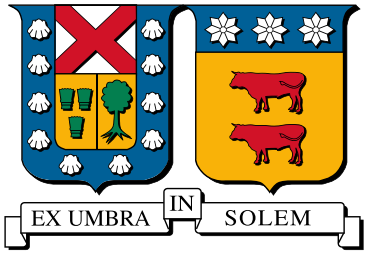
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE ELECTRÓNICA

VALPARAÍSO - CHILE



"AUTOMATIZACIÓN DE DETECCIÓN DE PATRÓN DEMOGRÁFICO EN SERVICIOS DE ATENCIÓN AL CLIENTE"

NICOLÁS GIANINI AGUILERA JIMÉNEZ

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELECTRÓNICO.

PROFESOR GUÍA: DR. WERNER CREIXELL FUENTES

PROFESOR CORREFERENTE: DR. MARCOS ZÚÑIGA BARRAZA

Marzo 2026

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: _**Automatización de detección de patrón demográfico en servicios de atención al cliente**_

Nombre del candidato(a): _**Nicolás Gianini Aguilera Jimenez**_

Carrera / Grado: _**Ingeniería Civil Electrónica**_

Campus: _Casa Central_ **Departamento:** _Electrónica_

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Werner Creixell Fuentes, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

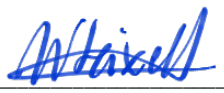
_____

_____

_____

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 20 de Abril 2026 **Firma:** 

Estudiante o Candidato(a):

Fecha: 22 de Abril 2026 **Firma:** 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Agradecimientos

En primer lugar, quisiera agradecer a mi profesor guía, Werner Creixel, por su orientación experta y dedicada. Su conocimiento y experiencia fueron vitales. Agradezco sinceramente su tiempo y compromiso para apoyarme en la elaboración de esta memoria.

A mis padres, quienes, junto a mis hermanos, supieron apoyarme en todo lo que pudieron desde el inicio hasta el fin de este camino, además de nunca perder su fe y orgullo hacia mí.

A mis compañeros, no puedo imaginar haber llegado hasta este punto sin su amistad y compañía a lo largo de la carrera. Gracias por los gratos momentos, las conversaciones, tanto las banales como las profundas y serias.

A mis amistades lejanas, con las que estoy infinitamente agradecido de poder mantener su amistad a día de hoy, a pesar de la distancia y la diferencia de nuestros caminos. Confío en que, a pesar de la distancia y lo dispersos que estamos, podremos continuar teniendo la gran amistad y cariño que hemos tenido.

A cada uno de ustedes, gracias de corazón. Su apoyo, amistad y amor han sido esenciales en mi camino hacia este punto. Sin ustedes, este viaje no habría sido tan significativo ni memorable.

Dedico este trabajo y titulo a mi abuela y mi padrino, quienes gracias a su dedicación y amor, fueron parte de la persona que soy hoy en día.

DEPARTAMENTO DE ELECTRONICA
UNIVERSIDAD TECNICA FEDERICO SANTA MARIA

DEPARTAMENTO DE ELECTRONICA

Automatización de detección de patrón demográfico en servicios de atención al cliente

Nicolás Gianini Aguilera Jiménez

Memoria para optar al título de Ingeniero Civil Electrónico, mención Computadores, submención Gestión.

Universidad Técnica Federico Santa María

Profesor Guía: Dr. Werner Creixell Fuentes

Profesor Correferente: Dr. Marcos Zúñiga Barraza

Marzo 2026

Resumen

En el ámbito de atención al cliente, mejorar la calidad del servicio requiere obtener métricas útiles para medir y controlar la calidad en la atención al usuario; por ello, este trabajo propone un enfoque basado en inteligencia artificial para analizar características faciales y tiempo de atención que permitan obtener estas medidas. Se desarrolla un software que integra técnicas de visión por computadora para detección y seguimiento, junto con clasificación de edad, género y etnia mediante modelos redes neuronales convolucionales multi-tarea, ajustados con transfer learning, para obtener datos faciales que luego serán procesados estadísticamente con el fin de obtener datos de medición relacionados con la interacción en el servicio.

Keywords: Atención al cliente, Calidad de servicio, Visión por computadora, Inteligencia artificial, Aprendizaje por transferencia, Redes neuronales convolucionales, Análisis estadístico.

Automating demographic pattern detection in customer service.

Nicolás Gianini Aguilera Jiménez

Thesis for the fulfillment of the requirements for the degree of Electronic Civil Engineer, mayor in Computers, minor in Management

Universidad Técnica Federico Santa María

Advisor: Dr. Werner Creixell Fuentes

Co-Advisor: Dr. Marcos Zúñiga Barraza

March 2026

Abstract

In the field of customer service, improving service quality requires obtaining useful metrics to measure and control the quality of user care. Therefore, this work proposes an artificial intelligence-based approach to analyze facial features and service time to obtain these measurements. Software is developed that integrates computer vision techniques for detection and tracking, along with age, gender, and ethnicity classification using multi-task convolutional neural network models, adjusted with transfer learning, to obtain facial data that will then be statistically processed to obtain measurement data related to service interaction.

Keywords: Customer service, Service quality, Computer vision, Artificial intelligence, Transfer learning, Convolutional neural networks, Statistical analysis.

Glosario

Banca Sector financiero que gestiona depósitos, préstamos y otros servicios económicos a individuos y empresas.. , 3, 5

BlazeFace Modelo de detección de rostros rápido y eficiente, desarrollado por Google.. , 3, 4, 6, 7, 17, 35, 38, 40

CNN Redes neuronales convolucionales o por sus siglas en ingles **Convolutional Neural Network**.. , 3, 4, 8, 13, 14, 17, 22, 27, 36, 38, 39

CV Visión por computadora o por sus siglas en ingles **Computer Vision**.. , 2

Deep Learning Rama del aprendizaje automático basada en redes neuronales profundas para el procesamiento de datos complejos.. , 2, 3, 13

Deep SORT Extensión de SORT que incorpora redes neuronales profundas para mejorar el seguimiento multiobjeto en videos.. , 3, 5, 7, 35, 38, 40

DenseNet Arquitectura de red neuronal convolucional que conecta cada capa con todas las anteriores para mejorar el flujo de información y la eficiencia del entrenamiento.. , 8, 10, 11, 17

EfficientNetB0 Modelo de red neuronal eficiente que optimiza el tamaño y la precisión.. , 3, 8, 11, 12, 17, 21, 26, 28, 29, 36

IA Inteligencia Artificial.. , 2

ImageNet Base de datos masiva de imágenes etiquetadas que se utiliza comúnmente para entrenar y evaluar modelos de visión por computadora. Contiene millones de imágenes organizadas jerárquicamente en miles de categorías.. , 8, 13

InceptionV3 Modelo de red neuronal convolucional optimizado para clasificación de imágenes, basado en módulos Inception que mejoran eficiencia y precisión.. , 8–12, 17, 19, 20, 24

Machine Learning Rama de la inteligencia artificial que permite a los sistemas aprender automáticamente a partir de datos, sin ser explícitamente programados. Utiliza algoritmos que identifican patrones y realizan predicciones o decisiones basadas en los datos de entrada.. , 3

OpenCV Biblioteca de código abierto para visión por computadora que permite el procesamiento de imágenes y video en tiempo real, incluyendo tareas de preprocesamiento, detección y seguimiento de objetos y rostros.. , 35, 36

Retail Sector comercial dedicado a la venta de productos o servicios directamente al consumidor final.. , 3, 5

Transfer Learning Técnica en aprendizaje automático donde se adapta un modelo preentrenado en una tarea especifica.. , 4, 8, 9, 11, 13, 14, 17–19, 21, 23–26

VGG16 Modelo de red neuronal profunda para clasificación de imágenes, conocido por su arquitectura simple y profunda, con 16 capas de procesamiento.. , 3, 8–12, 17, 23

Índice de contenidos

List of Figures

1 Introducción

En el sector de atención al cliente de negocios orientados a servicios masivos dirigidos a personas naturales, la calidad del servicio y la experiencia del usuario constituyen factores determinantes para la satisfacción, la fidelización y la reputación de la empresa. No obstante, este proceso suele verse afectado por diversas variables difíciles de controlar, tales como el estado emocional del personal, la variabilidad inherente a las interacciones humanas y, particularmente relevante para este trabajo, la presencia de sesgos cognitivos [1], entendidos como las tendencias inconscientes que distorsionan el juicio humano. Estas fluctuaciones pueden traducirse en inconsistencias en el trato, respuestas poco adaptadas a las necesidades del cliente y una disminución en la percepción de la calidad del servicio. [2, 3]

Una de las áreas críticas donde estas diferencias se hacen evidentes es en el trato hacia personas de distintos perfiles demográficos, como la edad, el género o la etnia. La percepción y el comportamiento del personal pueden estar influenciados, consciente o inconscientemente, por estos factores, lo que puede generar un trato desigual o sesgado hacia ciertos grupos [4]. En Chile, la presencia de sesgos cognitivos ha sido documentada en diversos ámbitos institucionales [5–7]. Aunque estos estudios no abordan directamente la atención al cliente, evidencian la relevancia del fenómeno a nivel nacional y refuerzan la necesidad de desarrollar herramientas que permitan analizar su posible manifestación en contextos de servicios. Este proyecto surge como una propuesta para automatizar la detección y clasificación del perfil demográfico de clientes mediante herramientas de inteligencia artificial, con el objetivo de reducir la intervención manual en el levantamiento de información y facilitar el acceso a métricas evaluativas objetivas. Para ello, se desarrolla un sistema capaz de detectar y clasificar rostros en tiempo real, extrayendo atributos demográficos como edad, género y raza, los cuales son posteriormente organizados y analizados de forma descriptiva.

De este modo, se busca entregar a las empresas una herramienta que permita monitorear y caracterizar el perfil demográfico de los clientes que interactúan con sus servicios, proporcionando una base cuantificable para la evaluación de patrones de atención y habilitando, a futuro, el análisis de posibles sesgos en la interacción cliente–empleado. [8].

1.1 Inteligencia artificial

La *IA* es una disciplina que busca desarrollar sistemas capaces de imitar capacidades cognitivas humanas como el aprendizaje, la toma de decisiones, la resolución de problemas y el reconocimiento de patrones. Su uso se ha expandido rápidamente en diversos campos, incluyendo la medicina, la industria, la seguridad y los servicios. [9]

La *IA* ha demostrado ser una herramienta poderosa en diversas aplicaciones, desde la automatización industrial hasta la medicina y la seguridad. En el contexto del servicio al cliente, la *IA* permite analizar grandes volúmenes de datos en tiempo real y tomar decisiones informadas con base en patrones identificados en los usuarios. [10]

En el contexto de atención al cliente, la *IA* se ha convertido en una herramienta clave para mejorar la eficiencia, personalización y calidad del servicio. [11, 12] A través del análisis automatizado de datos, es posible identificar patrones de comportamiento de los usuarios, adaptar respuestas en tiempo real y detectar posibles anomalías o sesgos en la interacción humana.

En este proyecto, la inteligencia artificial se aplica como base para la detección y análisis de características faciales mediante modelos de aprendizaje profundo o *Deep Learning* [13]. Al utilizar técnicas de detección y clasificación automática de rostros según atributos como edad, género y etnia, se busca identificar posibles patrones diferenciales en la atención brindada a distintos perfiles de clientes.

1.1.1 Visión por computador

La visión por computadora o *CV* es una rama de la inteligencia artificial cuyo objetivo es dotar a las máquinas de la capacidad de interpretar y analizar imágenes y videos, emulando ciertos aspectos de la percepción visual humana. Esta disciplina permite extraer información significativa a partir de datos visuales, reconociendo objetos, rostros, expresiones, patrones de movimiento y estructuras espaciales, entre otros. [14]

En el contexto de este trabajo, la visión por computadora cumple un rol fundamental al ser el punto de partida para la obtención de datos faciales. A través de técnicas de detección y segmentación, es posible localizar rostros en imágenes o videos en tiempo real [15] y, posteriormente, aplicar algoritmos de clasificación para estimar características demográficas como la edad, el género y la etnia de los individuos. [16]

Estas capacidades resultan especialmente útiles en entornos donde la interacción con el cliente es un factor clave, como en el *Retail* o la *Banca*, ya que permiten analizar información visual en escenarios dinámicos, caracterizados por movimiento constante y condiciones irregulares de iluminación. [17, 18]

En este proyecto, la visión por computadora se implementa mediante el uso de modelos como *BlazeFace* [19] para la detección facial, en conjunto con algoritmos de seguimiento como *Deep SORT* [20], integrando así un sistema automatizado de análisis visual aplicado al servicio al cliente.

1.1.2 DeepLearnig

En este trabajo, se utilizan redes neuronales convolucionales (*CNN*) como *VGG16* [21] y *EfficientNetB0* [22] para la clasificación de características faciales. Estas redes permiten analizar imágenes en múltiples niveles de abstracción, identificando detalles clave como la edad, el género y la etnia de los clientes con alta precisión. [23]

El *Deep Learning* (aprendizaje profundo) es una subárea del aprendizaje automático (*Machine Learning*) que utiliza redes neuronales artificiales con múltiples capas ocultas para procesar información de manera jerárquica y altamente abstracta. Este enfoque ha demostrado una eficacia sobresaliente en tareas complejas como el reconocimiento de imágenes, procesamiento de lenguaje natural y análisis de video.

En este proyecto, se utiliza *Deep Learning* como herramienta principal para la clasificación automática de rostros, empleando redes neuronales convolucionales (*CNN*), una arquitectura especialmente diseñada para el procesamiento de datos visuales. Estas redes son capaces de aprender representaciones espaciales jerárquicas de las imágenes, extrayendo características faciales relevantes para determinar la edad, el género y la etnia de los individuos detectados.

El uso de *Deep Learning* permite lograr altos niveles de precisión en entornos donde las características faciales pueden variar significativamente debido a la iluminación, la posición del rostro o la calidad de la imagen, lo cual es especialmente relevante para aplicaciones en escenarios reales de atención al cliente.

1.2 Análisis estadístico

Una vez detectados y clasificados los rostros mediante los modelos de visión por computadora e inteligencia artificial, el siguiente paso consiste en analizar estadísticamente los datos recopilados con el fin de medir e identificar diferencias sistemáticas. Para ello, se utilizan herramientas de análisis estadístico clásico, principalmente:

- Análisis descriptivo, para explorar la distribución de frecuencias, promedios y tiempos de atención asociados a cada grupo demográfico.
- Análisis de comparación, para detectar diferencias significativas en el tiempo atendido según variables como etnia, edad o género.
- Análisis de correlación y regresión, para evaluar relaciones entre los atributos faciales clasificados y parámetros como el tiempo de permanencia.

Este enfoque busca generar indicadores sobre el comportamiento del sistema de atención, como también métricas relacionadas con el patrón demográfico de atención del servicio. Así, la estadística actúa como puente entre el análisis automatizado y la interpretación de los resultados a posteriori, aportando evidencia para la medición del sistema y facilitando la futura detección de sesgos o patrones discriminatorios que podrían afectar la equidad del servicio.

1.3 Objetivos

Objetivo General: Desarrollar un modelo de clasificación facial basado en redes neuronales convolucionales, orientado en mejorar la atención al cliente mediante el análisis automatizado de características faciales y el tiempo de atención.

Objetivos Específicos:

- Implementar el modelo *BlazeFace* para la detección de rostros en tiempo real.
- Entrenar y evaluar modelos *CNN* mediante *Transfer Learning* para la clasificación de rostros en 3 atributos: edad (rango de entre 0 y 80 años), género (Femenino - Masculino) y raza o etnia (Asiático, Blanco, Hispano, Negro, Indio y Medio Oriente).

- Aplicar tecnologías de seguimiento como *Deep SORT* para el rastreo de rostros en tiempo real.
- Aplicar análisis estadístico a los datos generados para obtención de métricas relacionadas con el patrón demográfico de los clientes.

1.4 Alcances y proyecciones

El presente trabajo se orienta a su aplicación en sectores como el *Retail* y la *Banca*, donde la calidad del servicio y la atención personalizada tienen un impacto directo en la experiencia del cliente.

El sistema desarrollado se limita actualmente al análisis de atributos faciales estáticos, tales como edad, género y categorías étnico-raciales inferidas, y su funcionamiento depende en gran medida de la calidad de las imágenes capturadas en tiempo real. Asimismo, el análisis estadístico se emplea como una herramienta de apoyo basada en estadística descriptiva básica, derivada de las mediciones obtenidas durante la implementación del sistema, con el objetivo de facilitar la identificación y descripción de posibles diferencias observadas en la atención.

A futuro, este sistema podría complementarse mediante la incorporación de modelos LLM orientados a procesos de auditoría, permitiendo la interpretación automatizada de los resultados y la detección de patrones que puedan afectar la calidad del servicio. Estas capacidades podrían integrarse con sistemas CRM, proporcionando retroalimentación en tiempo real al personal de atención. De forma adicional, se contempla la inclusión de componentes de clasificación facial dinámica, tales como el análisis de emociones del usuario durante la interacción, con el fin de enriquecer la evaluación de la experiencia de atención.

2 Estado del arte

El presente capítulo expone el estado del arte de las tecnologías y enfoques utilizados en el desarrollo del sistema propuesto. Se revisan las principales herramientas de visión por computador y aprendizaje profundo empleadas actualmente para la detección, seguimiento y clasificación facial, con énfasis en aquellas que permiten un procesamiento eficiente en tiempo real.

En particular, se analizan modelos de detección facial rápida, algoritmos de seguimiento multiobjeto y arquitecturas de redes neuronales convolucionales utilizadas para la estimación de atributos demográficos como edad, género y etnia. Esta revisión tiene como objetivo contextualizar las decisiones técnicas adoptadas en el proyecto, justificando la selección de cada componente del sistema en función de criterios de precisión, eficiencia computacional y aplicabilidad en escenarios reales de atención al cliente.

Como referencia principal, se consideran trabajos previos enfocados en la estimación de atributos faciales mediante redes neuronales convolucionales y técnicas de aprendizaje por transferencia [24, 25], los cuales sirven como base comparativa para el diseño e implementación del sistema desarrollado.

2.1 BlazeFace

BlazeFace es un modelo de detección de rostros desarrollado por Google, diseñado específicamente para aplicaciones en tiempo real y dispositivos con recursos computacionales limitados. Su arquitectura liviana y altamente optimizada permite realizar detecciones faciales con latencias del orden de los milisegundos, lo que lo convierte en una alternativa adecuada para sistemas que requieren procesamiento continuo de video.

El modelo se basa en un enfoque de detección de una sola etapa (single-shot detector), en el cual las predicciones de ubicación facial y sus respectivas puntuaciones de confianza se generan directamente a partir de la imagen de entrada, sin necesidad de etapas intermedias de propuesta de regiones. Para lograr esta eficiencia, BlazeFace emplea convoluciones en profundidad (depthwise separable convolutions) y un diseño de red compacto, reduciendo significativamente el número de parámetros y operaciones requeridas.

En el contexto de este proyecto, BlazeFace es utilizado como la primera etapa del pipeline de procesamiento visual, encargada de localizar los rostros presentes en cada cuadro de video.

Gracias a su eficiencia y precisión, *BlazeFace* es ideal para aplicaciones donde se requiere procesamiento en tiempo real, como análisis en ambientes dinámicos, especialmente en escenarios con restricciones computacionales propias de dispositivos móviles [26].

2.2 Deep SORT

Deep SORT (Simple Online and Realtime Tracking with a Deep Association Metric) es un algoritmo de seguimiento multiobjeto diseñado para mantener la identidad de entidades a lo largo del tiempo, incluso en escenarios dinámicos y con oclusiones parciales. Este método extiende el algoritmo SORT incorporando descriptores de apariencia extraídos mediante redes neuronales profundas, lo que permite una asociación más robusta entre detecciones consecutivas y una reducción significativa de errores de reasignación de identidad [20].

El funcionamiento de *Deep SORT* se apoya en dos componentes fundamentales. En primer lugar, la predicción de movimiento se realiza mediante un filtro de Kalman, el cual estima la posición futura de cada objeto a partir de su estado previo. En segundo lugar, la asociación de detecciones combina información geométrica, basada en la intersección sobre unión (IoU), con una métrica de apariencia calculada como la distancia coseno entre embeddings visuales. Esta combinación permite resolver ambigüedades que no pueden ser abordadas únicamente con información espacial, como cruces entre individuos o desapariciones temporales del campo visual.

La incorporación de características de apariencia resulta clave para mantener identificadores persistentes (track IDs) a lo largo del tiempo, incluso cuando los objetos experimentan oclusiones parciales o variaciones en su posición relativa. Gracias a este enfoque, el algoritmo logra un seguimiento más estable y consistente en escenas con múltiples individuos, donde los objetos pueden aparecer y desaparecer de forma continua.

En el contexto de este proyecto, *Deep SORT* se integra como la etapa encargada de realizar el seguimiento temporal de los rostros detectados por *BlazeFace*. Cada rostro identificado es asociado a un identificador único que se mantiene a lo largo de múltiples cuadros, permitiendo registrar información longitudinal como la duración de permanencia, la frecuencia de aparición y la consolidación de atributos demográficos estimados por el modelo de clasificación facial en tiempo real [26].

Esta capacidad de seguimiento continuo resulta fundamental para el análisis estadístico posterior, ya que permite agrupar múltiples detecciones pertenecientes a un mismo individuo y filtrar aquellas cuya presencia en la escena es demasiado breve para ser considerada relevante.

2.3 Modelos de Aprendizaje Profundo

Las redes neuronales convolucionales (*CNN*) constituyen actualmente el enfoque predominante para el análisis automático de imágenes y videos, especialmente en tareas de clasificación y reconocimiento facial. Su capacidad para aprender representaciones jerárquicas a partir de datos visuales las ha convertido en el estándar de facto para la estimación de atributos demográficos como edad, género y etnia, tal como se reporta en diversos trabajos recientes del estado del arte.

En el contexto de la clasificación facial, las *CNN* permiten capturar patrones espaciales complejos asociados a rasgos faciales, variaciones de textura, iluminación y geometría del rostro, superando ampliamente a los métodos tradicionales basados en características manuales. Estas capacidades resultan particularmente relevantes en escenarios reales de atención al cliente, donde las condiciones de captura no están controladas y existe una alta variabilidad entre individuos.

Con el objetivo de evaluar distintas estrategias arquitectónicas y su impacto en precisión, eficiencia computacional y aplicabilidad en tiempo real, este trabajo considera cuatro modelos representativos y ampliamente utilizados: *VGG16*, *InceptionV3*, *DenseNet* y *EfficientNetB0*. Estos modelos presentan diferentes filosofías de diseño, que abarcan desde arquitecturas profundas y secuenciales hasta enfoques altamente optimizados en términos de conectividad y escalamiento.

Todos los modelos seleccionados han sido preentrenados sobre el conjunto de datos *ImageNet*, lo que permite aprovechar representaciones visuales genéricas mediante técnicas de *Transfer Learning*. Posteriormente, estas arquitecturas son adaptadas a la tarea específica de clasificación facial mediante la modificación de su arquitectura [27], reduciendo los requerimientos de datos y tiempo de entrenamiento. Este enfoque ha demostrado ser efectivo en estudios comparativos recientes, como el trabajo de referencia de de Kothari et al. (2022) [24], se analizan modelos CNN tradicionales y ajustados por transferencia para la predicción simultánea

de edad, género y etnia.

En las subsecciones siguientes se describen las principales características de cada arquitectura considerada, junto con sus ventajas y limitaciones desde el punto de vista del presente proyecto.

2.3.1 VGG16

VGG16 es una arquitectura de red neuronal convolucional propuesta por el Visual Geometry Group (VGG) de la Universidad de Oxford. Fue presentada en 2014 como parte del concurso ImageNet Large Scale Visual Recognition Challenge (ILSVRC), donde destacó por su simplicidad y alto rendimiento. [21]

La arquitectura de *VGG16* está compuesta por 13 capas convolucionales y 3 capas completamente conectadas, sumando un total de 16 capas de aproximadamente 15 mil parametros entrenables. Utiliza filtros pequeños de 3×3, activación ReLU y capas de pooling para reducir progresivamente la dimensión espacial, permitiendo la extracción de características de manera jerárquica.

Aunque es una red profunda y precisa, *VGG16* presenta una desventaja importante: su alto costo computacional, con más de 138 millones de parámetros. A pesar de esto, su estructura simple y modular facilita su adaptación y entrenamiento en tareas específicas mediante *Transfer Learning*.

En este proyecto, *VGG16* se utiliza como una de las arquitecturas de referencia para comparar el rendimiento de modelos en la tarea de clasificación facial. Se modificaron sus capas superiores para añadir salidas específicas para las variables de edad, género y etnia, y se ajustaron los últimos bloques convolucionales para permitir un aprendizaje más fino.

2.3.2 InceptionV3

InceptionV3 es una arquitectura de red neuronal convolucional propuesta por Google como parte de su investigación sobre redes profundas eficientes. Fue introducida en 2015 como una evolución del modelo GoogLeNet (InceptionV1), y es ampliamente reconocida por su capacidad de lograr alta precisión con un uso más eficiente de recursos computacionales que otras redes profundas tradicionales. [28]

A diferencia de arquitecturas secuenciales como *VGG16*, *InceptionV3* está basada en módulos Inception, los cuales permiten realizar múltiples operaciones convolucionales en paralelo (con tamaños de filtros 1×1, 3×3, 5×5, etc.), además de operaciones de reducción dimensional y pooling. Luego, estas salidas se concatenan, permitiendo al modelo capturar patrones a distintas escalas sin necesidad de aumentar drásticamente la profundidad ni el número de parámetros. En total, el modelo cuenta con aproximadamente 22 millones de parámetros entrenables.

InceptionV3 también incorpora optimizaciones como:

- Factorización de convoluciones (por ejemplo, separar una convolución 5×5 en dos 3×3 o usar 1×n y n×1),
- Uso de batch normalization para mejorar la estabilidad del entrenamiento,
- Y auxiliary classifiers, que ayudan a combatir el desvanecimiento del gradiente en redes muy profundas.

InceptionV3 destaca por su alta eficiencia computacional en relación con su profundidad, lo que le permite alcanzar un rendimiento competitivo sin incurrir en un crecimiento excesivo en el número de parámetros. Su capacidad de capturar características a múltiples escalas dentro de un mismo módulo lo hace especialmente adecuado para tareas donde los patrones visuales varían significativamente entre clases. Como desventaja, su arquitectura es más compleja de entender e implementar que modelos secuenciales como *VGG16*, y requiere imágenes de mayor tamaño (299×299), lo que puede incrementar los costos computacionales de preprocesamiento.

2.3.3 DenseNet

DenseNet forma parte de la familia de redes Dense Convolutional Networks (DenseNet), introducida por Gao Huang et al. en 2017. Su diseño propone una arquitectura basada en la conectividad densa entre capas: cada capa recibe como entrada la salida de todas las capas anteriores, y su propia salida es transmitida a todas las capas siguientes. [29]
Esta estrategia de conexión directa mejora el flujo de información y de gradientes durante el

entrenamiento, lo cual mitiga el problema del desvanecimiento del gradiente y facilita el aprendizaje de características más relevantes con menos parámetros. *DenseNet* también promueve la reutilización de características, lo que lo hace más eficiente que otras redes de profundidad comparable.

La arquitectura de *DenseNet* se organiza en bloques densos seguidos por transición layers que reducen la dimensionalidad mediante pooling y convolución 1×1. En el caso de DenseNet121, se utilizan 121 capas con pesos entrenables distribuidas en cuatro bloques principales, teniendo un total de 7 mil parámetros.

En comparación con otras arquitecturas como *VGG16* o *InceptionV3*, DenseNet121 ofrece un buen equilibrio entre eficiencia computacional y precisión, siendo más ligero en número de parámetros que *VGG16*, pero más profundo y con mejor propagación de información. Su diseño lo hace especialmente útil en tareas de clasificación donde los patrones visuales son sutiles o de alta variabilidad, como es el caso de la clasificación facial.

Como posible desventaja, la estructura altamente interconectada de *DenseNet* puede incrementar el consumo de memoria durante el entrenamiento, ya que todas las salidas intermedias deben conservarse para ser reutilizadas. Además, su implementación requiere una gestión cuidadosa del flujo de datos, especialmente cuando se adapta a tareas personalizadas mediante *Transfer Learning*.

2.3.4 EfficientNet

EfficientNetB0 es la versión base de la familia de arquitecturas EfficientNet, propuesta por Google en 2019. Su principal innovación es la introducción de una técnica llamada Compound Scaling, que permite escalar de forma eficiente la profundidad, el ancho y la resolución de entrada de la red de manera balanceada y sistemática. Esta estrategia produce modelos más compactos y precisos en comparación con arquitecturas previas de tamaño equivalente. [22]

EfficientNetB0 fue diseñado mediante un proceso automatizado de búsqueda de arquitectura (AutoML), optimizando tanto la precisión como el uso de recursos. Utiliza bloques MBConv (Mobile Inverted Bottleneck Convolution), similares a los empleados en redes móviles como MobileNetV2, junto con técnicas como:

- Depthwise separable convolutions, para reducir el costo computacional manteniendo la

precisión.

- Función de activación Swish, que ofrece mayor expresividad que ReLU.
- Bloques Squeeze-and-Excitation para ajustar dinámicamente la relevancia de los canales.

EfficientNetB0 tiene aproximadamente 4 millones de parámetros, una cantidad significativamente menor que la de *VGG16* o *InceptionV3*, lo que lo convierte en una opción ideal para aplicaciones en tiempo real o dispositivos con recursos limitados. A pesar de su tamaño reducido, logra una precisión comparable o incluso superior a modelos mucho más pesados.
Como desventaja, su arquitectura, al estar altamente optimizada, puede ser menos intuitiva de modificar, y su rendimiento óptimo depende de un preprocesamiento adecuado de las imágenes (tamaño típico: 224×224 píxeles) y una correcta adaptación de sus capas finales para tareas específicas como la clasificación facial.

Cada uno de estos modelos fue seleccionado por sus características particulares de arquitectura, eficiencia y aplicabilidad a tareas de clasificación facial. Su evaluación en este estudio busca contrastar diferentes enfoques de diseño, desde arquitecturas clásicas y pesadas como *VGG16*, hasta modelos más modernos y optimizados como *EfficientNetB0*. A partir de esta información, se tiene la siguiente tabla de comparación de los diversos modelos como referencia para los próximos pasos de entrenamiento.

Modelo	Año	Parámetros aprox.	Ventajas principales	Desventajas principales
VGG16	2014	∼15 millones	Arquitectura simple y modular. Fácil de adaptar con Transfer Learning.	Muy pesada. Alto uso de memoria y tiempo de entrenamiento.
InceptionV3	2015	∼22 millones	Captura patrones a múltiples escalas. Optimiza computación vs profundidad.	Arquitectura compleja. Requiere mayor tamaño de imagen.
DenseNet121	2017	∼7 millones	Reutilización de características. Mejor propagación de gradientes.	Mayor uso de memoria por conexiones densas. Implementación más delicada.
EfficientNetB0	2019	∼4 millones	Alta precisión con pocos parámetros. Ideal para tiempo real y dispositivos limitados.	Arquitectura menos intuitiva. Depende de escalamiento compuesto para su rendimiento óptimo.

Table 2.1: Comparación general de arquitecturas CNN utilizadas

2.4 Transfer Learning

Transfer Learning (aprendizaje por transferencia) es una estrategia ampliamente utilizada en *Deep Learning* que consiste en reutilizar el conocimiento aprendido por un modelo previamente entrenado en una tarea o dominio de gran escala, para adaptarlo a un nuevo problema relacionado [27]. Este enfoque resulta especialmente útil en escenarios donde la cantidad de datos disponibles para el entrenamiento es limitada o donde los recursos computacionales no permiten entrenar modelos profundos desde cero [30].

En este proyecto, el proceso de entrenamiento de los modelos de *CNN* se llevó a cabo mediante *Transfer Learning*, reutilizando las arquitecturas previamente entrenadas sobre el conjunto de datos *ImageNet*. Esta elección permite aprovechar representaciones jerárquicas ya aprendidas, como la detección de bordes, texturas y patrones faciales de alto nivel, y adaptarlas a la tarea específica de clasificación de atributos faciales, reduciendo significativamente tanto el costo computacional como la necesidad de grandes volúmenes de datos etiquetados. Este enfoque es recomendado en la literatura reciente para tareas de clasificación de edad, género y etnia así como para sistemas en tiempo real [24, 25].

La arquitectura a utilizar se plantea siguiendo las recomendaciones propuestas en **Deep Learning with Python** de François Chollet [27], particularmente en lo referente al uso de modelos preentrenados como extractores de características Y la incorporación de capas densas específicas para la tarea objetivo. Estas directrices han sido ampliamente validadas en la literatura como un enfoque efectivo para maximizar el desempeño de modelos basados en redes neuronales profundas cuando se dispone de conjuntos de datos limitados, permitiendo mantener un balance adecuado entre capacidad de generalización y costo computacional.

3 Diseño y entrenamiento

A partir del análisis del estado del arte presentado en el capítulo anterior, se decidió entrenar los modelos seleccionados bajo un esquema de aprendizaje multi-tarea (multi-task learning [31, 32]), utilizando una arquitectura *CNN* con salidas múltiples para la clasificación simultánea de edad, género y etnia. Esta estrategia, a diferencia del enfoque tradicional basado en modelos independientes por atributo, ha sido menos explorada en trabajos previos de clasificación facial, los cuales suelen entrenar modelos separados para cada característica demográfica. El uso de un modelo multi-salida presenta dos ventajas principales. En primer lugar, permite reducir significativamente el tiempo total de entrenamiento, al compartir las capas convolucionales profundas entre tareas relacionadas. En segundo lugar, disminuye la latencia de inferencia, ya que un único modelo es capaz de generar todas las predicciones necesarias, lo que resulta especialmente relevante para aplicaciones en tiempo real.

El objetivo principal de esta etapa es seleccionar el modelo entrenado mediante *Transfer Learning* que logre el mejor equilibrio entre precisión de clasificación y eficiencia computacional, de manera que pueda ser utilizado de forma estable en un sistema de análisis de atención al cliente en tiempo real.

3.1 Conjunto de datos de entrenamiento

El conjunto de datos utilizado para el entrenamiento y validación de los modelos corresponde al "FairFace Face Dataset" [33], el cual fue diseñado explícitamente para mitigar sesgos demográficos en tareas de clasificación facial. Este dataset contiene imágenes balanceadas según género, edad y etnia, lo que lo convierte en una referencia adecuada para aplicaciones orientadas a la equidad y análisis de sesgos.

El dataset está compuesto por un total de 97.698 imágenes, de las cuales 86.744 son utilizadas para entrenamiento y 10.954 para validación. Cada imagen se encuentra etiquetada con tres atributos:

- Edad, distribuida en nueve rangos etarios.
- Género, clasificado en masculino y femenino.

- Raza o etnia, distribuida en seis categorías: Asiático, Blanco, Hispano, Negro, Indio y Medio Oriente.

VAL	Edad	Genero	Etnia		TRAIN	Edad	Genero	Etnia
0-2'	199				0-2'	1792		
3-9'	1356				3-9'	10408		
10-19'	1181				10-19'	9103		
20-29'	3300				20-29'	25598		
30-39'	2330				30-39'	19250		
40-49'	1353				40-49'	10744		
50-59'	796				50-59'	6228		
60-69'	321				60-69'	2779		
'more than 70'	118				'more than 70'	842		
Male		5792			Male		45986	
Female		5162			Female		40758	
Asian			2965		Asian			23082
Black			1556		Black			12233
Hispanic			1623		Hispanic			13367
Indian			1516		Indian			12319
Middle Eastern			1209		Middle Eastern			9216
White			2085		White			16527
TOTAL	10954	10954	10954		TOTAL	86744	86744	86744

Como primera aproximación, las etiquetas de edad fueron transformadas a un valor numérico promedio por rango, con el objetivo de abordar la estimación de edad como un problema de regresión. Esta decisión permite explorar el impacto de combinar tareas de regresión y clasificación dentro de una misma arquitectura CNN.

3.2 Arquitectura

La arquitectura de entrenamiento utilizada fue común a todos los modelos evaluados, esta consiste en eliminar y reestructurar las capas superiores o top, incorporando capas adicionales entrenables:

- Capas de aplanamiento (Flatten): utilizadas para convertir las características faciales extraídas por la red en un formato adecuado para su procesamiento por las capas densas.
- Capa de abandono (Dropout): capa con una tasa de abandono definida, cuyo propósito es prevenir el sobreajuste mediante la desactivación aleatoria de un porcentaje de neuronas durante el entrenamiento, favoreciendo así una mejor capacidad de generalización.
- Capa completamente conectada (Dense): capa densa con función de activación ReLU, cuyo número de neuronas se definió según la complejidad y el tamaño de cada modelo.

Esta capa permite al modelo aprender representaciones más complejas a partir de las características extraídas.

Finalmente, se incorporaron las nuevas capas de salida diseñadas específicamente para una clasificación múltiple con tres salidas independientes, correspondientes a los atributos de interés junto a sus funciones de activación de acuerdo con la naturaleza de cada variable:

- Edad: abordada como un problema de regresión mediante una capa `Dense(1, activation='relu')`.
- Género: formulado como un problema de clasificación binaria mediante una capa `Dense(1, activation='sigmoid')`.
- Etnia: planteado como un problema de clasificación multiclase con seis categorías, utilizando una capa `Dense(6, activation='softmax')`.

Durante el proceso de entrenamiento se empleará el optimizador Adam, dada su eficiencia y estabilidad en problemas de aprendizaje profundo [34], junto con los respectivos criterios de desempeño para cada salida:

- Error absoluto medio (Mean Absolute Error, MAE) para la estimación de edad.
- Exactitud (accuracy) para la clasificación de género.
- Exactitud (accuracy) para la clasificación de etnia.

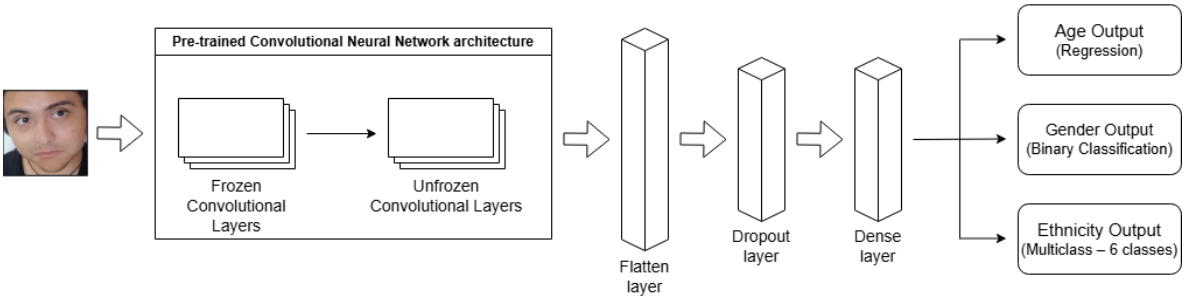


Figure 3.1: Diagrama de Arquitectura

3.3 Proceso de entrenamiento

El flujo general de entrenamiento se inicia con una etapa de preprocesamiento del dataset FairFace, seguida por la detección facial mediante el modelo *BlazeFace*. Los rostros detectados son utilizados como entrada de entrenamiento para las arquitecturas *CNN* congeladas —*VGG16*, *InceptionV3*, *DenseNet* y *EfficientNetB0*— sobre las cuales se aplica el proceso de *Transfer Learning*. Donde, las capas densas añadidas permiten generar las predicciones correspondientes a edad, género y etnia. Este flujo completo se resume en la Figura 3.2.

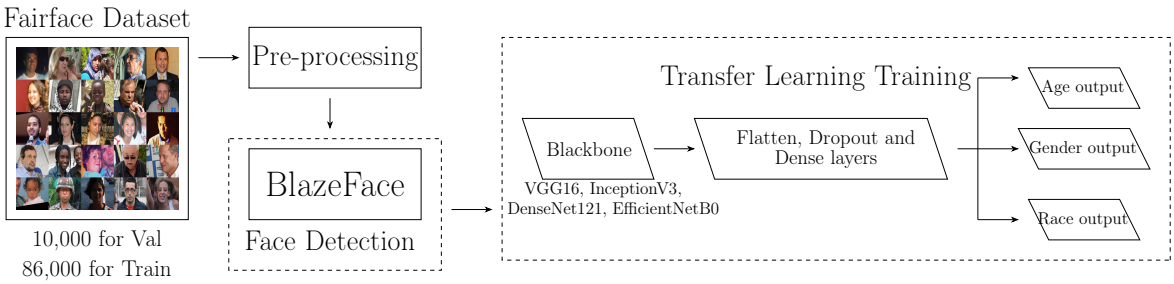


Figure 3.2: Training Pipeline

Los modelos fueron entrenados utilizando un máximo de 15 épocas, incorporando los callbacks Early Stopping y Model Checkpoint, con el objetivo de prevenir el sobreentrenamiento y optimizar el tiempo de cómputo. Asimismo, se fijó una tasa de abandono del 25 % en la capa Dropout, y se ajustaron parámetros como el tamaño del batch y la dimensión de la capa densa según la arquitectura evaluada.

Debido a limitaciones computacionales, los entrenamientos iniciales se realizaron utilizando un subconjunto del dataset, compuesto por 20.000 imágenes de entrenamiento, con el objetivo de obtener resultados preliminares y comparar el desempeño relativo de los distintos modelos. Posteriormente, el modelo seleccionado se plantea entrenar utilizando el conjunto completo de datos.

La partición para validación se definió según una proporción de 86/10 respecto al número de imágenes de entrenamiento utilizadas, dejando únicamente 10.000 imágenes para validación, mientras que las 1.698 imágenes restantes se reservaron para la evaluación de los modelos.

En las siguientes subsección se describen los resultados obtenidos durante los entrenamientos preliminares de los modelos, evaluados bajo la arquitectura definida en la Sección 3.2.

3.3.1 VGG16

El modelo fue entrenado utilizando imágenes de entrada con una resolución de 224×224 píxeles, de acuerdo con los requerimientos de su arquitectura original. La red base cuenta con un total de 14.714.688 parámetros distribuidos en 26 capas.

Para la aplicación de *Transfer Learning*, se congelaron 6 capas convolucionales profundas del modelo base, permitiendo el entrenamiento de 7.079.424 parámetros. Tras esta modificación y en conjunto con las capas añadidas, el modelo final alcanzó un total de 27.564.360 parámetros, de los cuales 19.929.096 fueron entrenables.

En la siguiente Tabla se resume la configuración de entrenamiento, incluyendo el tamaño total de parámetros, el número de parámetros entrenables y los detalles de los hiperparámetros relevantes utilizados durante el proceso de entrenamiento.

Total de parámetros originales	14714688 (56.13 MB) / 26 capas
Parámetros a entrenar	7079424 (27.01 MB) / 6 capas
Total de parámetros modelo final	27564360 (105.15 MB)
Total de parámetros a entrenar	19929096 (76.02 MB)
Capa Dense	512, activación: "relu"
Capa Dropout	0.25
Tamaño Batch	32

El modelo fue entrenado durante el máximo configurado de 15 épocas. El tiempo total de entrenamiento fue de 13:13:22, reflejando el costo computacional asociado a esta arquitectura bajo el esquema propuesto. Se lograron los siguientes resultados de evaluación sobre el conjunto de prueba:

Pérdida del test	255.2047
Edad - MAE	12.8155
Género - Exactitud	53.65%
Etnia - Exactitud	27.40%

3.3.2 InceptionV3

El modelo *InceptionV3* fue entrenado utilizando imágenes de entrada redimensionadas a una resolución de 299×299 píxeles, de acuerdo con la configuración original descrita en su trabajo de referencia. La arquitectura base cuenta con un total de 21.802.784 parámetros, distribuidos en 188 capas.

Para la aplicación de *Transfer Learning*, se congelaron las capas convolucionales profundas del modelo base, permitiendo el entrenamiento de 7.173.312 parámetros, correspondientes 27 capas. Tras esta modificación, el modelo final alcanzó un total de 55.493.160 parámetros, de los cuales 40.863.688 fueron entrenables.

En la siguiente Tabla se resume la configuración de entrenamiento utilizada para este modelo, incluyendo el número total de parámetros, los parámetros entrenables y los hiperparámetros relevantes empleados durante el proceso de entrenamiento.

Total de parámetros originales	21802784 (83.17 MB) / 188 capas
Parámetros a entrenar	7173312 (27.36 MB) / 27 capas
Total de parámetros modelo final	55493160 (211.69 MB)
Total de parámetros a entrenar	40863688 (155.88 MB)
Capa Dense	256, activación: "relu"
Capa Dropout	0.25
Tamaño Batch	16

El entrenamiento con 20.000 imágenes, por Early Stopping, duró 9 épocas y un tiempo total de entrenamiento de 4:24:15, con los siguientes resultados:

El entrenamiento realizado se detuvo de forma anticipada mediante Early Stopping tras 9 épocas, con un tiempo total de entrenamiento de 4:24:15. Los resultados de evaluación obtenidos sobre el conjunto de prueba fueron los siguientes:

Pérdida del test	1475.4966
Edad - MAE	34.9749
Género - Exactitud	51.43%
Etnia - Exactitud	15.35%

Cabe destacar que, de manera adicional al entrenamiento descrito con resolución de entrada 299×299 píxeles, se dispone de resultados correspondientes a un entrenamiento del modelo *InceptionV3* utilizando imágenes redimensionadas a 224×224 píxeles. Esta configuración no formaba parte del diseño experimental inicial y se originó a partir de una inconsistencia en el preprocesamiento de las imágenes.

No obstante, los resultados obtenidos bajo esta condición fueron incorporados al presente informe, dado que evidencian un mejor desempeño del modelo en términos de menor valor de pérdida, menor error en la estimación de edad (MAE) y una mayor precisión en las tareas de clasificación, en comparación con el entrenamiento realizado bajo la resolución de entrada oficial del modelo.

Total de parámetros originales	21802784 (83.17 MB) / 188 capas
Parámetros a entrenar	7173312 (27.36 MB) / 27 capas
Total de parámetros modelo final	35045928 (133.69 MB)
Total de parámetros a entrenar	20416456 (77.88 MB)
Capa Dense	256, activación: "relu"
Capa Dropout	0.25
Tamaño Batch	16

El entrenamiento se realizó durante el máximo configurado de 15 épocas, con un tiempo total de 3:45:52. Los resultados de evaluación obtenidos fueron:

Pérdida del test	310.4721
Edad - MAE	14.3514
Género - Exactitud	51.43%
Etnia - Exactitud	23.48%

3.3.3 DenseNet121

El modelo fue entrenado utilizando imágenes de entrada redimensionadas a una resolución de 224×224 píxeles, conforme a la configuración estándar de la arquitectura. El modelo base cuenta con un total de 7.037.504 parámetros, distribuidos en 362 capas.

Para la aplicación de *Transfer Learning*, se congelaron una cantidad de 26 capas del modelo base, permitiendo el entrenamiento de 641.408 parámetros. Tras las modificación de capas adicionales, el modelo final alcanzó un total de 7.301.960 parámetros, de los cuales 905.864 fueron entrenables.

La Tabla resume la configuración de entrenamiento utilizada para este modelo, incluyendo el número total de parámetros, los parámetros entrenables y los hiperparámetros relevantes empleados durante el proceso de entrenamiento.:

Total de parámetros originales	7037504 (26.85 MB) / 362 capas
Parámetros a entrenar	641408 (2.45 MB) / 26 capas
Total de parámetros modelo final	7301960 (27.85 MB)
Total de parámetros a entrenar	905864 (3.46 MB)
Capa Dense	256, activación: "relu"
Capa Dropout	0.25
Tamaño Batch	32

El proceso se completó durante el máximo configurado de 15 épocas, con un tiempo total de entrenamiento de 4:55:01. Los resultados de evaluación obtenidos sobre el conjunto de prueba fueron los siguientes:

Pérdida del test	1154.1836
Edad - MAE	29.8817
Género - Exactitud	56.99%
Etnia - Exactitud	22.40%

3.3.4 EfficientNetB0

EfficientNetB0 fue entrenado utilizando imágenes de entrada redimensionadas a una resolución de 224 × 224 píxeles, de acuerdo con la configuración estándar recomendada para esta arquitectura. El modelo base cuenta con un total de 4.049.571 parámetros, distribuidos en 211 capas.

Para la aplicación de *Transfer Learning*, se congelaron 23 capas del modelo base, permitiendo el entrenamiento de 1.462.752 parámetros. Tras las modificaciones, el modelo final alcanzó un

total de 12.078.891 parámetros, de los cuales 9.492.072 fueron entrenables.

En la Tabla siguiente se resume la configuración de entrenamiento utilizada para este modelo, incluyendo el número total de parámetros, los parámetros entrenables y los hiperparámetros relevantes empleados durante el proceso de entrenamiento.

Total de parámetros originales	4049571 (15.45 MB) / 211 capas
Parámetros a entrenar	1462752 (5.58 MB) / 23 capas
Total de parámetros modelo final	12078891 (46.08 MB)
Total de parámetros a entrenar	9492072 (36.21 MB)
Capa Dense	128, activación: "relu"
Capa Dropout	0.25
Tamaño Batch	16

El entrenamiento llevado a cabo se completo durante el máximo configurado de 15 épocas, con un tiempo total de entrenamiento de 2:58:02. Los resultados de evaluación obtenidos sobre el conjunto de prueba fueron los siguientes:

Pérdida del test	305.4417
Edad - MAE	13.6219
Género - Exactitud	51.43%
Etnia - Exactitud	14.16%

Resultados de entrenamiento

Los resultados evidencian diferencias significativas entre las arquitecturas evaluadas, tanto en la capacidad de estimación de edad como en las tareas de clasificación de género y etnia. En particular, se observa que la inclusión simultánea de una salida de regresión (edad) junto con salidas de clasificación introduce un aumento considerable en la función de pérdida total, lo que afecta de forma desigual a los distintos modelos. Este comportamiento sugiere que las arquitecturas *CNN* preentrenadas, originalmente diseñadas para tareas de clasificación, presentan limitaciones al incorporar salidas de tipo regresión junto con otras del tipo clasificación, dentro de un esquema multi-salida.

Con base en estas observaciones, y considerando la magnitud de la pérdida asociada a la

estimación de edad como variable continua, se plantea la necesidad de ajustar la estrategia de modelado. En particular, se propone modificar la arquitectura de las capas finales, reemplazando la capa de aplanamiento (Flatten) por una capa de pooling global del tipo GlobalAveragePooling2D, con el objetivo de reducir la cantidad de parámetros entrenables y mejorar la capacidad de generalización del modelo. Este tipo de capa permite condensar la información espacial extraída por la red convolucional de forma más robusta y estable, lo que ha sido ampliamente reportado en la literatura como una alternativa más adecuada que Flatten en arquitecturas profundas[35, 36].

Adicionalmente, se reformula la estimación de edad como un problema de clasificación por rangos, en concordancia con las etiquetas originales del dataset. Estos cambios motivan una segunda etapa de entrenamientos preliminares, descrita en las subsecciones siguientes, orientada a refinar la comparación entre modelos y seleccionar la arquitectura más adecuada para el entrenamiento final con el conjunto completo de datos.

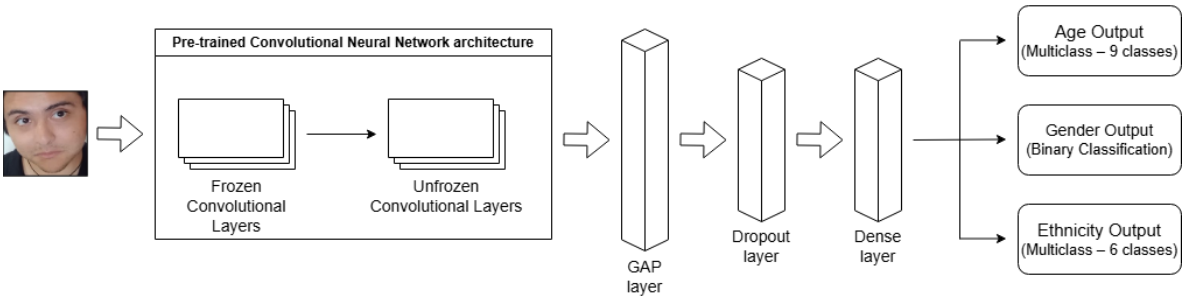


Figure 3.3: Diagrama actualizado de Arquitectura

3.3.5 VGG16

En esta segunda etapa de entrenamiento se mantiene la arquitectura original del modelo *VGG16*, junto con la estrategia de *Transfer Learning* planteada en el entrenamiento preliminar anterior.

Tras la incorporación de la nueva arquitectura de capas finales basada en Global Average Pooling, el modelo final alcanzó un total de 14.985.552 parámetros, de los cuales 7.350.288 fueron entrenables. La configuración de entrenamiento se resume a continuación:

Total de parámetros originales	14714688 (56.13 MB) / 26 capas
Parámetros a entrenar	7079424 (27.01 MB) / 6 capas
Total de parámetros modelo final	14985552 (57.17 MB)
Total de parámetros a entrenar	7350288 (28.04 MB)
Capa Dense	512, activación: "relu"
Capa Dropout	0.25
Tamaño Batch	32

El proceso de entrenamiento se detuvo de forma anticipada mediante Early Stopping tras 8 épocas, con un tiempo total de entrenamiento de 6:35:28. Los resultados obtenidos sobre el conjunto de prueba fueron los siguientes:

Pérdida del test	4.3396
Edad - Exactitud	27.96%
Género - Exactitud	51.43%
Etnia - Exactitud	26.11%

3.3.6 InceptionV3

El modelo *InceptionV3* fue reentrenado manteniendo su arquitectura base convolucional y la estrategia de *Transfer Learning* definida previamente, incorporando las modificaciones descritas en la sección de resultados de entrenamiento. En particular,la arquitectura de capas finales basada en GlobalAveragePooling2D, junto con la reformulación del atributo edad como un problema de clasificación por rangos etarios.

Si bien la arquitectura original de *InceptionV3* está diseñada para operar con imágenes de entrada de 299 $\times$ 299 píxeles, se optó por utilizar el redimensionamiento a 224 $\times$ 224 píxeles, dado que en entrenamientos anteriores esta configuración evidenció un mejor desempeño global. La configuración de entrenamiento utilizada se resume a continuación:

Total de parámetros originales	21802784 (83.17 MB) / 188 capas
Parámetros a entrenar	7173312 (27.36 MB) / 27 capas
Total de parámetros modelo final	22467120 (85.71 MB)
Total de parámetros a entrenar	7837648 (29.90 MB)
Capa Dense	256, activación: "relu"
Capa Dropout	0.25
Tamaño Batch	16

El entrenamiento se realizó durante el máximo configurado de 15 épocas, con un tiempo total de entrenamiento de 3:09:40. Los resultados obtenidos sobre el conjunto de prueba fueron los siguientes:

Pérdida del test	3.7456
Edad - Exactitud	35.66%
Género - Exactitud	69.00%
Etnia - Exactitud	39.07%

3.3.7 DenseNet121

En esta segunda etapa el modelo fue reentrenado manteniendo su arquitectura base convolucional y la estrategia de *Transfer Learning* definida previamente, incorporando las modificaciones propuestas.

Tras la incorporación de la nueva arquitectura de capas finales, el modelo final alcanzó un total de 7.304.016 parámetros, de los cuales 907.920 fueron entrenables. La configuración de entrenamiento utilizada se resume a continuación:

Total de parámetros originales	7037504 (26.85 MB) / 362 capas
Parámetros a entrenar	641408 (2.45 MB) / 26 capas
Total de parámetros modelo final	7304016 (27.86 MB)
Total de parámetros a entrenar	907920 (3.46 MB)
Capa Dense	256, activación: "relu"
Capa Dropout	0.25
Tamaño Batch	32

El entrenamiento se llevó a cabo durante 12 épocas, deteniéndose de forma anticipada mediante Early Stopping, con un tiempo total de entrenamiento de 3:40:01. Los resultados obtenidos sobre el conjunto de prueba fueron los siguientes:

Pérdida del test	3.4600
Edad - Exactitud	38.11%
Género - Exactitud	77.06%
Etnia - Exactitud	47.25%

3.3.8 EfficientNetB0

Para finalizar, *EfficientNetB0* fue reentrenado manteniendo su arquitectura base convolucional y la estrategia de *Transfer Learning* definida previamente.

Tras la incorporación de la nueva arquitectura de capas finales, el modelo final alcanzó un total de 4.287.795 parámetros, de los cuales 1.700.976 fueron entrenables. La configuración de entrenamiento utilizada se resume a continuación:

Total de parámetros originales	4049571 (15.45 MB) / 211 capas
Parámetros a entrenar	1462752 (5.58 MB) / 23 capas
Total de parámetros modelo final	4287795 (16.36 MB)
Total de parámetros a entrenar	1700976 (6.49 MB)
Capa Dense	128, activación: "relu"
Capa Dropout	0.25
Tamaño Batch	16

El entrenamiento se llevó a cabo durante 12 épocas, deteniéndose de forma anticipada mediante Early Stopping, con un tiempo total de entrenamiento de 2:06:32. Los resultados obtenidos sobre el conjunto de prueba fueron los siguientes:

Pérdida del test	2.6063
Edad - Exactitud	48.64%
Género - Exactitud	87.29%
Etnia - Exactitud	61.70%

Tras los resultados de los entrenamientos preliminares bajo la nueva configuración, se observó que EfficientNetB0 presentó el mejor desempeño global, destacando por:

- Menor valor de pérdida total.
- Mayor exactitud en las tres tareas de clasificación.
- Menor tiempo de entrenamiento.
- Menor cantidad de parámetros entrenables, favoreciendo la inferencia en tiempo real.

En función de estos resultados, se seleccionó EfficientNetB0 como modelo final a utilizar.

3.4 Modelo seleccionado

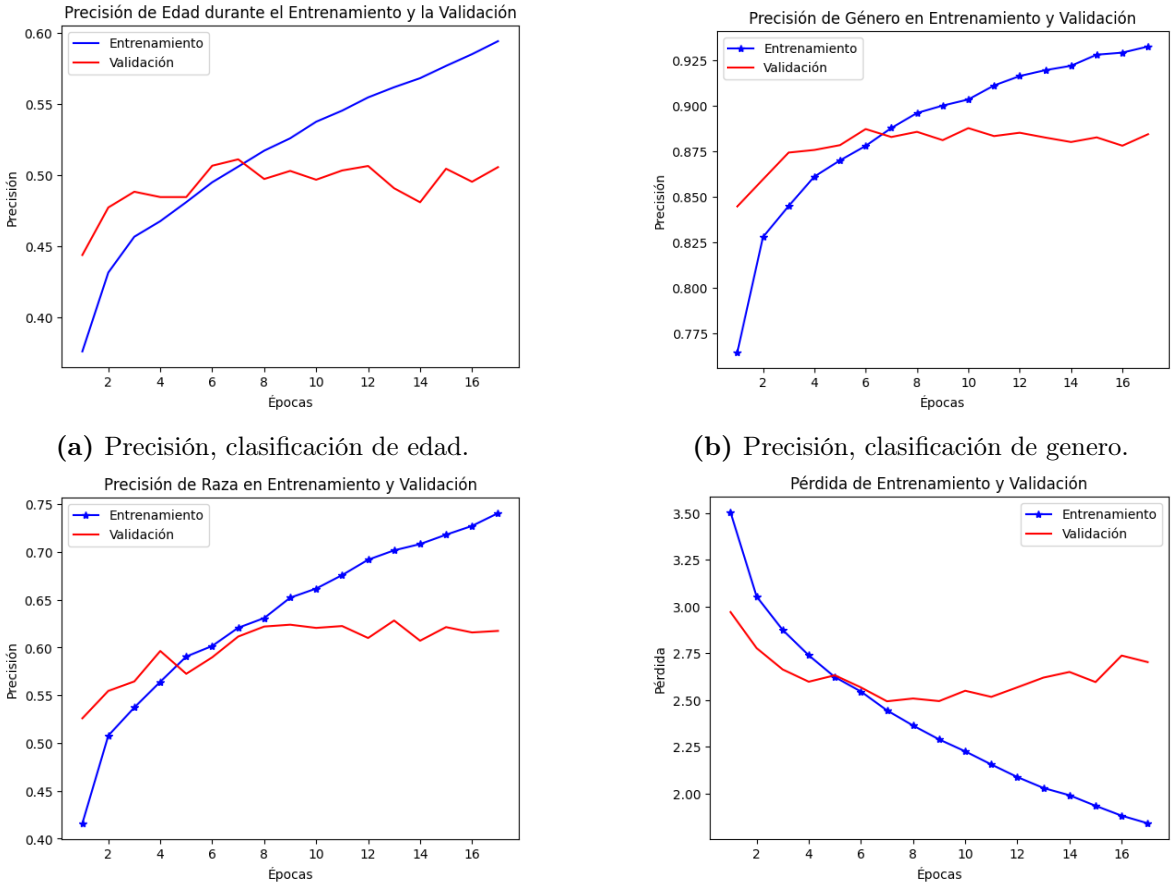
A partir de los entrenamientos preliminares realizados, se observa que reformular la estimación de edad desde una salida de regresión a una salida de clasificación por rangos reduce de manera considerable la pérdida total del modelo y aumenta la exactitud de la clasificación. Este comportamiento sugiere que las arquitecturas *CNN* preentrenadas, optimizadas originalmente para tareas de clasificación, tienden a presentar un desempeño menos estable cuando se integran salidas heterogéneas (regresión + clasificación) dentro de un esquema multi-salida, en comparación con un planteamiento completamente categórico.

Adicionalmente, los resultados evidencian una variabilidad relevante entre las arquitecturas entrenadas, en términos de tiempo de entrenamiento y cantidad de parámetros entrenables, factores directamente vinculados a la factibilidad de despliegue en un sistema operando en tiempo real y bajo restricciones computacionales.

En función del desempeño observado bajo la configuración basada en GlobalAveragePooling2D y salidas categóricas (edad, género y etnia), se seleccionó *EfficientNetB0* como arquitectura final. Esta elección se fundamenta en su mejor equilibrio global entre: menor pérdida, mayor exactitud en las tres tareas, menor tiempo de entrenamiento, y menor complejidad parametrizable, lo cual favorece la inferencia eficiente.

Una vez definido el modelo seleccionado, se incrementó el máximo de entrenamiento a 40 épocas, incorporando Early Stopping con paciencia de 10, con el objetivo de capturar la mejor época alcanzada sin incurrir en sobreentrenamiento. Sin embargo, debido a la capacidad de cómputo local disponible (16 GB de RAM) y al tamaño del dataset FairFace, el entrenamiento se restringió inicialmente a un subconjunto de 30.000 imágenes de entrenamiento y 10.000 imágenes de validación, manteniendo así un esquema viable para iteración y ajuste.

Los resultados de este entrenamiento, junto con la evolución de precisión por tarea y la pérdida durante las épocas, se presentan en la Figura 3.4.



(a) Precisión, clasificación de edad.

(b) Precisión, clasificación de genero.

(c) Precisión, clasificación de etnia.

(d) Perdida en entrenamiento.

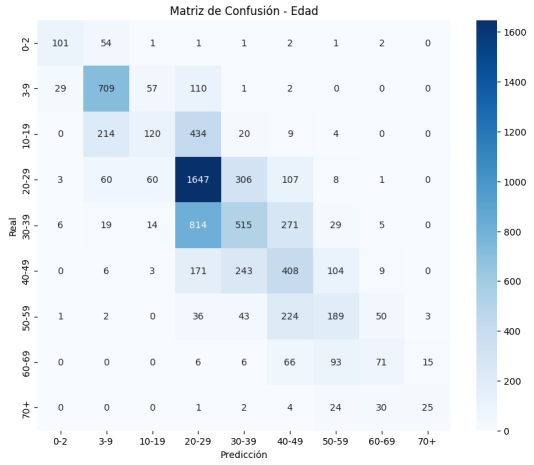
Figure 3.4: Gráficos de entrenamiento Modelo Final.

El entrenamiento del modelo *EfficientNetB0* bajo la configuración descrita se extendió hasta un total de 17 épocas con un tiempo de entrenamiento de 3:11:25, deteniéndose anticipadamente debido a la activación del criterio de Early Stopping. La mejor época registrada correspondió a la época 7, donde se obtuvo el menor valor de pérdida de validación

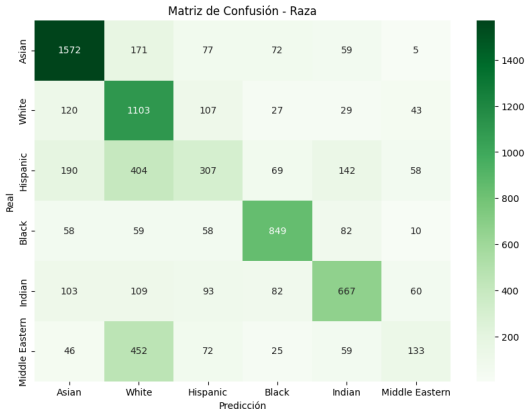
Como se aprecia en la Figura 3.4, la evolución de la pérdida muestra una tendencia decreciente durante las primeras épocas, seguida de una estabilización progresiva, mientras que las métricas de precisión asociadas a cada salida presentan un comportamiento consistente y convergente. Este patrón sugiere que el modelo logra capturar características relevantes del conjunto de datos sin incurrir en sobreajuste significativo, aun considerando el tamaño reducido del subconjunto utilizado para el entrenamiento inicial.

Se obtuvieron así, los siguiente resultados de evaluación del modelo:

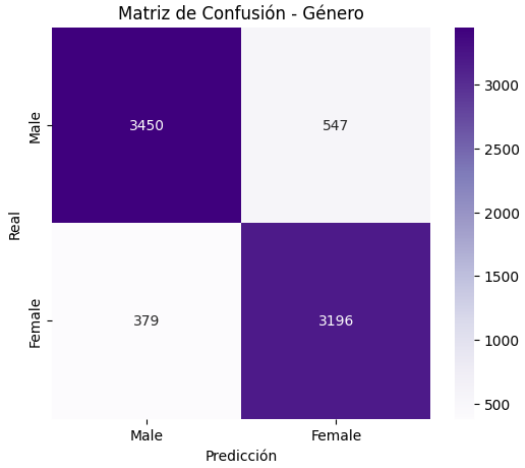
Pérdida del test	2.5159
Edad - Exactitud	49.98%
Género - Exactitud	87.77%
Etnia - Exactitud	61.16%



(a) Matriz de confusión, clasificación de Edad.



(b) Matriz de confusión, clasificación de Etnia.

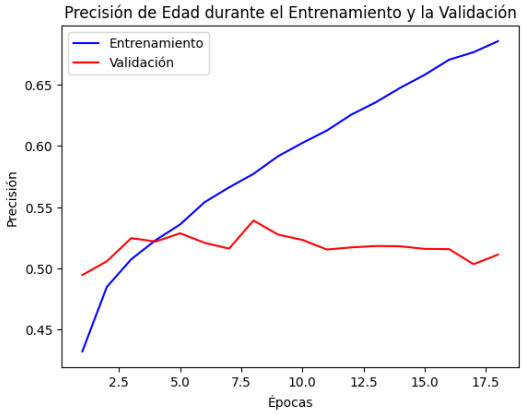


(c) Matriz de confusión, clasificación de Genero.

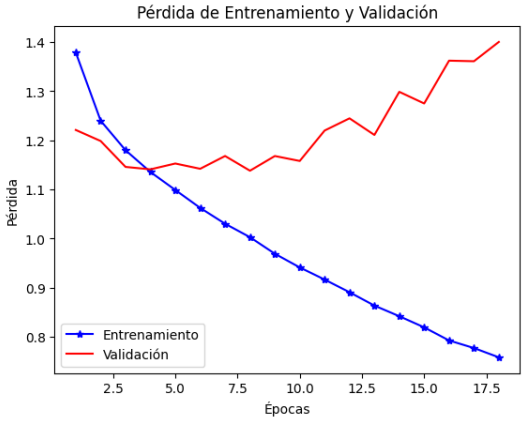
Figure 3.5: Matrices de confusión Modelo Final.

Con el objetivo de evaluar el impacto del esquema multi-salida en el desempeño global del sistema, se entrenaron adicionalmente modelos equivalentes considerando salidas individuales, es decir, arquitecturas dedicadas exclusivamente a una única tarea (edad, género o etnia). Esta comparación busca analizar no solo las métricas de precisión, sino también aspectos prácticos

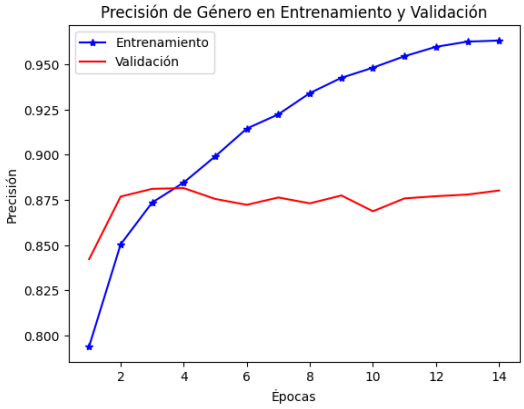
como el tiempo de entrenamiento, la complejidad computacional y la latencia de inferencia.



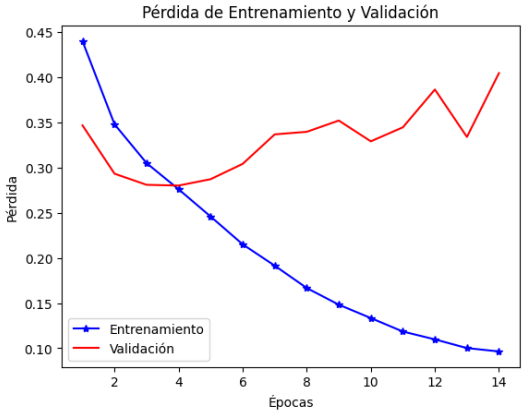
(a) Precisión, clasificación de edad.



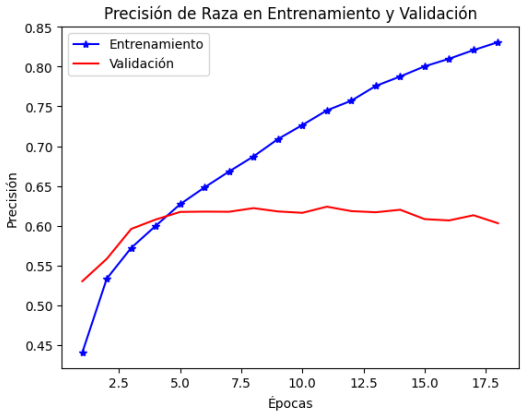
(b) Perdida, clasificación de edad.



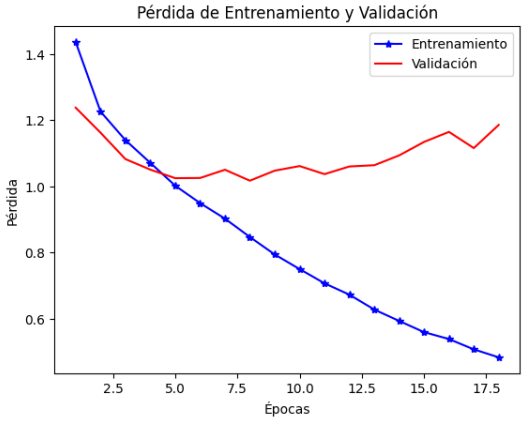
(c) Precisión, clasificación de genero.



(d) Perdida, clasificación de genero.



(e) Precisión, clasificación de etnia.



(f) Perdida, clasificación de etnia.

Figure 3.6: Gráficos de entrenamiento Modelos con salida única.

Logrando así, los siguiente resultados de evaluación sobre el conjunto de prueba de los modelos:

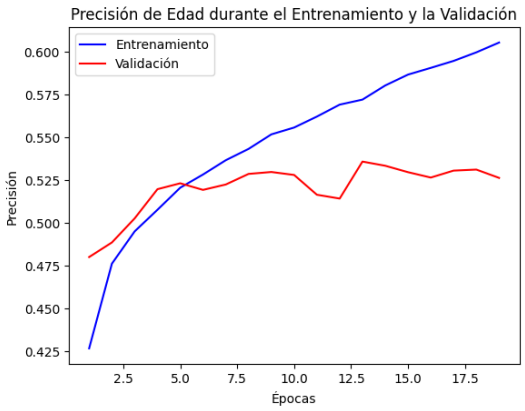
Modelo	Exactitud	Pérdida del test
EfficientNet - Edad	50.62%	1.1906
EfficientNet - Género	88.15%	0.2802
EfficientNet - etnia	62.22%	1.0167

Los resultados comparativos muestran que, si bien los modelos de salida única pueden alcanzar métricas ligeramente superiores en tareas específicas, el modelo multi-salida ofrece un mejor equilibrio global, al permitir la inferencia simultánea de múltiples atributos faciales con un costo computacional contenido. Esta característica resulta especialmente relevante para aplicaciones en tiempo real, donde la eficiencia del sistema completo prima sobre optimizaciones marginales en una sola tarea.

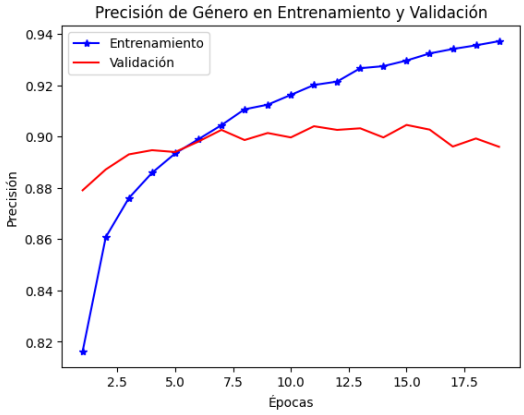
Adicionalmente, mediante pruebas de implementación en tiempo real, se observó que tanto el modelo multi-salida como los modelos de salida única presentan una latencia individual cercana a 60 ms por inferencia. No obstante, al considerar un esquema basado en tres modelos independientes ejecutados en serie, la latencia efectiva del sistema aumenta, afectando la fluidez y la escalabilidad del procesamiento en escenarios con múltiples individuos en escena.

En este contexto, el enfoque multi-salida se consolida como la alternativa más adecuada para el sistema propuesto, al ofrecer un mejor equilibrio entre precisión global, eficiencia computacional y simplicidad de implementación. La capacidad de estimar simultáneamente edad, género y etnia mediante un único modelo reduce la complejidad del pipeline y favorece su integración en aplicaciones de atención al cliente en tiempo real.

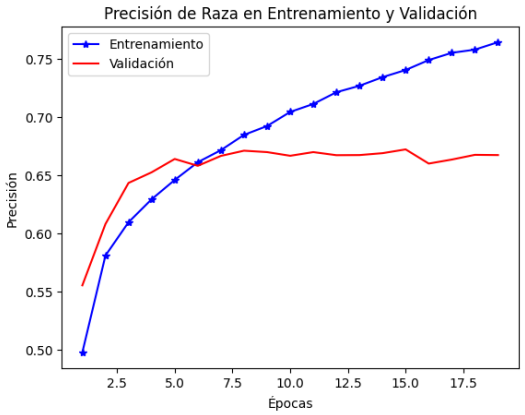
Con base en estas conclusiones, y disponiendo de mayores recursos computacionales mediante el uso de una instancia premium de Google Colab, se procedió a entrenar el modelo seleccionado utilizando el conjunto completo de datos, compuesto por 80.000 imágenes de entrenamiento y 10.000 imágenes de validación. Los resultados de este entrenamiento final, junto con su evaluación extendida y análisis de desempeño, se presentan a continuación:



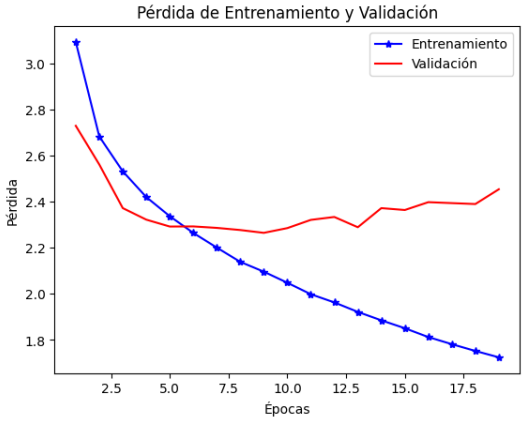
(a) Precisión, clasificación de edad.



(b) Precisión, clasificación de genero.



(c) Precisión, clasificación de etnia.

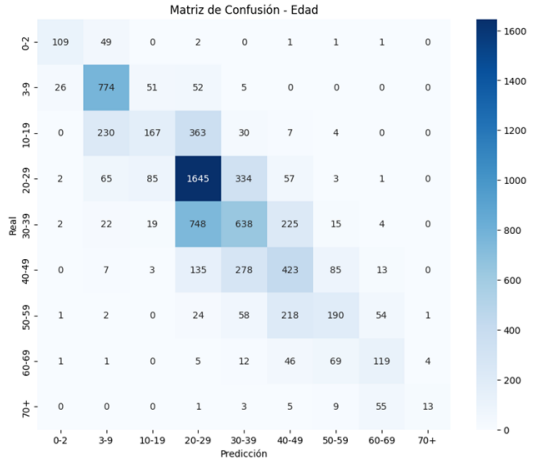


(d) Perdida en entrenamiento.

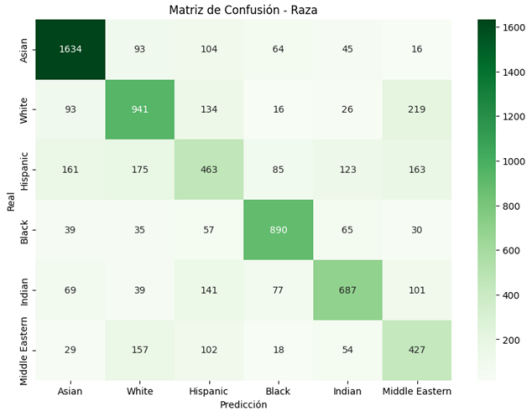
Figure 3.7: Gráficos de entrenamiento Modelo Final.

Se realizó un proceso de evaluación del modelo seleccionado, en el cual se utilizaron 7.698 imágenes de prueba, incorporando adicionalmente 6.000 imágenes que no habían sido utilizadas del dataset de entrenamiento a las 1.698 imágenes usadas en evaluaciones previas. A partir de este proceso, se obtuvieron los siguientes resultados de precisión global y las correspondientes matrices de confusión del modelo final:

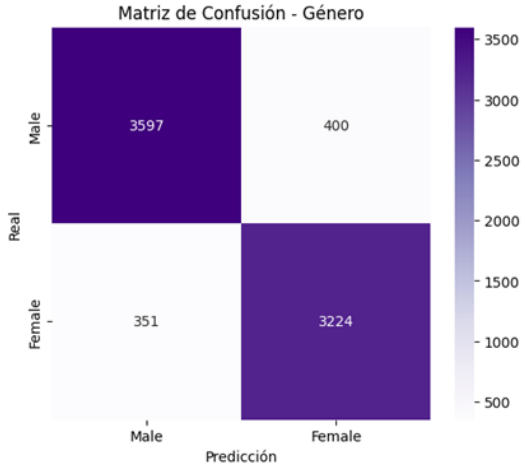
Pérdida del test	2.2644
Edad - Exactitud	53.86%
Género - Exactitud	90.08%
Etnia - Exactitud	66.59%



(a) Matriz de confusión, clasificación de Edad.



(b) Matriz de confusión, clasificación de Etnia.



(c) Matriz de confusión, clasificación de Genero.

Figure 3.8: Matrices de confusión Modelo Final.

Para finalizar, se realizaron pruebas de rendimiento en tiempo real del modelo, observándose una latencia de procesamiento cercana a 60,84 ms por inferencia.

4 Resultados

Esta sección presenta los resultados obtenidos tras la implementación del sistema completo.

4.1 Detección y seguimiento

Para la detección y seguimiento de individuos en escena se implementa el esquema basado en el uso combinado de *BlazeFace* para la detección de rostros y *Deep SORT* para el seguimiento de identidades. El sistema se configura para manejar un máximo de tres entidades detectadas de manera simultánea, asumiendo que este corresponde al número máximo de personas atendidas en la escena principal. Esta restricción permite evitar el seguimiento de individuos fuera del área de interés y reducir la latencia del sistema, favoreciendo un funcionamiento estable en tiempo real.

En el caso de *BlazeFace*, el proyecto utiliza el modelo cargado directamente desde el entorno local, sin recurrir a la implementación de MediaPipe. Para ello, se descargan previamente tanto los pesos del modelo como los anclajes (*anchors*), configurando un umbral de detección de 0.75. El preprocesamiento de las imágenes se realiza mediante la librería *OpenCV*, convirtiendo las imágenes desde el formato BGR (utilizado por defecto) a RGB, y posteriormente reescalándolas a una resolución de 128×128 píxeles antes de ser entregadas al modelo.

Para el seguimiento de identidades, se emplea el algoritmo *Deep SORT* cargado localmente desde la carpeta `deep_sort_realtime`, la cual es inicializada con parámetros personalizados: `max_age = 5`, que define el número máximo de cuadros sin detección antes de eliminar un objeto; `n_init = 3`, correspondiente a la cantidad de detecciones consecutivas necesarias para confirmar una identidad; y `max_iou_distance = 0,85`, que establece el umbral máximo de solapamiento para considerar que dos detecciones pertenecen al mismo individuo.

Adicionalmente, se implementa un proceso de filtrado que descarta aquellas identidades cuya duración en escena sea inferior a un umbral mínimo de tiempo (Por defecto, fijado en 10 segundos). Este criterio permite eliminar detecciones espurias o transitorias, mejorando la calidad y consistencia de los datos recolectados para el análisis estadístico posterior. La integración de *BlazeFace* y *Deep SORT* bajo estas condiciones permite mantener un seguimiento robusto de múltiples individuos en escena, lo cual resulta fundamental para aplicaciones orientadas a la atención al cliente.

4.2 Modelo CNN

Las imágenes de los rostros detectados en cada cuadro son sometidas a una etapa de preprocesamiento previa a su clasificación. Dicho preprocesamiento se realiza mediante la librería *OpenCV*, donde cada región facial es reescalada a una resolución de 224×224 píxeles, dimensión requerida por el modelo *CNN* y coherente con la configuración utilizada durante la etapa de entrenamiento.

De acuerdo con los resultados obtenidos durante la etapa de entrenamiento y evaluación descrita en el Capítulo 3, se seleccionó como modelo final *EfficientNetB0*, entrenado bajo el esquema de aprendizaje multi-tarea propuesto. Esta elección se fundamenta en su desempeño superior en términos de precisión global, menor valor de la función de pérdida y eficiencia computacional, aspectos críticos para su implementación en sistemas de procesamiento en tiempo real.

El modelo final alcanza una pérdida total de 2,264, con una exactitud de 90,1% en la clasificación de género, 66,6% en la clasificación de etnia y 53,9% en la clasificación de edad. Estos resultados se consideran adecuados para el objetivo del sistema, especialmente al considerar las condiciones no controladas propias de escenarios reales de atención al cliente.

En términos de rendimiento temporal, el modelo presenta una latencia promedio cercana a los 60 ms por rostro, lo que permite una ejecución fluida y estable en tiempo real. En pruebas prácticas, se observó una tasa aproximada de 7 FPS cuando se detecta un único rostro en escena, y alrededor de 2,7 FPS cuando se procesan simultáneamente hasta tres individuos. Este comportamiento confirma la viabilidad del modelo para aplicaciones en tiempo real, manteniendo un equilibrio adecuado entre precisión de clasificación y velocidad de inferencia.

La integración del modelo *CNN* dentro del sistema completo permite la obtención continua y consistente de datos clasificados, los cuales son almacenados en carpetas específicas definidas por el día de ejecución, en formato **CSV**. Estos registros son posteriormente utilizados como entrada para el análisis estadístico automatizado descrito en la sección siguiente.

4.3 Análisis Estadístico

A partir de los datos generados en cada sesión de captura, se implementa un sistema automatizado de análisis estadístico descriptivo, cuyo resultado se materializa en la generación de

informes individuales en formato PDF. Estos informes contienen métricas básicas asociadas al comportamiento de permanencia de cada individuo detectado, tales como el tiempo total de permanencia, así como la distribución de frecuencias y duraciones promedio según las categorías de edad, género y etnia.

El procesamiento de los datos se realiza mediante la agregación por identificador único (Face_ID) y los resultados de su clasificación mediante la moda detectada, permitiendo describir el comportamiento temporal de cada individuo dentro de una sesión. Sobre esta base, se incorporan visualizaciones que permiten profundizar en la exploración de los datos, entre las cuales se incluyen:

- Gráficos de Distribución, con el fin de observar categorías dominantes o minoritarias dentro del grupo, en conjunto de presencias de desbalances significativos [37, 38].
- Histogramas con estimación de densidad, empleados para analizar la distribución global de los tiempos de permanencia observados durante la sesión.
- Gráficos de violin (violin plots), utilizados para examinar la dispersión del tiempo de permanencia dentro de cada categoría y para identificar la presencia de valores atípicos. [39]
- Detalle individualizado por cada persona detectada, mostrando sus atributos predominantes, el intervalo temporal de permanencia y una imagen de referencia del rostro

Cuando se generan múltiples sesiones dentro de un mismo día, el análisis se extiende a un nivel temporal agregado, integrando todas las sesiones disponibles y produciendo un informe consolidado en formato PDF. Este informe resume el comportamiento global observado durante el período completo y constituye la base principal del análisis del sistema.

El enfoque propuesto prioriza el análisis de datos acumulados en períodos de tiempo prolongados, tales como un día, un mes o incluso un año, con el objetivo de obtener métricas representativas del patrón demográfico observado. Al trabajar con un mayor volumen de datos, se reduce la variabilidad asociada a sesiones individuales y se mejora la estabilidad de las métricas estimadas, permitiendo identificar tendencias y comportamientos recurrentes de manera más confiable.

De esta forma, el análisis temporal extendido se convierte en un componente clave del proyecto, ya que permite caracterizar de manera más precisa la distribución demográfica y el comportamiento general de los usuarios a lo largo del tiempo, fortaleciendo la validez de los

resultados obtenidos y su utilidad para la toma de decisiones.

4.4 Sistema completo

Se implementaron modelos de detección y seguimiento para la captura de individuos en tiempo real, observándose que el modelo *Deep SORT* presenta el mayor valor de latencia de procesamiento, con aproximadamente 30 ms, en comparación con el modelo de detección *BlazeFace*, cuya latencia es cercana a 0,9 ms. A medida que aumenta la cantidad de personas presentes en escena, ambas latencias se incrementan; para un máximo de tres individuos detectados simultáneamente, *Deep SORT* alcanza valores cercanos a 70 ms, mientras que *BlazeFace* presenta latencias del orden de 11 ms.

El modelo *CNN* seleccionado fue integrado al sistema completo para su uso en condiciones reales de operación, demostrando su capacidad de clasificación en tiempo real. No obstante, su desempeño se encuentra condicionado por factores propios del entorno, tales como la iluminación disponible y el número máximo de individuos detectados y clasificados de manera simultánea. Como resultado de este proceso, se genera un conjunto estructurado de datos de detección y clasificación de los individuos, el cual es utilizado posteriormente para el análisis estadístico.

Sobre los datos recolectados se realizan análisis que permiten caracterizar los resultados de clasificación a lo largo de las distintas sesiones registradas durante un mismo día. Estos análisis al detectar discrepancias, generan tablas y gráficos que facilitan la interpretación del comportamiento de la atención, tanto a nivel individual como agregado.

El sistema permite efectuar los análisis de forma local, por sesión, así como de manera global en un periodo determinado. Este último constituye el enfoque principal del proyecto, ya que proporciona una mayor cantidad de datos y, en consecuencia, una base más robusta para el análisis. El proceso culmina en la generación de un informe global que resume el comportamiento de la atención durante el período diario, a partir del cual se habilita la posibilidad de evaluar la existencia de preferencias en la atención o, alternativamente, detectar desigualdades en el trato que puedan afectar la calidad o el tiempo de atención entregado.

5 Conclusiones

El desarrollo del presente trabajo evidenció que un enfoque inicial basado en entrenamiento de redes neuronales convolucionales mediante aprendizaje por transferencia, siguiendo esquemas tradicionales descritos en la literatura, resultó suficiente para implementar un sistema funcional de clasificación facial. Sin embargo, este enfoque inicial no permitió alcanzar niveles de precisión comparables con los reportados en otros trabajos, principalmente debido a la complejidad inherente a la tarea de clasificación multi-tarea, como también la inclusión de salidas de distinta naturaleza, tales como regresión, clasificación binaria y clasificación multiclase.

A partir de estas limitaciones, se propusieron y evaluaron modificaciones estructurales orientadas a mejorar el desempeño y la estabilidad del modelo. Entre ellas, se destaca la reformulación de la estimación de edad desde un enfoque de regresión hacia una clasificación por rangos, así como la sustitución de la capa de aplanamiento por una capa de Global Average Pooling, lo que permitió mejorar la capacidad de generalización del modelo y evitar comportamientos de colapso hacia predicciones dominantes.

Adicionalmente, las restricciones computacionales presentes durante las etapas iniciales del proyecto condicionaron el uso de subconjuntos de datos y entrenamientos preliminares. No obstante, estas pruebas resultaron fundamentales para analizar el comportamiento de las arquitecturas *CNN* en tareas de clasificación facial multi-tarea y para orientar la selección del modelo final. Posteriormente, al disponer de mayores recursos de cómputo, fue posible entrenar y desplegar el sistema completo, logrando un equilibrio adecuado entre precisión de clasificación, tamaño del modelo y velocidad de inferencia.

En conjunto, los resultados obtenidos demuestran que las decisiones de diseño adoptadas permiten la implementación efectiva de un sistema de análisis automatizado de atención al cliente en condiciones reales, validando la viabilidad del enfoque propuesto y su aplicabilidad práctica dentro de un contexto ingenieril.

5.1 Trabajo futuro

Como proyección del trabajo desarrollado, se identifican diversas líneas de mejora y extensión del sistema propuesto, abarcando tanto aspectos de visión por computador como de seguimiento, modelado y análisis de datos. Estas proyecciones buscan fortalecer el desempeño del sistema y ampliar su capacidad de análisis en escenarios reales de atención al cliente.

Una primera línea de trabajo corresponde a la mejora en la obtención de características faciales mediante el uso de modelos de segmentación facial, lo que permitiría superar las limitaciones inherentes al enfoque basado en *bounding boxes* utilizado por *BlazeFace*. La segmentación del rostro posibilitaría un recorte más preciso de las regiones faciales relevantes y reduciría la influencia del fondo, lo que podría traducirse en una mejora en la calidad de las características empleadas para la clasificación.

En relación con el seguimiento de individuos, resulta pertinente explorar alternativas más livianas y eficientes que *Deep SORT*, tales como ByteTrack, con el objetivo de disminuir la latencia del sistema en escenarios con múltiples personas presentes simultáneamente en la escena. La adopción de este tipo de métodos permitiría mejorar el desempeño temporal del sistema sin comprometer la consistencia del seguimiento.

Desde el punto de vista del modelado, una posible mejora consiste en la incorporación de capas específicas para cada una de las salidas del modelo multi-tarea, permitiendo una especialización parcial de las representaciones aprendidas para la clasificación de edad, género y etnia. Este enfoque podría contribuir a mejorar el desempeño individual de cada tarea, manteniendo al mismo tiempo la eficiencia global del sistema.

Adicionalmente, se considera la ampliación del conjunto de atributos analizados, incorporando variables de carácter dinámico relacionadas con el comportamiento humano, tales como expresiones faciales, estados emocionales o información acústica, con el fin de enriquecer el análisis de la interacción entre clientes y personal.

Finalmente, se proyecta una integración más robusta del análisis estadístico con modelos LLM orientados a la interpretación automatizada de los informes, así como con sistemas de gestión de clientes (Customer Relationship Management, CRM). Esta integración permitiría generar indicadores y mecanismos de retroalimentación en tiempo real, abriendo la posibilidad de utilizar los resultados del sistema no solo como una herramienta de análisis posterior, sino también

como un apoyo activo para la mejora continua de la calidad de la atención al cliente.

References

[1] D. Kahneman, *Thinking, fast and slow*. macmillan, 2011.

[2] K. N. Lemon and P. C. Verhoef, "Understanding customer experience throughout the customer journey," *Journal of Marketing*, vol. 80, no. 6, pp. 69–96, 2016. [Online]. Available: https://doi.org/10.1509/jm.15.0420

[3] M. L. Scott, S. A. Bone, G. L. Christensen, A. Lederer, M. Mende, B. G. Christensen, and M. Cozac, "Revealing and mitigating racial bias and discrimination in financial services," *Journal of Marketing Research*, vol. 61, no. 4, pp. 598–618, 2024. [Online]. Available: https://doi.org/10.1177/00222437231176470

[4] M. Bertrand and S. Mullainathan, "Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination," *American Economic Review*, vol. 94, no. 4, p. 991–1013, September 2004. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/0002828042002561

[5] D. Pacheco Rodríguez, "Estereotipos, prejuicios y sesgos y su impacto en la valoración de pruebas declarativas en procedimientos penales y de familia," Master's thesis, Universidad de Chile, 2021, tesis de Licenciatura en Ciencias Jurídicas y Sociales.

[6] Universidad de Los Lagos, "Abordar los sesgos de género en el quehacer universitario," 2020, documento institucional. Dirección de Igualdad de Género. Accessed: Jan. 21, 2026. [Online]. Available: https://direcciondegenero.ulagos.cl/wp-content/uploads/2021/01/Abordar-los-sesgos-de-ge%CC%81nero-en-la-institucio%CC%81n.pdf

[7] Ministerio de Energía, "Sesgos inconscientes en la empleabilidad," 2021, informe institucional, Gobierno de Chile. Accessed: Jan. 21, 2026. [Online]. Available: https://energia.gob.cl/sites/default/files/documentos/estudio_de_sesgos_2021.pdf

[8] K. N. Lemon and P. C. Verhoef, "Understanding customer experience throughout the customer journey," *Journal of Marketing*, vol. 80, no. 6, pp. 69–96, 2016. [Online]. Available: https://doi.org/10.1509/jm.15.0420

[9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841517301135

[10] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2021.

[11] J. Wirtz, P. G. Patterson, W. H. Kunz, T. Gruber, V. N. Lu, S. Paluch, and A. Martins, "Brave new world: Service robots in the frontline," *Journal of Service Management*, vol. 29, no. 5, pp. 907–931, 2018.

[12] T. Davenport, A. Guha, D. Grewal, and T. Bressgott, "How artificial intelligence will change the future of marketing," *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 24–42, 2020.

[13] J. D. Kelleher, *Deep learning*. MIT press, 2019.

[14] R. Szeliski, *Computer vision: algorithms and applications*. Springer Nature, 2022.

[15] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, p. 399–458, Dec. 2003. [Online]. Available: https://doi.org/10.1145/954339.954342

[16] Y. Bengio, I. Goodfellow, A. Courville *et al.*, *Deep learning*. MIT press Cambridge, MA, USA, 2017, vol. 1.

[17] J. Liu, Y. Gu, and S. Kamijo, "Customer behavior recognition in retail store from surveillance camera," in *2015 IEEE International Symposium on Multimedia (ISM)*, 2015, pp. 154–159.

[18] A. Generosi, S. Ceccacci, and M. Mengoni, "A deep learning-based system to track and analyze customer behavior in retail store," in *2018 IEEE 8th International Conference on Consumer Electronics - Berlin (ICCE-Berlin)*, 2018, pp. 1–6.

[19] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "Blazeface: Sub-millisecond neural face detection on mobile gpus," *arXiv preprint arXiv:1907.05047*, 2019.

[20] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[21] E. Donelly, "Very deep convolutional networks for large-scale image recognition," *International Journal of Artificial Intelligence and Machine Learning*, vol. 2, no. 1, 2012.

[22] M. Tan, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, pp. 6105–6114, 2019.

[23] M. Uysal and M. DEMİRAL, "Gender, age and ethnicity estimation by image processing," *DÜMF Mühendislik Dergisi*, vol. 15, pp. 49–59, 03 2024.

[24] S. Kothari, S. Deshmukh, and S. Mehta, "Comparison of age, gender and ethnicity prediction using traditional cnn and transfer learning," in *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2022, pp. 1–4.

[25] A. Kanwar and K. D. Singh, "Prediction of age, gender, and ethnicity using cnn and facial images in real-time," in *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*, 2023, pp. 668–674.

[26] s. Balaji and P. Cp, "A comparative study of lightweight face detection models for real-time mobile applications," *IJEDR Indonesian Journal of Education and Development Research*, vol. Volume 13, p. 22, 06 2025.

[27] F. Chollet, *Deep Learning with Python*, 2nd ed. Manning Publications, 2021.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *Computer Vision Foundation*, 2015. [Online]. Available: https://arxiv.org/abs/1512.00567

[29] G. Li, M. Zhang, J. Li, F. Lv, and G. Tong, "Efficient densely connected convolutional neural networks," *Pattern Recognition*, vol. 109, p. 107610, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320320304131

[30] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International journal of computer vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[31] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017. [Online]. Available: https://arxiv.org/abs/1706.05098

[32] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task cnn model for attribute prediction," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1949–1959, 2015.

[33] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1548–1558.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[35] A. Zafar, M. Aamir, N. Mohd Nawi, A. Arshad, S. Riaz, A. Alruban, A. K. Dutta, and S. Almotairi, "A comparison of pooling methods for convolutional neural networks," *Applied Sciences*, vol. 12, no. 17, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/17/8643

[36] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. [Online]. Available: https://arxiv.org/abs/1312.4400

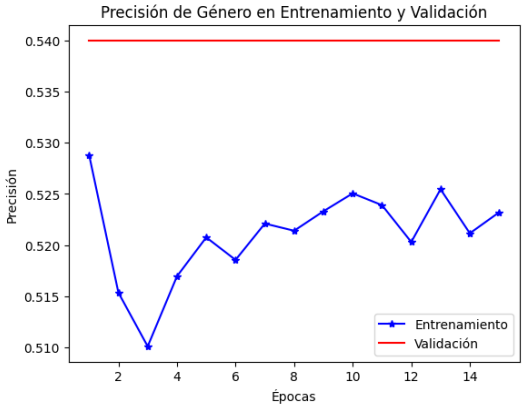
[37] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[38] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

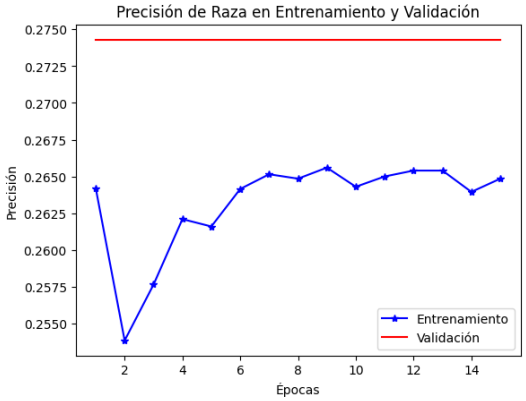
[39] C. Chatfield, "Exploratory data analysis," *European Journal of Operational Research*, vol. 23, no. 1, pp. 5–13, 1986. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377221786902092

Anexos

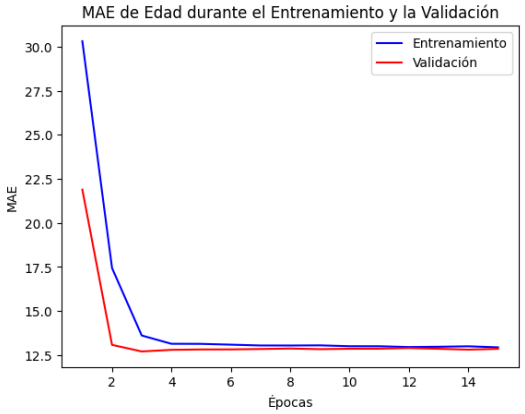
Gráficos de entrenamiento VGG16. Clasificación con regresión.



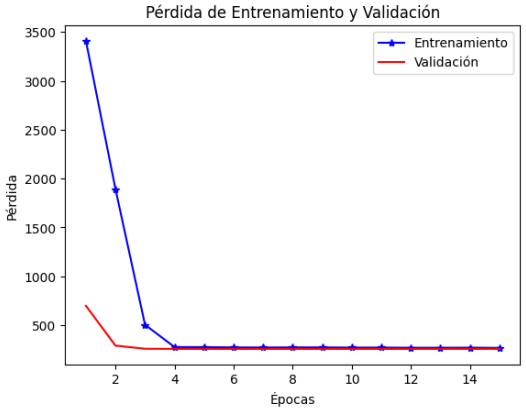
(a) Precision, clasificación de genero.



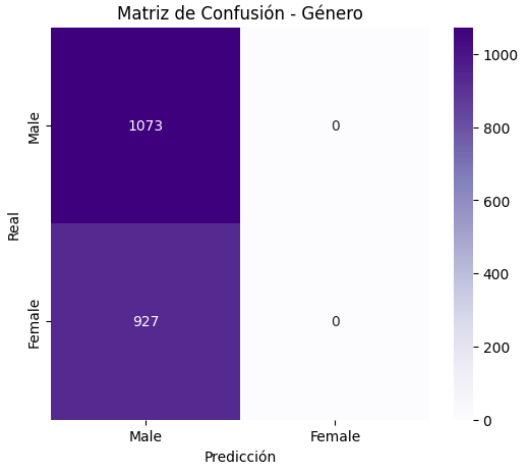
(b) Precision, clasificación de etnia.



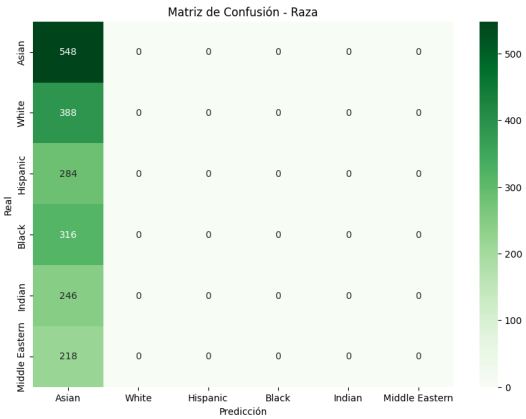
(c) MAE de edad.



(d) Perdida en entrenamiento.



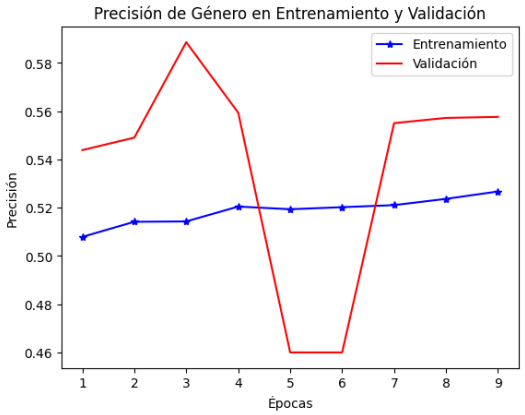
(e) Matriz de confusión, clasificación de genero.



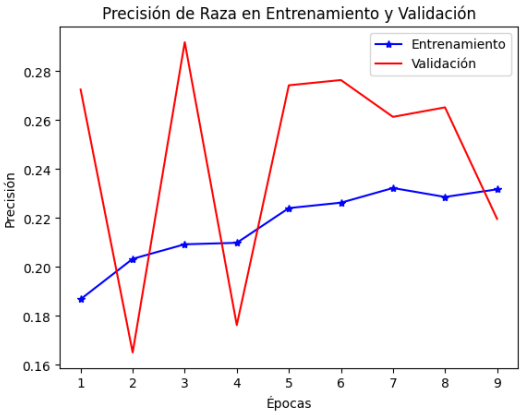
(f) Matriz de confusión, clasificación de genero.

Figure 5.9: Gráficos de entrenamiento VGG16, con edad con regresión.

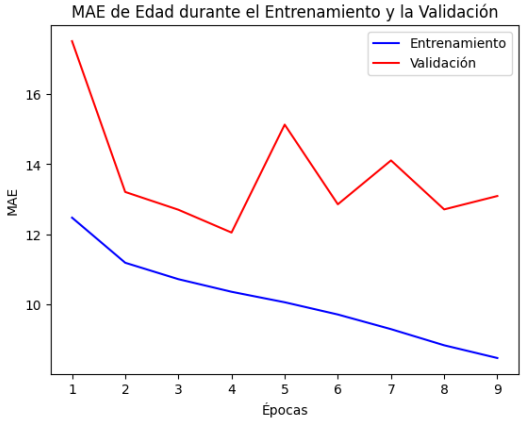
Gráficos de entrenamiento InceptionV3. Clasificación con regresión.



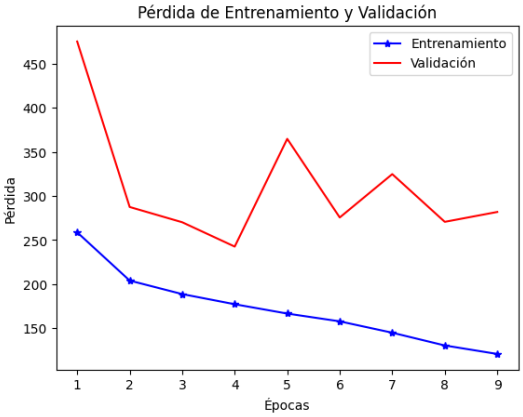
(a) Precision, clasificación de genero.



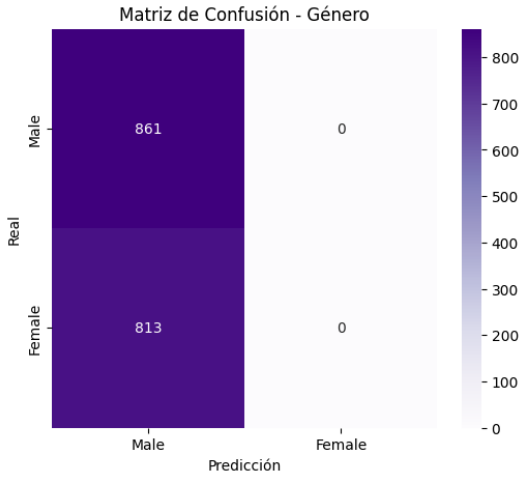
(b) Precision, clasificación de etnia.



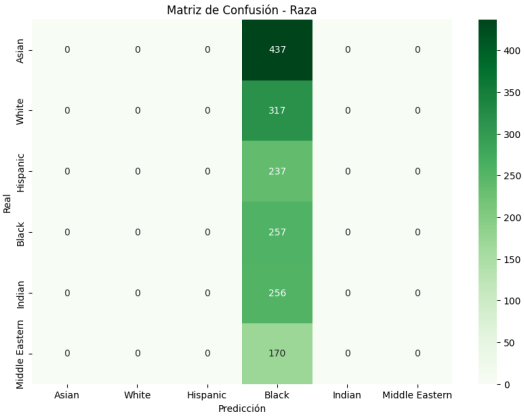
(c) MAE de edad.



(d) Perdida en entrenamiento.



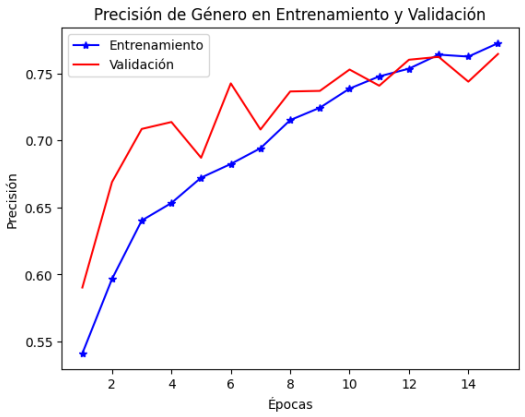
(e) Matriz de confusión, clasificación de genero.



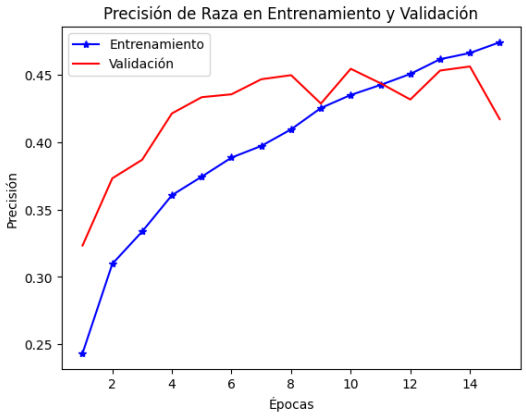
(f) Matriz de confusión, clasificación de genero.

Figure 5.10: Gráficos de entrenamiento InceptionV3, con edad con regresión.

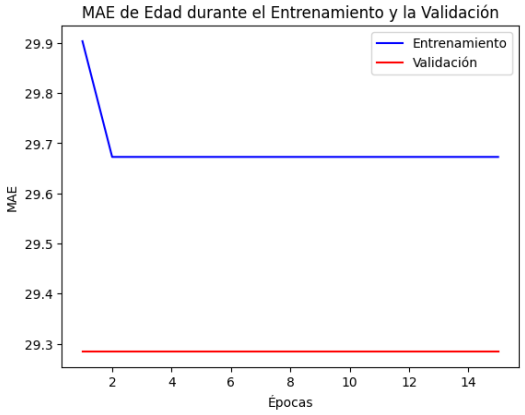
Gráficos de entrenamiento DenseNet121, Clasificación con regresión.



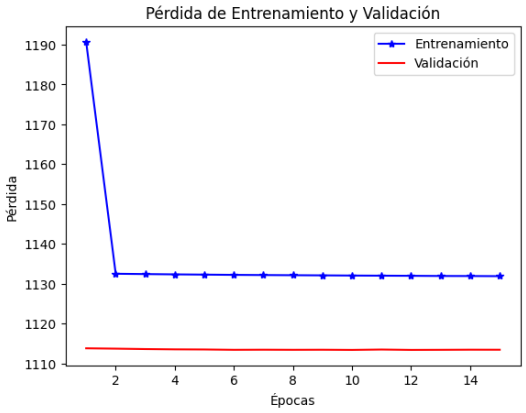
(a) Precision, clasificación de genero.



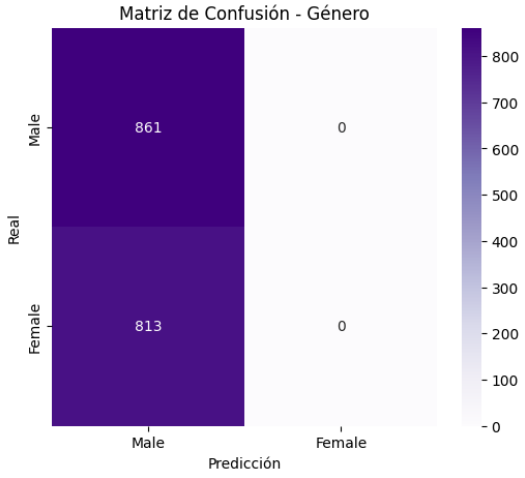
(b) Precision, clasificación de etnia.



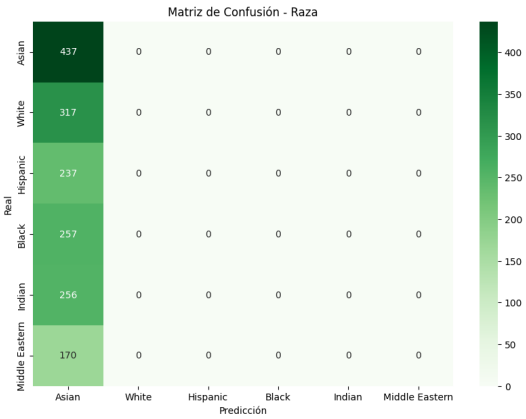
(c) MAE de edad.



(d) Perdida en entrenamiento.



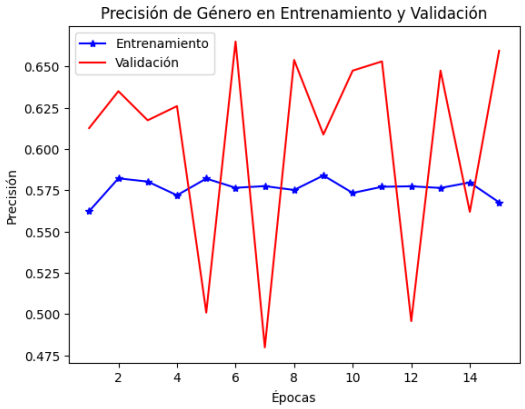
(e) Matriz de confusión, clasificación de genero.



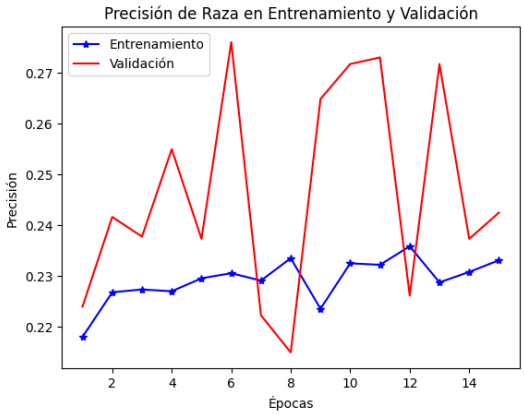
(f) Matriz de confusión, clasificación de genero.

Figure 5.11: Gráficos de entrenamiento DenseNet121, con edad con regresión.

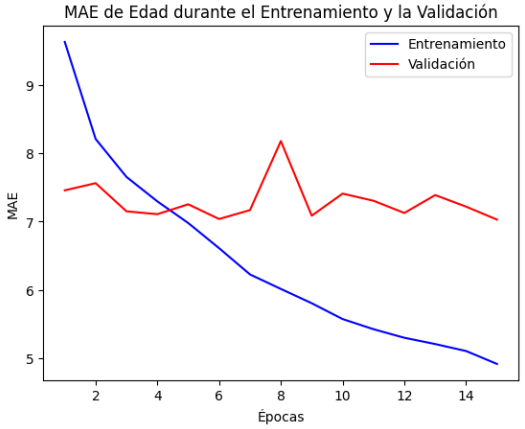
Gráficos de entrenamiento EfficientNetB0, Clasificación con regresión.



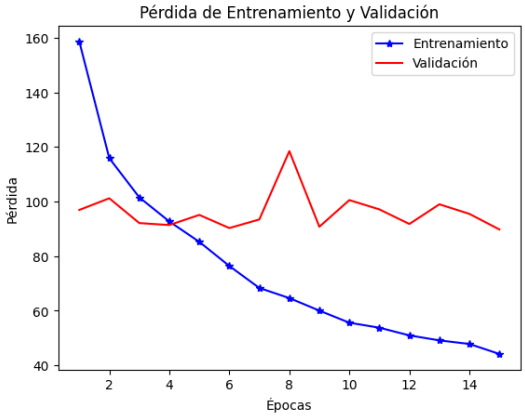
(a) Precision, clasificación de genero.



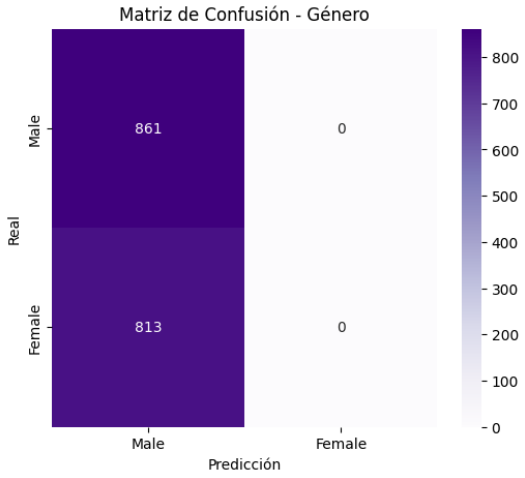
(b) Precision, clasificación de etnia.



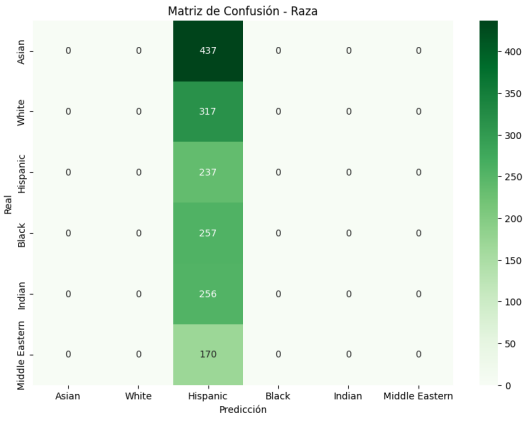
(c) MAE de edad.



(d) Perdida en entrenamiento.



(e) Matriz de confusión, clasificación de genero.



(f) Matriz de confusión, clasificación de genero.

Figure 5.12: Gráficos de entrenamiento EfficientNetB0, con edad con regresión.