajes de información y errores durante la ejecución de la aplicación, lo que facilita el seguimiento y la depuración de problemas.

- Configuración del entorno de desarrollo: Se utilizó un contenedor de desarrollo (dev container) para proporcionar un entorno de desarrollo consistente y aislado. Se configuraron las dependencias necesarias y se instalaron en el contenedor para garantizar un entorno de desarrollo funcional.

En resumen, este proyecto utiliza FastAPI para crear una aplicación web de detección de emociones en imágenes, permite a los usuarios subir imágenes y utiliza los dos modelos de

clasificación relacionados en este trabajo (CNN y ViT) para predecir las emociones presentes en las imágenes. La aplicación sigue una estructura modular, maneja errores de manera adecuada y proporciona una interfaz de usuario básica para interactuar con la funcionalidad de detección de emociones.

Para realizar una evaluación cualitativa, con imágenes ajenas al conjunto de datos FER2013, se toman un total de 34 imágenes de rostros donde se pueden identificar de manera no estadística sus emociones. Estas imágenes son tomadas del sitio Unsplash, que ofrece fotos de libre acceso para uso comercial y personal.
Se seleccionaron 34 imágenes distribuidas cualitativamente en las 7 categorías de emociones, para posteriormente utilizar el prototipo funcional y clasificar las mismas con los modelos basados en CNN y ViT. Estas clasificaciones de los modelos son comparadas con la emoción a la que se relacionó la imagen de manera cualitativa.
Las clasificaciones son relacionadas por el prototipo de manera porcentual en la parte inferior de cada imagen, como se muestra en la figura de “Clasificación con modelos para imágenes de emociones”, y en la totalidad de las imágenes que se encuentran agrupadas por su emoción respectiva en los anexos del presente trabajo.

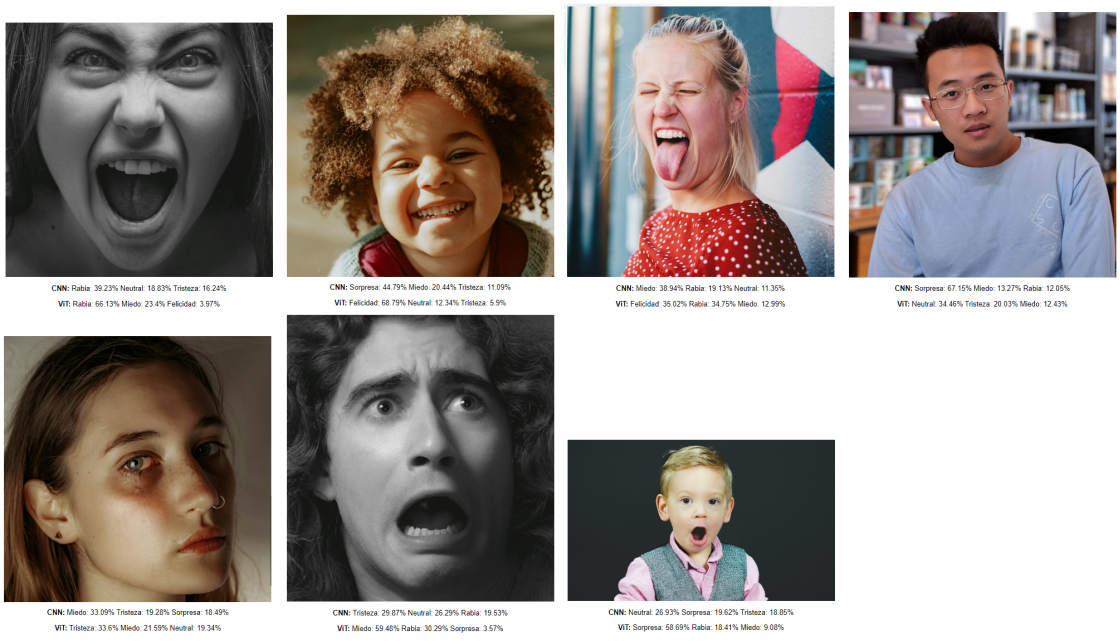

Figura 3-6: Clasificación con modelos para imágenes de emociones. Fuente: Unsplash.com

Una vez obtenidas las clasificaciones por CNN y ViT, estas se comparan con la emoción asociada a cada imagen, obteniendo el total de verdaderos positivos o aciertos para cada modelo. Definiéndose como un verdadero positivo cuando la emoción clasificada con mayor porcentaje por el modelo coincide con la emoción que se ha asociado cualitativamente a ella.

Emoción	Imágenes	Aciertos CNN	Aciertos ViT
Asco	1	0	0
Enojo	7	3	5
Felicidad	10	4	7
Miedo	1	0	1
Neutralidad	6	2	5
Sorpresa	5	4	3
Tristeza	4	0	3
Exactitud CNN			0.35
Exactitud ViT			0.70

Tabla 3-4: Informe de clasificación de la evaluación cualitativa.

Se encuentra así para ViT que su porcentaje de aciertos con aproximadamente un 70 % se acerca en este análisis cualitativo al porcentaje de precisión encontrado con las imágenes de validación del conjunto de datos, que fue del 76 %. Por otro lado, CNN cuenta con un porcentaje de aciertos de aproximadamente 35 % en este análisis, quedando significativamente lejos de la precisión encontrada con las imágenes de validación del conjunto de datos, que fue del 63 %.
En esta evaluación, en ambos modelos las emociones de “asco”, “tristeza” y “miedo” tienen un comportamiento muy diferente frente a la precisión obtenida con el conjunto de validación, lo que se puede asociar a la baja cantidad de imágenes con dichas emociones encontradas de manera cualitativa. Mientras que ambos presentan un comportamiento similar en la clasificación de las demás emociones, pese a que porcentualmente su precisión no es igual a la encontrada con el conjunto de validación.

3.7. Comparación de resultados de CNN y ViT

Los hallazgos del presente trabajo muestran tanto logros como áreas susceptibles de mejora. Unas precisiones cercanas al 63 % para CNN y al 76 % para ViT demuestran las capacidades y limitaciones inherentes de los modelos.
Con ambos modelos se observó una alta precisión en la identificación de emociones como la “felicidad”, mientras que otras, como el “miedo” y la “tristeza”, presentaron mayores dificultades de reconocimiento.
Cuando se desarrollan ambos modelos se ve una mejora significativa entrenando con Fer2013 y ViT para la clasificación de emociones. Este muestra un rendimiento superior que puede seguir mejorando por con otros conjuntos de datos, ya que tenemos un 13 % de mejores resultados con el conjunto de evaluación y un 35 % en la evaluación cualitativa con imágenes ajenas a FER2013.

4 Conclusiones y trabajo futuro

- Las conclusiones extraídas de este proyecto refuerzan la viabilidad de las CNN para el reconocimiento de emociones en imágenes, destacando tanto su potencial como las áreas susceptibles de mejora. Se concluye que mientras ciertas emociones se detectan con alta exactitud, otras requieren un enfoque más refinado que podría incluir el análisis de secuencias temporales o datos multimodales para capturar la dinámica de las expresiones faciales.

- La capacidad de los ViT para aprender representaciones ricas y detalladas a partir de los datos sugiere que el uso de conjuntos de datos más grandes y diversos podría mejorar aún más su rendimiento. La exploración de técnicas de aumentación y balanceo de datos más sofisticadas y la inclusión de datos multimodales podrían proporcionar al modelo una comprensión más holística de las emociones humanas.

- La implementación de visión transformadora (ViT) ha demostrado ser prometedora en el campo del reconocimiento de emociones faciales, sobrepasando en el presente trabajo las limitaciones de las CNN en aspectos clave como la captura de relaciones globales en la imagen. Esto subraya el potencial de los ViT para mejorar la clasificación de emociones complejas y sugiere una dirección fructífera para la investigación futura en el procesamiento avanzado de imágenes.

- De cara al futuro, se propone explorar modelos más complejos, como redes neuronales recurrentes o arquitecturas de atención, que podrían ofrecer una mejor interpretación del contexto y la temporalidad de las expresiones faciales. Además, la expansión y diversificación del conjunto de datos podrían ayudar a aumentar la robustez del modelo ante variaciones demográficas y culturales. Finalmente, la implementación de técnicas de aprendizaje profundo más avanzadas, como el aprendizaje por transferencia o las redes generativas adversarias, podría mejorar significativamente la capacidad de generalización del modelo.

- El conjunto de datos FER2013 proporciona una base sólida para entrenar y evaluar modelos de reconocimiento de emociones faciales, pero sus resultados no garantizan una generalización completa a otros conjuntos de datos de FER. Aunque FER2013 es variado y contiene una buena cantidad de imágenes etiquetadas en diferentes categorías de emociones, tiene ciertas limitaciones, como la baja resolución de las imágenes, la falta de diversidad en términos de etnias y entornos, y posibles sesgos en la recolección

de datos. Por lo tanto, se debe tener precaución al generalizar los resultados descritos entre CNN y ViT frente a otros conjuntos de datos de FER.

- Para futuras investigaciones, sería beneficioso comparar el rendimiento de diferentes variantes de ViT y explorar el efecto de distintos esquemas de preentrenamiento. La adaptación de modelos específicamente diseñados para comprender mejor las sutilezas y variaciones en las expresiones faciales podría conducir a avances significativos en el campo.

- La exploración de enfoques híbridos, que combinen las fortalezas de los ViT con las de otros modelos, como las redes neuronales recurrentes (RNN) para capturar la temporalidad y el contexto, podría ofrecer nuevas perspectivas sobre el reconocimiento efectivo de emociones. Esta integración de diferentes arquitecturas podría ser clave para abordar las emociones que se manifiestan sutilmente a lo largo del tiempo.

- Finalmente, la incorporación de feedback de aplicaciones del mundo real y la colaboración interdisciplinaria con expertos en psicología y ciencias cognitivas pueden enriquecer el desarrollo de modelos de reconocimiento de emociones, asegurando que las soluciones tecnológicas se alineen estrechamente con los entendimientos humanísticos de la emoción.

Bibliografía

[Abdulhussein and Saud, 2023] Abdulhussein, D. M. and Saud, L. J. (2023). An analysis of a hybrid algorithm for face detection and emotion recognition. *International Journal of Mathematics and Computer Science*, 19:163â€“–169.

[Ahammed et al., 2023] Ahammed, T., Ghosh, S., Rahman, A., Chandra, P., Shuvo, A. I., and Balaji, P. (2023). Meta-transfer learning for contextual emotion detection in face affirmation. In *Recent Trends in Artificial Intelligence and IoT*, pages 107–121.

[Ahmadi, 2023] Ahmadi, S. (2023). Open ai and its impact on fraud detection in financial industry. *Journal of Knowledge Learning and Science Technology*, 2:2959–6386.

[Akiba et al., 2019] Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *arXiv preprint arXiv:1907.10902*.

[Anas et al., 2020] Anas, H., Rehman, B., and Ong, W. H. (2020). Deep convolutional neural network based facial expression recognition in the wild. 1.

[Barrett, 2006] Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*.

[Barrett, 2012] Barrett, L. F. (2012). Emotions are real. *Emotion*.

[Barrett, 2016] Barrett, L. F. (2016). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*.

[Barrett, 2017a] Barrett, L. F. (2017a). Emotion is personal: An affective neuroscience approach. *Emotion Review*.

[Barrett, 2017b] Barrett, L. F. (2017b). *How Emotions are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.

[Barrett, 2019] Barrett, L. F. (2019). Variety is the spice of life: A psychological construction approach to understanding variability in emotion. *Cognition and Emotion*.

[Bhat and Jain, 2023] Bhat, A. and Jain, S. (2023). Face recognition in the age of clip & billion image datasets.

[Bird and Lotfi, 2024] Bird, J. J. and Lotfi, A. (2024). Cifake: Image classification and explainable identification of ai-generated synthetic images. *IEEE Access*, 12:15642–15650.

[Calvo and D'Mello, 2010a] Calvo, R. A. and D'Mello, S. (2010a). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*.

[Calvo and D'Mello, 2010b] Calvo, R. A. and D'Mello, S. (2010b). Automatic setection and classification of facial expressions of emotion. *Annual Review of Cybertherapy and Telemedicine*.

[Cardoso et al., 2021] Cardoso, A. C., Talame, L., Amor, M., and Monge, A. (2021). Aplicación de técnicas avanzadas de aprendizaje sutomático para identificar emociones en textos. In *XXIII Workshop de Investigadores en Ciencias de la Computación (WICC 2021, Chilecito, La Rioja)*.

[Carrera et al., 2021] Carrera, H. A., Maita, S. S., and Lascano, P. H. (2021). Modelo para detectar el uso correcto de mascarillas en tiempo real utilizando redes neuronales convolucionales. *Revista de Investigación en Tecnologías de la Información: RITI*, 9(17):111–120.

[Centeno et al., 2023] Centeno, T.-B., Ferreira, C., Inga, J.-G., Vélez, A., Huacho, R., Vidal, O.-D., Moya, S.-M., Reyes, D.-C., Goytendia, W.-E., Ascue, B.-S., and Tomazello-Filho, M. (2023). Herramientas de corte para optimizar parámetros de clasificación de especies maderables con redes neuronales convolucionales. *Revista de Biología Tropical*, 71.

[Chavan and Kulkarni, 2020] Chavan, U. B. and Kulkarni, D. (2020). Optimizing deep convolutional neural network for facial expression recognition. *European Journal Of Engineering Research And Science*, 5.

[Darwin, 1872] Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. John Murray.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[Ehsan et al., 2022] Ehsan, A., Abuhaliqa, M. A. M., Catal, C., and Mishra, D. (2022). Restful api testing methodologies: Rationale, challenges, and solution directions. *Applied Sciences*, 12(9):4369.

[Ekman, 1992] Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*.

[Ekman, 1999] Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*.

[Galety et al., 2022] Galety, M. G., Almukhtar, F. H., Maaroof, R. J., Rofoo, F., and Arun, S. (2022). Marking attendance using modern face recognition (fr): Deep learning using the opencv method. In *2022 8th International Conference on Smart Structures and Systems (ICSSS)*, pages 1–6. IEEE.

[Godishala et al., 2022] Godishala, A. K., Yassin, H., Veena, R., and Lai, D. T. C. (2022). Breast cancer tumor image classification using deep learning image data generator. In *2022 7th International Conference on Image, Vision and Computing (ICIVC)*, pages 418–423. IEEE.

[Gross, 2002] Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*.

[Gross and Muñoz, 1995] Gross, J. J. and Muñoz, R. F. (1995). Emotion regulation and mental health. *Clinical psychology: Science and practice*, 2(2):151.

[Guo et al., 2013] Guo, P., Luo, Y., and Weng, Y. (2013). Ist557 final report - human facial expression recognition.

[Harvey, 2016] Harvey, A. (2016). Seeing ourselves through technology: How we use selfies, blogs and wearable devices to see and shape ourselves.

[Huber et al., 2019] Huber, S., Wiemer, H., Schneider, D., and Ihlenfeldt, S. (2019). Dmme: Data mining methodology for engineering applications â€“ a holistic extension to the crisp-dm model. *Procedia CIRP*, 79:403â€“408.

[Japkowicz, 2013] Japkowicz, N. (2013). Assessment metrics for imbalanced learning. *Imbalanced learning: Foundations, algorithms, and applications*, pages 187–206.

[Juárez Trujillo et al., 2023] Juárez Trujillo, I. A., Zavala de Paz, J. P., Palillero Sandoval, O., and Castillo Velásquez, F. A. (2023). Calibración de cámara multiespectral utilizando redes neuronales convolucionales. *Computación y Sistemas*.

[Khan et al., 2022] Khan, A. A., Yadav, S., and Kumar, L. (2022). Face mask detection using opencv and machine learning. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 80–85. IEEE.

[Khare et al., 2024] Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., and Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014â€“2023) and research recommendationsn. *Information Fusion*, 102.

[Krumm and Davies, 2017] Krumm, J. and Davies, N. (2017). The social implications of affective computing. *IEEE Pervasive Computing*.

[Kuwahara et al., 2023] Kuwahara, M., Fujima, J., Takahashi, K., and Takahashi, L. (2023). Improving scientific image processing accessibility through development of graphical user interfaces for scikit-image. *Digital Discovery*, 2(3):775–780.

[Lasri et al., 2019] Lasri, I., Solh, A. R., and Belkacemi, M. E. (2019). Facial emotion recognition of students using convolutional neural network.

[LeDoux, 2000] LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience*.

[Lemaître et al., 2017] Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

[Li et al., 2023] Li, C., Li, X., Chen, M., and Sun, X. (2023). Deep learning and image recognition. In *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*, pages 557–562. IEEE.

[Martínez-Plumed et al., 2019] Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., and Flach, P. (2019). Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*.

[McStay, 2018] McStay, A. (2018). Emotional ai: The rise of empathic media.

[Meena et al., 2023] Meena, G., Mohbey, K. K., Indian, A., Khan, M. Z., and Kumar, S. (2023). Identifying emotions from facial expressions using a deep convolutional neural network-based approach. *Multimedia Tools and Applications*, pages 1–22.

[Organización Mundial de la Salud, 2023] Organización Mundial de la Salud (2023). Preguntas frecuentes. Último acceso: 2023.

[Pantic and Rothkrantz, 2008] Pantic, M. and Rothkrantz, L. J. (2008). Machine analysis of facial expressions. *Face recognition*.

[Pennington et al., 2017] Pennington, J., Schoenholz, S., and Ganguli, S. (2017). Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *Advances in neural information processing systems*, 30.

[Plazas López et al., 2022] Plazas López, J. A., Gutiérrez Leguizamón, J. J., Suárez Barón, M. J., and González Sanabria, J. S. (2022). Reconocimiento de lengua de señas colombiana mediante redes neuronales convolucionales y captura de movimiento. *Tecnura*, 26:70–86.

[Pomazan et al., 2023] Pomazan, V., Tvoroshenko, I., and Gorokhovatskyi, V. (2023). Face recognition in the age of clip & billion image datasets. *International Journal of Academic Information Systems Research*, 7:25–36.

[Rasheed et al., 2022] Rasheed, L., Khadam, U., Majeed, S., Ramzan, S., Bashir, M. S., and Iqbal, M. M. (2022). Face recognition emotions detection using haar cascade classier and convolutional neural network. *Research Square*, 1.

[Russell, 2003] Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*.

[Sánchez, 2018] Sánchez, S. (2018). Recuperación de imágenes por contenido usando descriptores generados por redes neuronales convolucionales. *Revista Cubana de Ciencias Informóticas*, pages 78– 90.

[Sathyamoorthy et al., 2023] Sathyamoorthy, B., Snehalatha, U., and Rajalakshmi, T. (2023). Facial emotion detection of thermal and digital images based on machine learning techniques. *Biomedical Engineering: Applications, Basis and Communications*, 35(01):2250052.

[Savva and Stylianou, 2023] Savva, A. and Stylianou, V. (2023). Real-time emotional analysis. In *2023 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, pages 1–5. IEEE.

[Schröer et al., 2021] Schröer, C., Kruse, F., and Marx Gómez, J. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, 181:526–â€“534.

[Shafique and Qaiser, 2014] Shafique, U. and Qaiser, H. (2014). A comparative study of data mining process models (kdd, crisp-dm and semma). *International Journal of Innovation and Scientific Research*, 12:217–222.

[Smith, 2017] Smith, L. N. (2017). Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE.

[Smys et al., 2020] Smys, S., Chen, J. I. Z., and Shakya, S. (2020). Survey on neural network architectures with deep learning. *Journal of Soft Computing Paradigm (JSCP)*, 2(03):186–194.

[van der Vlist et al., 2022] van der Vlist, F. N., Helmond, A., Burkhardt, M., and Seitz, T. (2022). Api governance: the case of facebookâ€™s evolution. *Social Media+ Society*, 8(2):20563051221086228.

[Wachter et al., 2021] Wachter, J. F. C., Villas, C. F. H., Villa, F. H., and Arango, D. A. G. (2021). Análisis del aporte del aprendizaje de máquinas a la seguridad de la información. *InGente Americana*, 1(1):9–20.

[Wäldchen and Mäder, 2018] Wäldchen, J. and Mäder, P. (2018). Machine learning for image based species identification. *Methods in Ecology and Evolution*, 9(11):2216–2225.

[Wang et al., 2021] Wang, P., Fan, E., and Wang, P. (2021). Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters*, 141:61–67.

[Wang et al., 2019] Wang, W., Sun, Q., Chen, T., Cao, C., Zheng, Z., Xu, G., Qiu, H., and Fu, Y. (2019). A fine-grained facial expression database for end-to-end multi-pose facial expression recognition. 1.

[Wessels et al., 2022] Wessels, H., Böhm, C., Aldakheel, F., Hüpgen, M., Haist, M., Lohaus, L., and Wriggers, P. (2022). Computational homogenization using convolutional neural networks. In *Current Trends and Open Problems in Computational Mechanics*, pages 569–579. Springer.

[Wiemer et al., 2019] Wiemer, H., Drowatzky, L., and Ihlenfeldt, S. (2019). Data mining methodology for engineering applications (dmme)â€"a holistic extension to the crisp-dm model. *applied sciences*, 9.

[Wirth and Hipp, 2000] Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining.

[Wu et al., 2020] Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., and Vajda, P. (2020). Visual transformers: Token-based image representation and processing for computer vision.

[Yang et al., 2022] Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., and Shen, F. (2022). Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*.

[Zheng et al., 2023] Zheng, Y., Ding, J., Liu, F., and Wang, D. (2023). Adaptive neural decision tree for eeg based emotion recognition. *Information Sciences*, 643:119160.

5 Anexos

En este apartado se presentan las 34 imágenes de rostros que fueron introducidas en el prototipo funcional para realizar la evaluación cualitativa de los modelos de clasificación. Estas imágenes se encuentran agrupadas cualitativamente en “Asco”, “Enojo”, “Felicidad”, “Miedo”, “Neutralidad”, “Sorpresa” y “Tristeza”.

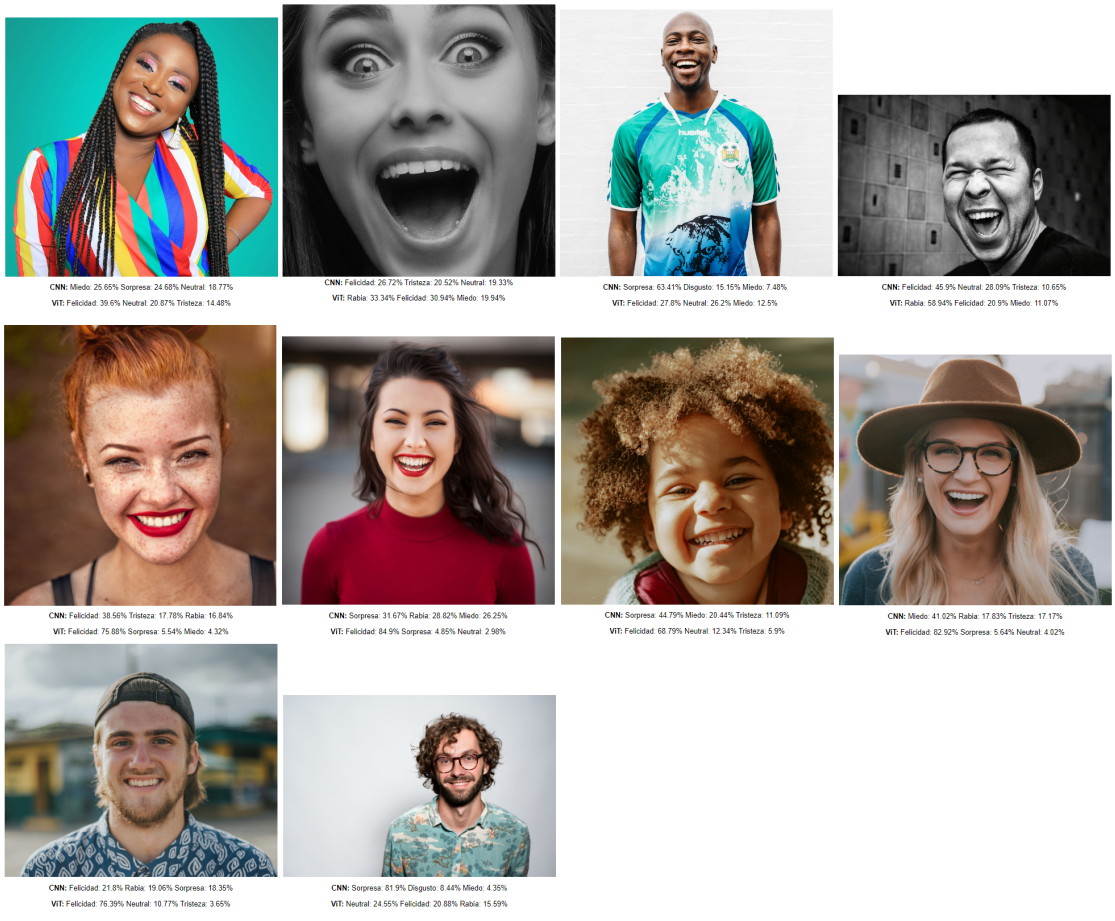

Figura 5-1: Clasificación con modelos para imágenes de felicidad. Fuente: Unsplash.com

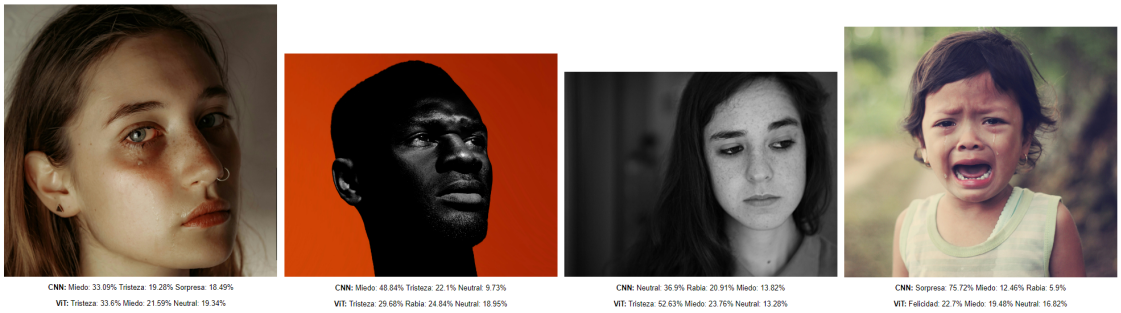

Figura 5-2: Clasificación con modelos para imágenes de tristeza. Fuente: Unsplash.com

Figura 5-3: Clasificación con modelos para imágenes de sorpresa. Fuente: Unsplash.com

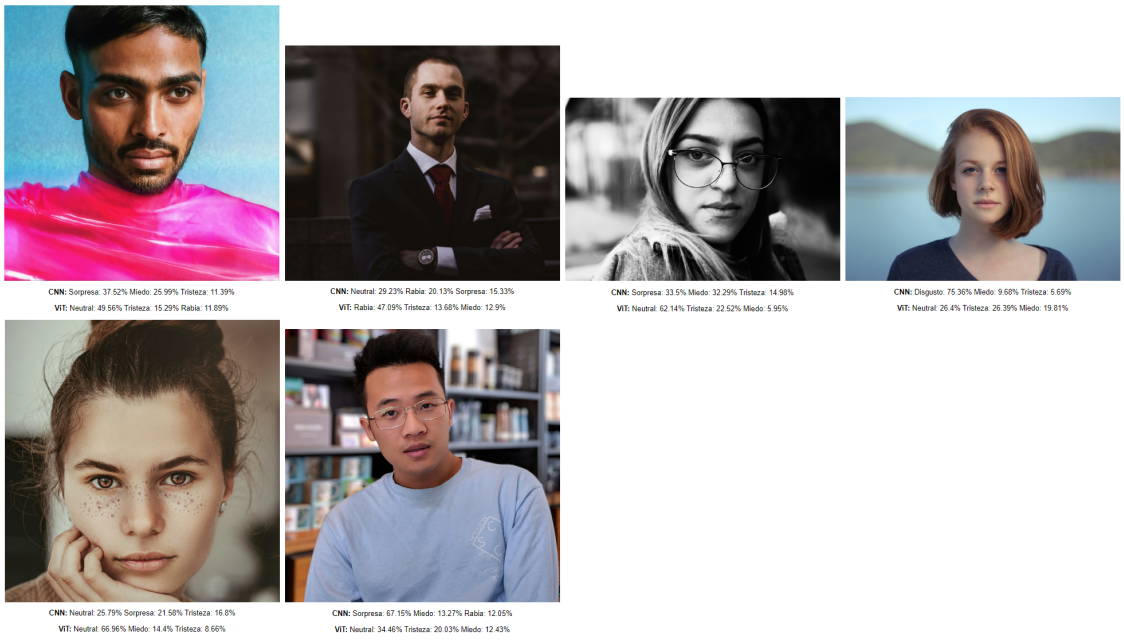

Figura 5-4: Clasificación con modelos para imágenes de neutralidad. Fuente: Unsplash.com

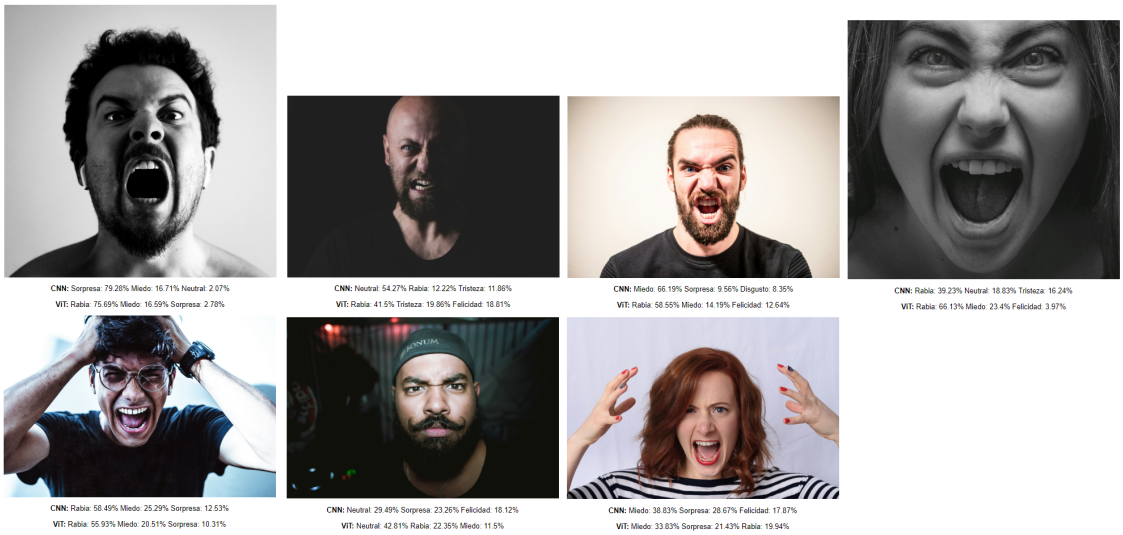

Figura 5-5: Clasificación con modelos para imágenes de enojo. Fuente: Unsplash.com

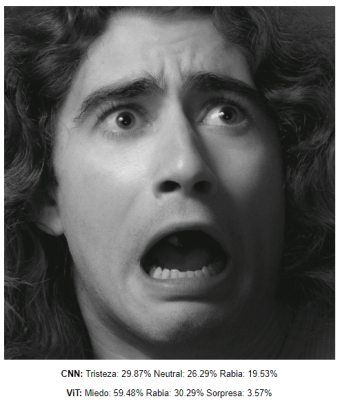

Figura 5-6: Clasificación con modelos para imagen de miedo. Fuente: Unsplash.com

Figura 5-7: Clasificación con modelos para imagen de asco. Fuente: Unsplash.com