Modelo de Machine Learning para clasificación de pacientes con glaucoma en la población del Valle del Cauca

PROYECTO DE GRADO

Juan Camilo Cardona Suárez
Fabio Nelson Fernández Agudelo

Asesores
Carlos Rivera Hoyos
Oftalmólogo especialista en glaucoma

Edgar Muñoz
M.Sc en Epidemiología

FACULTAD DE INGENIERÍA
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI
2022

Modelo de Machine Learning para clasificación de pacientes con glaucoma en la población del Valle del Cauca

Juan Camilo Cardona Suárez
Fabio Nelson Fernández Agudelo

Trabajo de grado para optar al título de
Magister en Ciencia de Datos

Asesores
Carlos Rivera Hoyos
Oftalmólogo especialista en glaucoma

Edgar Muñoz
M.Sc en Epidemiología

FACULTAD DE INGENIERÍA
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI
2022

CONTENIDO

pág.

LISTA DE TABLAS

LISTA DE FIGURAS

RESUMEN

En este trabajo de grado se formuló una propuesta para abordar el problema de modelos de Machine Learning (ML) no adaptados a las características raciales/étnicas de la población del Valle del Cauca para clasificar pacientes con Glaucoma. Para hacerlo, se usó la metodología CRIPS-DM (CRoss Industry Standard Process for Data Mining) Project que aborda las seis fases del ciclo de un proyecto de analítica de datos. Los modelos, técnicas y herramientas de la ciencia de datos usados para abordar la solución al problema fueron modelos de Deep Learning usando Redes Neuronales Convolucionales y de Transfer Learning usando Inception V3. La validación a la que fue sometida la propuesta consistió en evaluar los modelos entrenados en la muestra de test que fue reservada y se analizaron los resultados obtenidos en una matriz de confusión, obteniendo que el mejor modelo para clasificación del glaucoma es el modelo Inception V3 como el mejor clasificador, con un AUC ROC en el set de validación del 0.8706 y 0.9084 en el set de test, esto se logró al contar con un gran número de imágenes para entrenamiento y un modelo que fue previamente preentrenado, disminuyendo los efectos adversos de contar con una baja cantidad de datos y clases desbalanceadas. Finalmente, se puede afirmar que el enfoque de solución propuesto y la metodología empleada para obtener los resultados reportados son aceptables y permiten a futuro seguir explorando modelos más precisos.

1. INTRODUCCIÓN

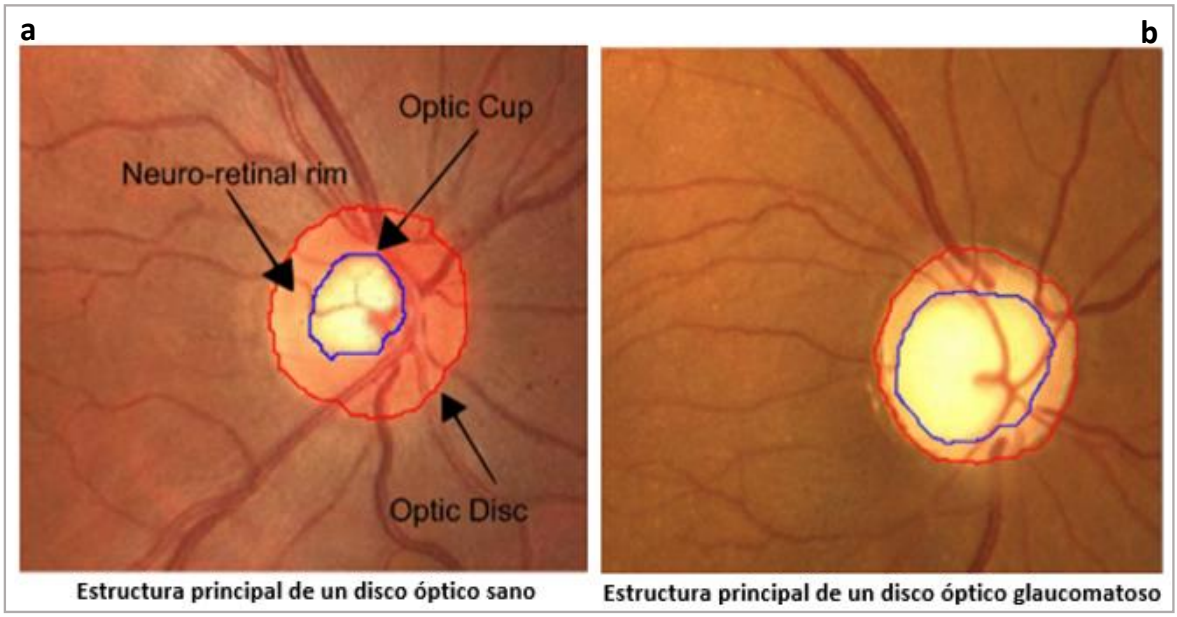
1.1 Contexto y Antecedentes

El glaucoma es una enfermedad ocular neurodegenerativa irreversible que se considera una de las principales causas de discapacidad visual en el mundo después de las cataratas (Kingman, 2004). Este término agrupa un número heterogéneo de enfermedades que se manifiestan y alteran el ojo anterior y posterior de diferentes formas. El glaucoma implica, por un lado, pérdida progresiva de las células ganglionares de la retina, células más propensas al daño glaucomatoso en el tejido del borde neurorretiniano (neuroretinal rim) de la cabeza del nervio óptico (optic nerve) que viene dada por el aumento de la presión intraocular (PIO) y/o pérdida de flujo sanguíneo al nervio el cual lleva a una disminución progresiva del campo visual. Sin embargo, la medición de la PIO no es suficientemente específica ni sensible para ser un indicador efectivo de glaucoma, ya que el daño visual puede estar presente sin aumento de la PIO.

Esta neuropatía se relaciona con diversos factores de riesgo tales como la presión intraocular elevada, disminución del flujo sanguíneo ocular, distorsión de la arquitectura de la red trabecular y aumento/reducción del drenaje del humor acuoso, junto con inflamación retiniana asociada con microglía activada, células de Müller y astrocitos, y degeneración de las células ganglionares de la retina (Križaj, 2019).

Por otra parte, la cabeza del nervio óptico es donde los axones de las células ganglionares salen del ojo y forman el disco óptico (optic disc). En una imagen de fondo de ojo, el disco óptico se puede separar visualmente en dos zonas, una zona brillante y central llamada copa óptica (optic cup) y una parte periférica llamada borde neurorretiniano (Bock, Meier, Nyúl, Hornegger, & Michelson, 2010) (Ver figura 1, panel a).

Figura 1. *Imágenes digitales de fondo de ojo recortadas alrededor del disco óptico*



Nota. Tomado de (Diaz-Pinto, y otros, 2019).

En todos los individuos tanto el disco óptico y la copa están presentes, sin embargo, una característica de un ojo glaucomatoso es la presencia de un tamaño anormal de la copa con respeto al disco óptico (excavación aumentada del nervio óptico), tal como se muestra en la figura1, paneles a y b.

Existen diversos tipos de glaucoma, los dos principales son el primario y el secundario, cada uno tiene dos subtipos principales: ángulo abierto y ángulo cerrado (Harasymowycz, y otros, 2016). El glaucoma primario o idiopático es la conjunción del glaucoma de ángulo abierto (Open Angle Glaucoma, OAG) y el glaucoma de ángulo cerrado (Angle Closure Glaucoma, ACG) sin causa identificable; por su parte, el tipo secundario es debido a causas específicas de la distorsión de la anatomía ocular (PIO).

En la siguiente tabla, se presentan los dos tipos de glaucoma con la clasificación y las causas de estos.

Tabla 1. *Clasificación y causas del glaucoma*

Clasificación del Glaucoma		
Tipos de glaucoma	***Clasificación***	***Causas***
Glaucoma de Ángulo Abierto (OAG)	Glaucoma de Ángulo Abierto Primario (POAG)	Aumento de la PIO con progresión al nervio óptico
	Glaucoma de Tensión Normal (NTG)	PIO estable con progresión y neuropatía óptica
	Glaucoma Secundario de Ángulo Abierto (SOAG)	PIO alta y/o neuropatía óptica secundario a alteraciones anatómicas del segmento anterior
Glaucoma de Ángulo Cerrado (ACG)	Glaucoma de Ángulo Cerrado Primario (PACG) - Agudo	Cierre del ángulo de la cámara anterior con un aumento repentino de la PIO
	Glaucoma de Ángulo Cerrado Primario (PACG) - Crónico	Cierre del ángulo de la cámara anterior con un aumento gradual de la PIO o desarrollo de sinequias anteriores periféricas
	Glaucoma de Ángulo Cerrado Secundario (SACG)	Cierre del ángulo de la cámara anterior con aumento de la PIO debido a causas anatómicas específicas del segmento anterior y posterior

Un estudio publicado en el año 2006 por Quigley et al., tomaron datos desde 1996 a 2006 y llegaron a la conclusión que para el año 2010 habría 60,5 millones de personas con OAG y ACG, incrementando a 79,6 millones en 2020, y de estos, el 74% tendrá OAG, llegando a ser la segunda causa principal de ceguera en el mundo (Quigley & Broman, 2006). En su estudio, determinaron que más del 80% de las personas con ACG vive en el continente asiático, principalmente las mujeres son las más afectadas; así mismo, determinaron que el OAG es más común en personas de origen africano. Uno de los hallazgos más importantes de este estudio es conforme aumente la proporción de adultos mayores a 40 años mayor probabilidad de aumentar el glaucoma en el mundo y específicamente en personas de raza negra que no pueden acceder a un sistema de salud de forma prioritaria.

En otro estudio realizado en 2021 por Tielsch et al., se evidenciaron las disparidades que sufren las minorías en los Estados Unidos, tanto comunidades negras e hispanas se ven afectadas por inequidades en salud y un racismo sistémico, el glaucoma en este caso no es la excepción (Tielsch, y otros, 1991). Las personas de raza negra tienen entre 4 y 5 veces más probabilidad de tener glaucoma que personas de raza blanca; las personas negras tienen mayor probabilidad de

desarrollar problemas de visión; y las personas hispanas, tienen de 2 a 3 veces más probabilidad de tener glaucoma que las personas de raza blanca aunado a una mayor tasa de casos no diagnosticados que su contraparte la raza blanca (Nathan & Joos, 2016).

Estos grupos étnicos presentan características anatómicas específicas tales como la presencia de una espesor corneal más delgado, factores genéticos, factores vasculares y factores socioeconómicos de inequidad en el acceso a los servicios de salud. Esta compleja interacción de factores heterogéneos y la diversidad de causas asociadas al glaucoma, hace que en los últimos años se haya volcado la atención a encontrar métodos para detectar el glaucoma con una alta probabilidad de exactitud mediante el análisis e investigación de imágenes del fondo del ojo y escaneos de tomografía de coherencia óptica (Optical Coherence Tomography, OCT) utilizando métodos de Machine Learning para el procesamiento automatizado. Para 2005 hubo únicamente un total de 6 artículos que abordaron el uso del Machine Learning como método para el procesamiento de imágenes de glaucoma, comparado al año 2020 con 81 artículos que abordaron este método.

Sin embargo, los distintos estudios e investigaciones donde han enfocado su análisis al uso del Machine Learning para caracterizar el glaucoma no han tenido una representación adecuada de personas negras e hispanas, que como se dijo anteriormente, son entre 3 a 4 veces más propensas a tener glaucoma (Tielsch, y otros, 1991).

Lo anterior, hace que se requiera tomar en cuenta variables que se ha demostrado influyen en cada individuo en la posibilidad de tener glaucoma, tales como: edad, sexo y raza/etnia; práctica que hoy en día no es estándar. Por ejemplo, el Biobanco del Reino Unido (Biobank UK) utilizan datos de imágenes para exploración, los cuales incluyen principalmente datos de individuos de ascendencia europea mayoritariamente (Fry, y otros, 2017); de igual forma, una revisión de conjuntos de

datos disponibles públicamente para imágenes oftalmológicas descubrió que 75 de los 94 conjuntos de datos de acceso público (80%) son de ciudadanos provenientes de Europa, América del Norte y Asia, con baja representación de individuos de países de ingresos medianos-bajos. Así mismo, 74% de estos conjuntos de datos no tomaron bajo estudio variables como raza/etnia, edad, sexo (Khan, y otros, 2021).

Sin duda el excluir de un análisis el estudio de variables como raza o etnia y la exclusión de población hispana y negra hace que los modelos solo se adapten a pacientes que por lo general son de origen asiático y europeo. Un estudio para comprobar la magnitud de esto hecho revisó las publicaciones más recientes de glaucoma (excluyendo artículos publicados antes de 2016 y estudios con menos de 500 participantes), utilizando términos de búsqueda de PubMed (base de acceso especializada en ciencias de la salud) que incluían “glaucoma” y “Machine Learning” e incluyendo artículos que utilizaron el aprendizaje automático para clasificar el glaucoma o caracterizarlo por medio de imágenes diagnósticas. Sus resultados concluyeron que el 59% de las investigaciones con estas especificaciones no incluyo la variable raza/etnia (Sekimitsu & Zebardast, 2021).

1.2 Planteamiento del Problema

El estudio del glaucoma ha aumentado su importancia en los últimos años, numerosos artículos han utilizado diferentes métodos de aprendizaje automático para caracterizar el glaucoma y las estructuras glaucomatosas mediante fotografías de escaneo del ojo y escaneos OCT. Sin embargo, estos estudios en su gran mayoría carecen de representación de la población y con el agravante que ni siquiera toman datos de variables sobre raza/etnia; por otra parte, las investigaciones que incluyeron representación de poblaciones negras o hispanas se enfocaron gran parte en un solo conjunto de datos que posiblemente genera sesgos en sus análisis.

En el caso colombiano, el último Censo Nacional de Población y Vivienda (2018), elaborado por el DANE (Departamento Administrativo Nacional de Estadísticas), la población Negra, Afrocolombiana, Raizal y Palenquera (NARP) era de 4.671.160 personas, lo que correspondía al 9,34% de la población total nacional. Específicamente, el Valle del Cauca es la región donde más se concentra la población NARP con 21,7% de participación (Gobernación del Valle del Cauca, 2020).

Es de vital importancia desarrollar y entrenar modelos que no generalicen las características de las diferentes razas/etnias para la detección del glaucoma. Se necesita recopilar datos de oftalmología que no sean sesgados como en su gran mayoría hasta ahora han sido implementados y sin consideración de las características demográficas y genéticas que hacen del glaucoma una enfermedad más agresiva en personas de raza negra, hispana y en minorías étnicas.

1.3 Objetivo General

Formular y validar un modelo de Machine Learning para el tamizaje de glaucoma, adaptado a la población del Valle del Cauca.

1.4 Objetivos Específicos

1. Recolectar la información obtenida de la población objeto de estudio utilizando las variables sociodemográficas, las comorbilidades y su posible relación con el glaucoma.
2. Entrenar, comparar y validar modelos de Machine Learning para la detección del glaucoma por medio de fotos a color del nervio óptico.
3. Seleccionar un modelo de clasificación de la población objeto de estudio que permita predecir con la mayor precisión la detección del glaucoma.

1.5 Organización del Documento

La estructura del presente documento inicia en el capítulo 2 con el marco teórico, donde se describe el dominio del problema y la solución a este, se encuentran las definiciones del glaucoma con sus características y una breve descripción de la población del Valle del Cauca. También se presenta en contexto la solución al problema con cada una de las características como lo son los modelos de Machine Learning, Deep Learning e Inteligencia Artificial y las respectivas métricas de evaluación. Adicionalmente en el capítulo 2 se muestra el estado del arte y una tabla comparativa con las características de los trabajos relacionados con el actual proyecto.

En el capítulo 3 se muestra la metodología abordada para el problema, la cual fue CRIPS-DM. En el capítulo 4 se estructura la propuesta de solución, así como el entendimiento y preparación de los datos, la selección de los modelos y las métricas de clasificación que se usaron. El capítulo 5, está dirigido al diseño de validación de los modelos y de los instrumentos utilizados para llegar a la solución del proyecto. En el capítulo 6, se presentan los resultados que se obtuvieron luego de la realización de los capítulos anteriores; primero se efectuó un análisis descriptivo de tipo univariado y bivariado de las imágenes, se compararon los experimentos propuestos del capítulo 4, se realizó la discusión del proyecto y se explicó el despliegue que se llevará a cabo para los próximos años. Por último, se tienen las conclusiones y algunas recomendaciones para trabajos futuros.

2. ANTECEDENTES

El problema por resolver consiste en formular y validar un Modelo de Machine Learning (ML) aplicado a la clasificación de pacientes con Glaucoma en la población del Valle del Cauca. Se busca predecir un umbral de probabilidad a partir del cual una persona puede pertenecer a una posible categoría de diagnóstico, sea positiva (tiene la enfermedad) o negativa (no tiene la enfermedad).

Dado el objetivo principal planteado, es relevante abordar algunos conceptos claves (marco teórico) e investigaciones anteriores (estado del arte) que se presentan a continuación.

2.1 Marco Teórico

2.1.1 Dominio del Problema

2.1.1.1 Glaucoma

El glaucoma es un grupo de enfermedades que se caracterizan por un daño en el nervio óptico del ojo y constituye la principal causa de ceguera irreversible en el mundo (McMonnies, 2017). El ojo produce constantemente un fluido llamado humor acuoso que se drena a través de un área llamada ángulo de drenaje o malla trabecular. Cuando esta área no funciona correctamente el fluido en exceso se acumula, produce un aumento en la presión interior del ojo lo cual lleva al daño progresivo de las fibras nerviosas del nervio óptico (Boyd, 2021). El glaucoma es una enfermedad asintomática y se considera la principal causa de ceguera en adultos mayores de 60 años, sin embargo, mediante estrategias de detección temprana se puede diagnosticar oportunamente y evitar la ceguera irreversible.

En sus inicios el glaucoma es asintomático, pero a medida que evoluciona suele causar diversas manifestaciones clínicas como un deterioro en la calidad de la visión

debida a la sensibilidad al contraste, disminución en la agudeza visual y calidad de la imagen, y ceguera irreversible de la persona (López Rojas, Belalcázar Rey, & Dávila Ramírez, 2015).

Los principales tipos son el glaucoma de ángulo abierto y el glaucoma de ángulo cerrado. El glaucoma de ángulo abierto es el más común y se presenta cuando el ojo no drena adecuadamente el fluido, por una distorsión en la malla trabecular, lo que aumenta la presión ocular y genera un daño en el nervio óptico. El glaucoma de ángulo cerrado ocurre cuando el iris de la persona está muy cerca del ángulo de drenaje del ojo, lo que puede bloquear el drenaje y aumentar rápidamente la presión ocular. Los principales síntomas son disminución de la visión, dolor ocular intenso, cefalea frontal y la presencia de visión halos o glare (Boyd, 2021).

El glaucoma de ángulo abierto se clasifica en glaucoma de ángulo abierto primario (Primary Open Angle Glaucoma, POAG), glaucoma de tensión normal (Normal Tension Glaucoma, NTG) y glaucoma secundario de ángulo abierto (Secondary Open Angle Glaucoma, SOAG). El POAG se debe a un aumento de la PIO con progresión al nervio óptico; el NTG cuenta con un PIO estable con progresión y neuropatía óptica; y, por último, el SOAG se caracteriza por una PIO alta y/o neuropatía óptica secundario a alteraciones anatómicas del segmento anterior (Harasymowycz, y otros, 2016).

El glaucoma de ángulo cerrado se puede clasificar en glaucoma de ángulo cerrado primario (Primary Angle Clousure Glaucoma, PACG) y glaucoma de ángulo cerrado secundario (Secondary Angle Closure Glaucoma, SACG). El PACG se clasifica en agudo y crónico, el primero hace referencia al cierre del ángulo de la cámara anterior con un aumento repentino de la PIO y el segundo al cierre del ángulo de la cámara anterior con un aumento gradual de la PIO o desarrollo de sinequias anteriores periféricas; el SACG se origina por el cierre del ángulo de la cámara anterior con

aumento de la PIO debido a causas anatómicas específicas del segmento anterior y posterior (Harasymowycz, y otros, 2016).

2.1.1.2 Características de las poblaciones afrocolombianas del Valle del Cauca

El Valle del Cauca es uno de los 32 departamentos de Colombia y se encuentra ubicado al sur occidente del país. Posee 42 municipios y para el año 2020, el Departamento Administrativo Nacional de Estadística (DANE) indicó que la población se estimaba en 4.532.152 habitantes, 9% del total de la población del territorio colombiano (Gobernación del Valle del Cauca, 2020).

La población departamental estimada, según el DANE es del 27.2% afrodescendiente, concentrando la cuarta parte de la población del país. La presencia indígena es estimada en 24.422 personas distribuidas entre las etnias Embera, Inga, Wounan y Chami Nasa. Los blancos y mestizos ocupan el 72.2% de la población vallecaucana (Gobernación del Valle del Cauca, 2020).

2.1.1.3 La fotografía del fondo del ojo como método de detección del glaucoma

La discapacidad visual debido a la enfermedad del glaucoma se puede prevenir con un diagnóstico oportuno. Son varios los métodos que usan los oftalmólogos para su diagnóstico. En nuestros días se reconoce que la valoración oftalmoscópica resulta insuficiente para este propósito.

La introducción de imágenes del ojo ha mejorado el proceso de diagnóstico, sin embargo, su interpretación depende de la experiencia del observador. Un método de diagnóstico que ha evolucionado en el tiempo es la fotografía del fondo del ojo la cual permite visualizar un área mucho mayor que la observada con un oftalmoscopio portátil y con un mayor grado de nitidez y resolución (Pérez Molina & León Veitía, 2017).

La fotografía del fondo del ojo se toma de la superficie interior del ojo e incluye la retina, el nervio óptico, la mácula y el polo posterior. Esta fotografía es generada por una cámara de fondo de ojo (una especie de microscopio con cámara adjunta) con la ventaja que la imagen puede ser analizada después por otros especialistas para el diagnóstico del glaucoma, de otras enfermedades oculares como por ejemplo la presencia de retinopatía diabética y/o hipertensiva o el desprendimiento de retina. Además, como se mencionó previamente, la fotografía del fondo del ojo entrega un mayor campo de visualización (Pérez Molina & León Veitía, 2017).

2.1.2 Dominio de la solución

2.1.2.1 Machine Learning, Inteligencia Artificial y Deep Learning

El aprendizaje automático (Machine Learning - ML) es un área que relaciona tanto la cibernética como la informática (ciencia de control o ciencia de la computación) (Fradkov, 2020). El estudio relacionado con este campo de la inteligencia artificial ha traído un especial interés conforme se han ido desarrollando las computadoras y softwares especializados que permiten trabajar con grandes volúmenes de datos (big data).

Para entender el origen de este concepto, es necesario remontarse a 1943. El matemático Walter Pitts y Warren McCulloch, en un estudio analizaron el cerebro como un órgano computacional y cuál era la forma de crear computadoras que fueran capaces de equiparar o superar la red neuronal humana. En 1950, Alan Mathison Turing, creó el “Test de Turing”, con el objetivo de medir la inteligencia de una computadora poniéndola a prueba al responder una conversación imitando el comportamiento humano. Para 1952, Arthur Samuel, creó un software para jugar damas chinas, con la particularidad de ser el primer programa con la capacidad de aprender a partir de la recopilación de información de partidas, haciéndose más preciso a medida que jugaba más.

También, en la década de los cincuenta, Martin Minsky, John McCarthy y otro grupo de profesionales, acuñaron el nombre de “Inteligencia Artificial" (Artificial Intelligence); y posteriormente, el psicólogo Frank Rosenblatt de la Universidad de Cornell creó una máquina llamada “Perceptrón” para reconocer las letras del alfabeto Rosenblatt (Fradkov, 2020). Este artefacto, fue el primer prototipo de redes neuronales artificiales modernas (Artificial Neural Network, ANN) con tecnología que se asemejaba al cerebro humano.

Para el principio de los años sesenta, se creó el algoritmo “Nearest neighbor” (vecinos más cercanos), capaz de reconocer patrones para dar una respuesta efectiva, y en ese entonces fue usada principalmente con fines comerciales. Esta herramienta de inteligencia artificial dio origen al Machine Learning. En los 10 años siguientes, hubo un estancamiento en nuevos procesos e innovaciones, y solo hasta finales de la década de los setenta, estudiantes de la Universidad de Stanford (1979), crearon el robot “Stanford Car” capaz de desplazarse por una habitación sorteando los obstáculos en ella (Fradkov, 2020).

En 1981, Gerald Dejong presentó un modelo de software bajo el concepto de “Explanation-Based Learning” (Aprendizaje Basado en Explicación, EBL). Este modelo consistía en una forma de aprendizaje automático, capaz de trabajar con las variables ingresadas; así mismo, tenía la funcionalidad de recibir y almacenar nuevas variables. En 1985, Terry Sejnowski, profesor e informático teórico creó el programa “NetTalk”.

Durante los próximos años, se presenció un estancamiento en nuevos descubrimientos y avances en este campo, los sectores empresariales que más se habían beneficiado, carecían de recursos y capacidad de promoción. Solo hasta el año 2003, Google publica un estudio sobre ficheros distribuidos llamado Google File System (GFS) y más tarde, en 2004, presenta una nueva forma de hacer procesamiento llamada “‘Map & Reduce”’.

Los próximos años fueron un frenesí en la revolución del procesamiento de datos; cada vez nuevos competidores en especial en el campo tecnológico estaban interesados en nuevas formas de procesamiento, softwares multifuncionales y aplicaciones web. Para 2008, IBM salió al mercado con el ordenador Watson y Microsoft, llevando el procesamiento y almacenamiento de datos a una escala global, creando servicios en la nube.

Geoffrey Hinton (2006) introduce el término “Deep Learning” (aprendizaje profundo), con nuevas arquitecturas como lo son las redes neuronales convolucionales (Convolutional neural networks), las redes de creencias profundas (Deep belief networks) y las redes neuronales recurrentes (Recurrent neural networks) (Wang & Raj, 2017), permitiendo a las computadoras distinguir objetos y texto en imágenes y videos.

2.1.2.2 Procesamiento de imágenes

La fotografía del fondo del ojo es utilizada para el diagnóstico, seguimiento de la enfermedad o para programas de detección y evaluación de patologías del polo posterior (Öhnell, Heijl, Anderson, & Bengtsson, 2017).

El aprendizaje automático ha permitido en la última década aumentar la capacidad para procesar datos con alta dimensionalidad mediante la detección de imágenes. Kirzhevsky et al. en el año 2012, expusieron los beneficios de usar redes neuronales convolucionales (Convolutional Neural Networks, CNN) en la competencia ImageNet, clasificando de forma correcta 1.000 imágenes con la tasa de error más baja lograda hasta la fecha (Krizhevsky, Sutskever, & Hinton, 2012).

El Deep Learning es una forma de aprendizaje de representación, también conocido como aprendizaje de características (Smits, Elze, Wang, & Pasquale, 2019), la computadora presenta las imágenes como una matriz dimensional en tres niveles de valores de pixeles: longitud, ancho y profundidad, esta última incluye los colores:

rojo, verde, y azul. La arquitectura de la CNN está compuesta de tres capas: capa convolucional (Convolutional Layer, CONV), capa de pooling o sub-muestreo (POOL) y capa completamente conectada (Fully Conected, FC). Al usar las capas CONV y POOL se extraen las características de la imagen cuya salida (output) es la entrada (input) a un proceso de clasificación compuestas por capas FC; la unión de ambos procesos se conoce como la CNN (Chen, Xu, Wong, Wong, & Liu, 2015).

El proceso inicia con una capa convolucional (CONV), teniendo como input una imagen de dimensiones *m*n*r*, siendo *m* y *n* el alto y ancho respectivamente y *r* el número de canales, para una imagen RGB por ejemplo el número de canales sería 3 (profundidad). Esta capa cuenta con *K* filtros (también conocidos como Kernels), cada Kernel aplica un proceso de transformación en la imagen original, arrojando como output *k* mapas de características; estos mapas se pasan a una capa de funciones de activación (Activation Functions, AF). Para problemas de clasificación, se utilizan funciones de activación rectificadoras (Rectified Linear Unit, ReLu) para las capas convolucionales y para la capa de salida se aplica una función de activación de tipo Softmax (Chen, Xu, Wong, Wong, & Liu, 2015).

La capa POOL posterior a la capa ReLu realiza una operación de submuestro para las dimensiones *m*r* (alto y ancho) de la imagen, cuyo resultado es una imagen de tamaño reducido. Este submuestro se realiza principalmente con la operación max-pooling o con mean-pooling. Por lo general, se realizan múltiples procesos de Conv, ReLu y POOL antes de aplicar el submuestreo.

Los outputs de características del proceso anterior se pasan a una capa denominada Flatten, la cual convierte en un vector plano los mapas de características para su posterior uso como inputs para clasificación. Usualmente, la etapa se compone de múltiples capas FC (Fully Connected) que computan las probabilidades de cada clase de la imagen de entrada. Finalmente, el output que se obtiene de las capas FC representa la distribución de probabilidad calculada a partir

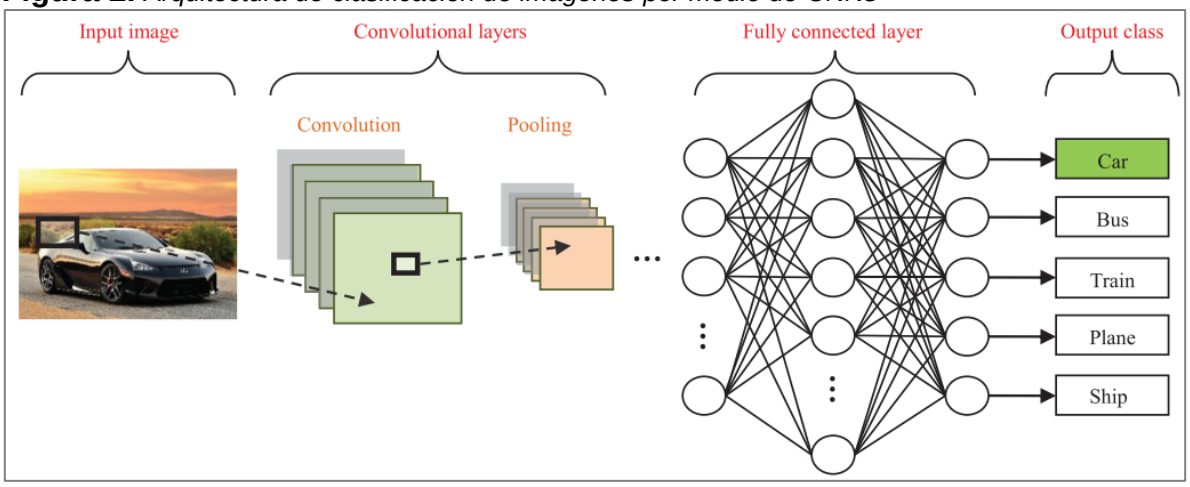
de la función Softmax, esto aplicado al contexto del glaucoma, indicaría por ejemplo que la imagen de entrada es muy probable que sea de un individuo con esta enfermedad.

2.1.2.3 Redes Neuronales Profundas

En la última década se ha demostrado que los modelos de aprendizaje profundo CNN permiten procesar múltiples capas no lineales, utilizados para la extracción y transformación de características, así como para el análisis de patrones; convirtiéndose en la arquitectura líder para tareas de reconocimiento, clasificación y detección de imágenes (LeCun, Bengio, & Hinton, 2015).

La CNN es un tipo de red neuronal artificial con aprendizaje supervisado que procesa las capas imitando al procesamiento que realiza la zona cerebral encargada de decodificar la percepción y transformarla en visión (corteza visual). La CNN está compuesta por múltiples capas ocultas que se especializan en diferentes tareas, desde las más sencillas para las primeras capas, hasta tareas complejas para las capas más profundas. Ver figura 2.

Figura 2. *Arquitectura de clasificación de imágenes por medio de CNNs*



Nota: Tomado de Rawat, W., & Wang, Z. (2017)

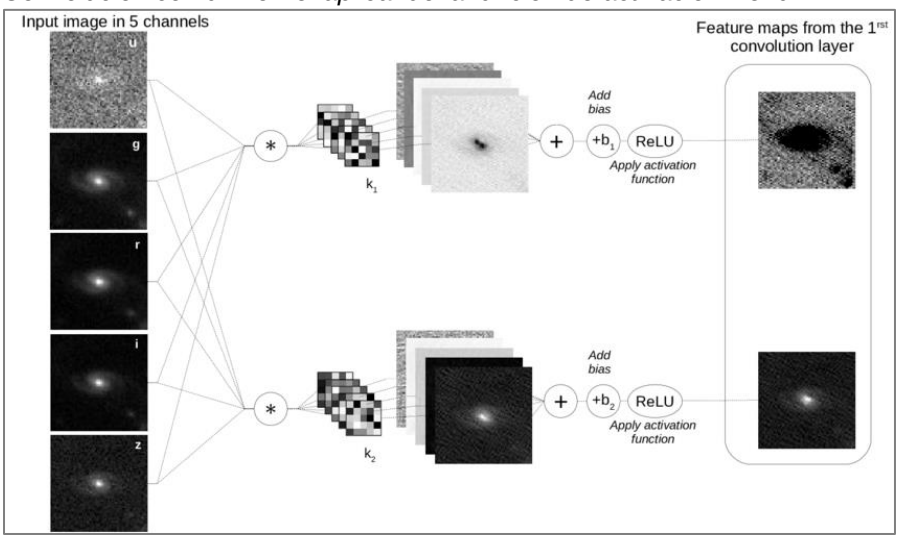
Programar la CNN requiere un gran volumen de imágenes que permita que el algoritmo por sí solo aprenda a diferenciar las características principales de cada objeto y así poder generalizarlo. Cada imagen es una matriz de pixeles que la red toma como entrada (input), oscila en un rango entre 0 a 255 pixeles y es normalizado por la red entre 0 y 1.

Si la imagen por ejemplo tiene 30*30 pixeles (alto y ancho), equivale a su multiplicación, es decir, 900 neuronas, en el caso de que fuera en un solo color (escala de grises). Para imágenes RGB (red, green, blue) se usaría 30*30*3 = 2.700 neuronas de entrada (LeCun, Bengio, & Hinton, 2015). Esta es la capa de entrada de la CNN, la cual debe ser normalizada por el valor (input) sobre 255 límite máximo del píxel.

El proceso de convolución consiste en tomar grupos de pixeles cercano a la entrada y operar matemáticamente con diferentes kernels. El kernel realiza un producto escalar con todas las neuronas de entrada generando una nueva matriz de salida (output), que será una nueva capa de neuronas ocultas. Esta salida es conocida como feature mapping. Por ejemplo, si del conjunto de kernel obtenemos un total de 34 matrices, al final tendríamos un total de 30*30*34 = 30.600 neuronas para la primera capa oculta de neuronas.

Lo anterior significa que para la primera imagen se traduce en tener 34 imágenes filtradas “nuevas”. Cada “nueva” imagen, representa ciertas características de la imagen original.

Figura 3. *Convolución con un kernel aplicando la función de activación ReLu*



Nota: Tomado de Pasquet, J et al. (2019). *Astronomy & Astrophysics.*

La función de activación ReLu (figura 3) es usualmente la más usada para trabajar con CNN, dado que los valores negativos obtenidos en cada capa de convolución se pasan a cero y los valores positivos los pasa a la siguiente capa convolucional (LeCun, Bengio, & Hinton, 2015). Si continuamos procesando las distintas imágenes de nuestro conjunto de datos, tendríamos un problema con el procesamiento de todas las imágenes, como se mostró, una sola imagen a blanco y negro en su primera convolución son 900 neuronas * 34 kernels = 30.600 neuronas. Para reducir el tamaño de la capa se hace un proceso de subsampling para quedarse con las características más importantes que se detectaron en los filtros. El tipo de subsampling más utilizado es el Max-Pooling (LeCun, Bengio, & Hinton, 2015).

Si se toma un Max-Pöoling de tamaño 2*2 (2 de alto por 2 de ancho), ahora el kernel no recorrerá las imágenes píxel por pixel, sino que se preservará el valor más alto de la matriz 4*4. Esto permite reducir nuestro conjunto de 34 imágenes pasando de dimensión 30*30 a 15*15. Ahora se tendrá 15*15*34 = 7.650 neuronas de las 30.600 que se obtuvieron sin realizar este proceso.

El proceso de convolución se itera en más capas convolucionales, pasando de reconocer patrones básicos de la imagen (líneas y curvas) a características más complejas. Al llegar a la última capa, se integra en una red neuronal multicapa (Fully Connected) la cual se le aplica una función softmax que conecta con la capa de salida final que tendrá las neuronas correspondientes con las clases que se está clasificando. El formato de salida se conoce como “one-hot-encoding” que son las probabilidades que determinada neurona de salida pertenezca a cierta clase en un rango de 0 a 1 (LeCun, Bengio, & Hinton, 2015).

Finalmente, para ajustar el valor de los pesos de los kernels, se utiliza el algoritmo de backpropagation para mejor los pesos de las interconexiones entre capas ocultas de neuronas con el objetivo de encontrar los pesos óptimos.

2.1.2.4 Modelo Inception

Entre los años de 2012 a 2014 existían diferentes arquitecturas de redes neuronales profundas, algunas de estas son AlexNet, VGG16 y GoogleNet. Desde 2014, las redes convolucionales muy profundas comenzaron a generalizarse, aplicadas con éxito en tareas de visión artificial como son la detección y seguimiento de objetos, clasificación de imágenes y localización de objeto único.

En 2015 surge el modelo Inception, su nombre se deriva del artículo científico, Network in network (Lin, Chen, & Yan, 2013) y del meme en internet “We need to go deeper”. Esta arquitectura tuvo gran acogida por sus resultados sobresalientes en el desafío de clasificación del ImageNet Large Scale Visual Recognition Challenge 2014 (ILSVRC 2014), plataforma usada para comparar algoritmos de reconocimiento y detección de imágenes, con aproximadamente 1 millón de imágenes y 1.000 clases de objetos. Este enfoque se describió en un artículo publicado por Christian Szegedy y otros autores, titulado “*Going Deeper with Convolutions”* (Szegedy, y otros, 2015).

A partir del año de publicación el modelo Inception ha tenido diferentes transformaciones, cada una se basa en la filosofía de reducir el coste computacional, modificando sus arquitecturas anteriores. Para esto, se usan diferentes técnicas para optimizar las redes, estas son convoluciones factorizadas, regularización, reducción de dimensiones y cálculos en paralelo, todas con el propósito de optimizar la red y facilitar la adopción del modelo.

2.1.2.5 Curva ROC

La clasificación binaria para problemas de clasificación se basa principalmente en seleccionar entre dos posibles alternativas, 0 y 1, “sí” y “no” o “positivo” y “negativo”. Este es un componente fundamental en la toma de decisiones en modelos de Machine Learning, Ciencia de datos y Econometría (Feng, Hong, Tang, & Wang, 2019).

El concepto usado para evaluar la calidad de la clasificación binaria y la predicción se denomina Curva Característica Operativa del Receptor (Receiver Operating Characteristic Curve, ROC curve). La curva ROC mide cuánto es capaz el modelo de distinguir entre dos clases a medida que varía su umbral de corte.

La primera aparición de los gráficos ROC fue en el artículo “The use of multiple measurements in taxonomic problems”, (Fisher, 1936), y desde entonces se han usado muy a menudo para representar la compensación entre las tasas de acierto y las tasas de falsas alarmas de los clasificadores (Isaac, 1976) como es el caso de la comunidad médica, la cual tiene una extensa literatura sobre el uso de gráficos ROC para pruebas de diagnóstico (Alemayehu & Zou, 2012).

En los últimos años, ha sido exponencial el aumento de uso de gráficos ROC en modelos de Machine Learning, dado que se ha demostrado que la precisión de clasificación simple (simple classification accuracy) es una métrica deficiente para medir el rendimiento de un modelo (Provost & Fawcett, 1997); por el contrario, se

ha probado que la curva ROC ha resultado bastante eficaz para dominios con distribución de clases sesgadas y costos de error de clasificación desiguales (Feng, Hong, Tang, & Wang, 2019).

Los problemas de clasificación de dos clases consisten en que cada instancia (i) se asigna a un elemento del conjunto {p, n}, de etiquetas de clase positiva (p) y negativa (n). Dependiendo del modelo de clasificación se pueden dar como resultado dos posibles escenarios, el primero, se calcula una estimación de la probabilidad de pertenencia a una clase de instancias, a la que se pueden aplicar diferentes umbrales para predecir la probabilidad a estar en determinada clase. La segunda, es una etiqueta discreta que solo arroja el resultado de la clase predicha para cada instancia.

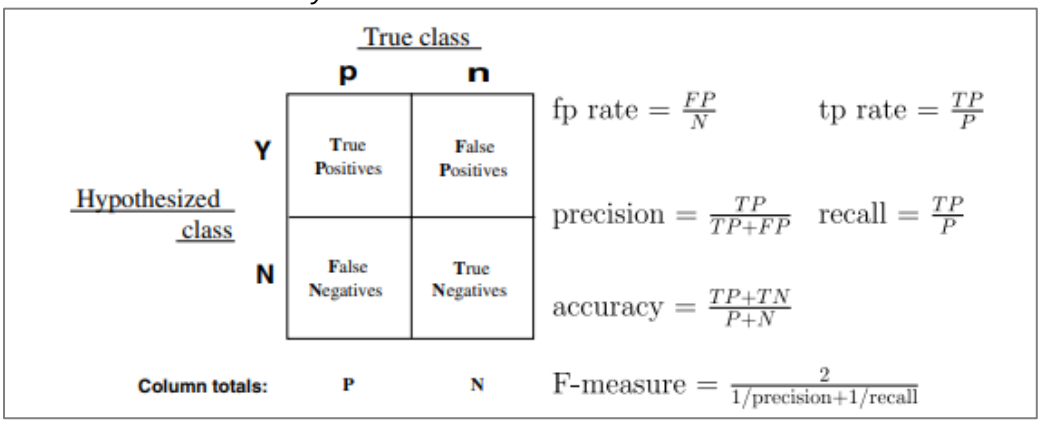
Para diferenciar entre la clase real y la pronosticada, se usan las etiquetas {Y, N}, para las predicciones de clases dadas por el modelo. Para cada instancia (i) y dado un clasificador, hay cuatro posibles resultados:

• Si la instancia es positiva y se clasifica como positiva, se cuenta como un verdadero positivo (true positive).
• Si la instancia es positiva y se clasifica como negativa, se cuenta como un falso negativo (false negative).
• Si la instancia es negativa y se clasifica como negativa, se cuenta como un verdadero negativo (true negative).
• Si la instancia es negativa y se clasifica como positiva, se cuenta como un falso positivo (false positive).

Para un conjunto de instancias (conjunto de prueba) y un clasificador dado, se genera una matriz de confusión de 2*2, llamada también tabla de contingencia, que representa las disposiciones del conjunto de instancias. La figura 4 es un ejemplo de una matriz de confusión y diferentes ecuaciones para diferentes métricas que se

pueden calcular a partir de esta. Los números de la diagonal {p, Y} y {n, N} representan las decisiones correctas del modelo y la otra diagonal son los errores, es decir, la confusión entre las clases.

Figura 4. *Matriz de confusión y métricas de rendimiento*



Nota: Tomado de (Fawcett, 2006).

2.1.2.6 Data Augmentation

En las últimas dos décadas se ha suscitado un gran interés por el aprendizaje profundo (Deep Learning) lo que implica que las CNN se hayan convertido en la herramienta más usada para análisis y clasificación de imágenes. Sin embargo, existen desafíos con este tipo de modelos. Por un lado, está el hecho que son impulsadas por redes muy grandes que implican el entrenamiento de miles de parámetros, y por otra parte está la falta de conjuntos de datos de entrenamiento confiables, lo que origina problemas de sobreajuste y poca capacidad de generalización. Por lo general, el problema más común es la falta de datos, que en algunos casos pueden ser difíciles de obtener.

Una alternativa a este problema, es aumentar los datos (data augmentation) realizando transformaciones afines y elásticas tradicionales: crear nuevas imágenes mediante la rotación o el reflejo de la imagen original, acercar y alejar, desplazar, aplicar distorsión y cambiar la paleta de colores, esto hace que el modelo sea más

robusto y evita que la red neuronal aprenda patrones irrelevantes, permitiendo un aumento del rendimiento en general (Mikołajczyk & Grochowski, 2018).

Antes de mostrar las diferentes técnicas de aumento de datos, es imprescindible definir en qué parte del Pipeline del modelo de Machine Learning se deben aumentar los datos, aquí hay dos posibles alternativas. La primera, se conoce como aumento sin conexión (offline augmentation), método usado para conjunto de datos relativamente pequeños, porque cada transformación aumenta el total de imágenes en un factor igual al número de transformaciones. La segunda opción, se conoce como aumento en línea (online augmentation), usada para conjuntos relativamente más grandes de datos, por consiguiente, se debe evitar un ascenso abrupto en el tamaño de los datos realizando transformaciones en mini-lotes (mini-batches) antes de pasarlo al modelo.

Para cada técnica de transformación se debe especificar el factor por el cual se aumentaría el tamaño de su conjunto de datos (factor de aumento de datos), suponiendo que no se necesita conocer qué hay más allá de la imagen. Las técnicas más utilizadas de aumento son:

1. Voltear (Flip): este parámetro permite voltear la imagen horizontal o verticalmente.

2. Rotación (Rotation): permite girar la imagen en un rango de 0 a 180 grados. Esto puede hacer que las dimensiones de la imagen no se conserven después de la rotación.

3. Escala (Escale): esta opción permite escalar hacia afuera o hacia adentro. Si se escala hacia afuera, el tamaño final de la imagen será mayor que el tamaño original, caso contrario, si se escala hacia adentro, la nueva imagen se recortará en determinadas secciones, pero con un tamaño igual que la original.

4. Recortar (Crop): a diferencia del escalado, parte de la imagen original se recorta y se crea una nueva imagen del tamaño de la imagen original. Este método se conoce como recorte aleatorio (random cropping).

5. Translación (Translation): implica mover la imagen a lo largo de las coordenadas X o Y o ambas. Esto permite a la red convolucional buscar casi que en cualquier parte de la imagen un objeto.

6. Ruido Gaussiano (Gaussian Noise): su objetivo es distorsionar las características de alta frecuencia, cuyos patrones se repiten en relativamente un número grande de imágenes del conjunto de datos y pueden no ser útiles. El ruido gaussiano, tiene media cero y puede distorsionar características de alta frecuencia y baja frecuencia, permitiendo a la red mejorar su capacidad de aprendizaje.

Técnicas avanzadas de aumento:

1. Redes adversarias generativas (Generative Adversarial Network): es una herramienta nueva y poderosa para realizar la generación no supervisada de nuevas imágenes utilizando la estrategia min-max (Engstrom, Tran, Tsipras, Schmidt, & Madry, 2018). Se ha descubierto que son bastante útiles para problemas de generación y manipulación de imágenes, superresolución, traducción de imagen a imagen (ejemplo bocetos a imágenes), fusión de imágenes, e imagen en pintura (restauración de partes faltantes de una imagen).

   Las GAN pueden transformar una imagen de un dominio a una imagen a otro dominio, utilizando dos redes antagónicas (G(z) y D(x)), donde una genera una imagen fotorrealista para engañar a la otra red (generador G(z)) entrenada para distinguir mejor las imágenes falsas de las reales (discriminador D(z)) (Mikołajczyk & Grochowski, 2018).

2. Transferencia de textura (Texture transfer): es una alternativa al GAN, dado que es costoso computacionalmente. Esta toma la textura, ambiente y apariencia de una imagen (se conoce como el "estilo") y lo mezcla con el contenido de otra (Efros & Freeman, 2001). Esta técnica produce un efecto al final similar al de usar GAN. Como desventaja está el hecho que el resultado parece menos realista y más artístico.

Otros enfoques:

1. Técnica de borrado aleatorio (random erasing technique): se basa en pintar aleatoriamente un rectángulo lleno de ruido en una imagen, lo que resulta en cambios en los valores de píxeles originales (Zhong, Zheng, Kang, Li, & Yang, 2020).

2.2 Estado del arte/trabajos relacionados

El glaucoma afecta principalmente al disco óptico al aumentar el tamaño de la copa y conduce a la ceguera irreversible si este no es detectado a tiempo (Nayak, Acharya, Bhat, Shetty, & Lim, 2009). Hay formas de detectarlo como la Tomografía de Coherencia Óptica, la cual permite proporcionar una imagen más completa de la verdadera naturaleza del daño y la progresión de la enfermedad (Nayak, Acharya, Bhat, Shetty, & Lim, 2009).

Nayak et al. utilizaron imágenes digitales del fondo del ojo para el preprocesamiento, los datos morfológicos y el umbral para la detección automática del disco óptico. Extrajeron las características correspondientes para validarlas usando imágenes positivas y negativas para glaucoma, por medio de un clasificador de CNN. Como resultados reportan que las características extraídas son clínicamente significativas para detectar el glaucoma. El modelo creado es capaz de detectar automáticamente la enfermedad, con una sensibilidad del 100% y especificidad del 80%.

En la actualidad, diferentes algoritmos de aprendizaje automático se han aplicado para la detección del glaucoma. En el año 2019 Pahn et al, investigaron el desempeño de las CNN para observar la discriminación usando imágenes a color del fondo del ojo. En sus estudios incluyeron imágenes con glaucoma, con sospecha de glaucoma, ojos no glaucomatosos e imágenes de baja calidad (despixeladas). Su estudio arrojó que el área que más discrimina el modelo es el disco óptico, el modelo CNN mostró un área bajo la curva de 0.9 entre los ojos con glaucoma comparados con los ojos no glaucomatosos, a diferencia de los ojos con sospecha comparados con los no glaucomatosos que mostraron un área bajo la curva de 0.7 (Phan, Satoh, Yoda, Kashiwagi, & Oshika, 2019).

Ese mismo año, Diaz-Pinto et al. validaron modelos de CNN por su alta capacidad de discriminación y exploraron 5 modelos entrenados en ImageNet (VGG16, VGG19, InceptionV3, ResNet50 y Xception). Usaron 1.707 imágenes, arrojando un AUC de 0.96, especificidad de 0.85 y sensibilidad de 0.93 usando Xception. La alta especificidad y sensibilidad son respaldadas por una estrategia de validación cruzada, ratificando que los modelos entrenados con ImageNet son una alternativa para este sistema automático de detección del glaucoma.

En 2020 Barros et al. compararon métodos que emplearon y otros que no utilizaron Deep Learning, con el objetivo de describir los pasos necesarios para el desarrollo de un sistema de diagnóstico automatizado para la detección de glaucoma en imágenes de la retina. Con base en las investigaciones evaluadas, las arquitecturas utilizadas para Machine Learning en el procesamiento de imágenes retinales, algunos estudios aplicaron extracción de características y reducción de dimensionalidad para detectar y aislar partes importantes de la imagen analizada; de manera diferente, otros trabajos utilizaron una red convolucional profunda. Sus resultados indican que las técnicas computacionales recientes, como el Deep Learning, han demostrado ser tecnologías prometedoras en la obtención de imágenes de fondo de ojo. Aunque tal técnica requiere una base de datos extensa y altos costos computacionales, los estudios demuestran que las técnicas de

aprendizaje de transferencia y aumento de datos se han aplicado como una forma alternativa de optimizar y reducir el entrenamiento de redes (Barros, y otros, 2020).

En su investigación encontraron diferentes arquitecturas empleadas para los modelos de CNN. Cada uno de estos es diferente en aspectos específicos tales como: número y tamaño de las capas, función de activación y profundidad de la red. De esta forma, no es posible determinar la arquitectura más eficiente para la clasificación del glaucoma. No obstante, la prueba empírica demostró ser la mejor manera de realizar la tarea. Si bien se evidenciaron las diferencias entre ambas arquitecturas genéricas, el desarrollo de investigaciones indica que es posible desarrollar un sistema de tamizaje automatizado para el diagnóstico de glaucoma (Barros, y otros, 2020).

También en 2020, Mursch-Edlmayr et al. revisaron los desarrollos contemporáneos de las estrategias de IA que utilizan fotografía de fondo de ojo, imágenes de tomografía de coherencia óptica (OCT) y perimetría en el diagnóstico de glaucoma y la detección de la progresión del glaucoma, y contextualizaron su potencial para ayudar a dar forma al futuro de la prestación de servicios de glaucoma. El uso de imágenes de OCT, la evaluación del campo visual (VF) mediante perimetría automatizada estándar (SAP) y el examen clínico del disco óptico respaldan el diagnóstico de lesión glaucomatosa del nervio óptico en un entorno clínico. Los algoritmos de IA aplicados a fotografías de fondo de ojo con fines de detección pueden proporcionar buenos resultados utilizando una prueba simple y ampliamente accesible. Sin embargo, para los pacientes que probablemente tengan glaucoma, se deben usar métodos más sofisticados que incluyan datos de OCT y perimetría (Mursch-Edlmayr, y otros, 2020).

Se encontró que la mayoría de los estudios no pudieron demostrar superioridad en la precisión diagnóstica en comparación con el uso del mejor parámetro de OCT convencional único (p. ej., área del borde y grosor promedio de RNFL). Sin embargo, transformaciones más complejas de los datos de OCT, incluida la

segmentación de superpíxeles en Machine Learning supervisado, un enfoque de Deep Learning híbrido, y el uso de la distancia de Mahalanobis, fueron capaces de demostrar superioridad en comparación con el uso de parámetros de OCT convencionales logrando valores AUC entre 0.86 y 0.99 (Mursch-Edlmayr, y otros, 2020).

La metodología usada por Permita Mehta et al, en abril de 2021 fue CNN, usando datos demográficos y clínicos para combinar estas características. Sus resultados fueron altamente precisos, obteniendo un área bajo la curva de 0.97, al utilizar variables relacionadas con la enfermedad como la edad, la presión intraocular y la morfología del disco óptico (Mehta, y otros, 2021). Este modelo tiene características desconocidas y cuestionadas como la función pulmonar y las capas externas de la retina, encontrando que la edad y la función pulmonar cambian con la progresión del glaucoma.

En agosto de 2021, Wu et al. utilizan la Tomografía de Coherencia Óptica (OCT) para detectar los cambios estructurales a nivel detallado que causa el glaucoma. Esta metodología arrojó demasiados parámetros los cuales confundieron a los médicos y profesionales en el área, por lo cual fue útil el uso de modelos de clasificación automática (MLC) para generar diagnósticos confiables. Se compararon diferentes modelos de clasificación basados en los parámetros OCT de Spectralis, los MLC utilizados fueron árbol de inferencia condicional, árbol de modelo logístico, arboles de decisión, bosques aleatorios y aumento de gradiente extremo (Wu, Shen, Lu, Chen, & Chen, 2021).

La regresión logística fue usada como punto de referencia para realizar esta comparación, esta regresión arrojó resultados que resaltaban los bosques aleatorios como los mejores clasificadores y hubo predictores importantes para la detección temprana del glaucoma como la medición de capa de células ganglionares. Finalizan recomendando a los médicos y profesionales en

oftalmología incluir resultados de OCT Spectralis como diagnóstico para el Glaucoma.

Tabla 2. *Resumen de estudios que utilizan ML para detectar el glaucoma a partir de fotografías de fondo de ojo*

Autores	Título	Criterio	Similitud	Diferencia	Resultado
Nayak, J., Acharya, U. R., Bhat, P. S., Shetty, N., & Lim, T. C. (2009).	Automated diagnosis of glaucoma using digital fundus images	Clasificador de CNN con características del disco óptico en imágenes positivas y negativas	* Métricas de sensibilidad y especificidad * Imágenes digitales del fondo del ojo	* Datos morfológicos utilizados para la detección del disco óptico	* Sensibilidad 1 * Especificidad 0.8
Phan, S., Satoh, S. I., Yoda, Y., Kashiwagi, K., & Oshika, T. (2019).	Evaluation of deep convolutional neural networks for glaucoma detection	Clasificación de imágenes usando modelos de CNN para la discriminación de glaucoma	* CNN para clasificar Imágenes del fondo del ojo * Mapa de calor para determinar que parte del ojo contribuye a la discriminación del modelo * Método de evaluación AUC de la curva ROC	* Uso de imágenes de sospecha de glaucoma * Efectos del tamaño de la imagen en la capacidad de discriminación	* AUC: 0.9 * Una imagen de baja calidad reduce el AUC * El disco óptico es el área más importante para la discriminación del glaucoma
Diaz-Pinto, A., Morales, S., Naranjo, V., Köhler, T., Mossi, J. M., & Navea, A. (2019).	CNNs for automatic glaucoma assessment using fundus images: an extensive validation	Detección del glaucoma usando ImageNet y modelos de Transfer Learning tales como: VGG16, VGG19, InceptionV3, ResNet50 y Xception	* CNN e Inception V3 para clasificar Imágenes del fondo del ojo	* Utilizaron los modelos VGG16, VGG19, ResNet50 y Xception	* AUC: 0.96 para Xception
Barros, D., Moura, J. C., Freire, C. R., Taleb, A. C., Valentim, R. A., & Morais, P. S. (2020).	Machine learning applied to retinal image processing for glaucoma detection: review and perspective	Revisión de diferentes desarrollos de un sistema de diagnóstico automatizado empleando métodos que utilizan la extracción de características tales como: K-NN, LS-SVM, Random Forest, Naive Bayes y SVM: y métodos de Deep Learning tales como: Inception-V3, Disc-aware ensemble network (DENet), ResNet50, MB-NN, entre otros	* CNN e Inception V3 para clasificar Imágenes del fondo del ojo * Técnicas de aprendizaje de transferencia y aumento de datos para optimizar y reducir el entrenamiento de redes * Método de evaluación AUC de la curva ROC	* Emplear métodos que utilizan la extracción de características *Emplear métodos de Deep Learning tales como ResNet50, MB-NN, entre otros	* Para los métodos de extracción de características, los resultados para el AUC están entre un rango de 0.993 y 0.901; la especificidad en un rango entre 0.994 y 0.856 y la sensibilidad entre 0.969 y 0.943 * Para los métodos de Deep Learning, los resultados para el AUC están entre un rango de 0.996 y 0.92; la especificidad en un rango entre 0.977 y 0.956 y la sensibilidad entre 0.962 y 0.923
Mursch-Edlmayr, A. S., Ng, W. S., Diniz-Filho, A., Sousa, D. C., Arnould, L., Schlenker, M. B., . . . Jayaram, H. (2020).	Artificial intelligence algorithms to diagnose glaucoma and detect glaucoma progression: translation to clinical practice	Revisión de los desarrollos contemporáneos de las estrategias de IA que utilizan fotografía de fondo de ojo, imágenes de tomografía de coherencia óptica (OCT) y perimetría en el diagnóstico de glaucoma y la detección de la progresión del glaucoma.	* CNN para clasificar Imágenes del fondo del ojo * Método de evaluación AUC de la curva ROC * Métricas de sensibilidad y especificidad	* Emplear métodos que utilizan la extracción de características * Modelos de ML para detectar glaucoma a partir de conjuntos de datos perimétricos * Modelos de ML que usaron tomografía de coherencia óptica (OCT) o perimetría	* Algoritmos de ML desarrollados con casi 50,000 imágenes de fondo de ojo identifican neuropatía óptica glaucomatosa referible con un AROC de 0.90. * Otros algoritmos de DL entrenados en imágenes de fondo de ojo y OCT coincidentes de más de 30,000 ojos pueden discriminar entre ojos

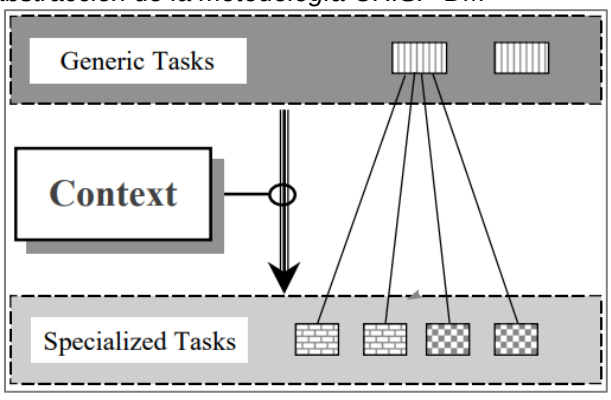
Autores	Título	Criterio	Similitud	Diferencia	Resultado
				automatizada estándar (SAP) solas o en combinación para el diagnóstico de glaucoma * Modelos de AI para detectar la progresión en ojos glaucomatosos	glaucomatosos y sanos con un AROC de 0.98. *Los algoritmos que incorporaron más parámetros clínicos e información de las pruebas de FV y las imágenes de OCT pudieron identificar a los pacientes con glaucoma con un AROC de 0.98, incluso cuando solo se utilizaron menos de 200 sujetos.
Mehta, P., Petersen, C. A., Wen, J. C., Banitt, M. R., Chen, P. P., Bojikian, K. D., . . . & Vision Consortium. (2021).	Automated detection of glaucoma with interpretable Machine Learning using clinical data and multimodal retinal images	Modelo para automatizar detección del Glaucoma usando CNN	* CNN para Imágenes del fondo del ojo * Uso de características sociodemográficas * Método de evaluación AUC de la curva ROC	* Relación de la enfermedad del glaucoma con la edad y la función pulmonar	AUC: 0.97
Wu, C. W., Shen, H. L., Lu, C. J., Chen, S. H., & Chen, H. Y. (2021).	Comparison of Different Machine Learning Classifiers for Glaucoma Diagnosis Based on Spectralis OCT	Clasificadores de aprendizaje automático (MLC) en función de los parámetros de Spectralis OCT	* Método de evaluación AUC de la curva ROC * Métricas de sensibilidad y especificidad	* Se propusieron cinco MLC: árboles de inferencia condicional (CIT), árbol de modelo logístico (LMT), árbol de decisión C5.0, bosque aleatorio (RF) y refuerzo de gradiente extremo (XGBoost) * Se utilizó la regresión logística (LGR) como punto de referencia para la comparación	* Accuracy promedio 0.881 * Sensibilidad promedio 0.916 * AUC promedio 0.945

3. METODOLOGÍA

Para abordar el ciclo de vida de este proyecto de analítica se empleó la metodología CRIPS-DM (CRoss Industry Standard Process for Data Mining) Project (Wirth & Hipp, 2000). Su aplicación es un proceso estándar de la industria de minería de datos para transformar problemas comerciales de negocio en proyectos de análisis de datos que incluyen la elaboración de modelos. Esta metodología surgió en la década de los 90, a partir del término KDD (Knowledge Discovery in Databases) que significaba la extracción del conocimiento a partir de la búsqueda de patrones, anomalías y correlaciones que permitían predecir resultados con el objetivo de generar valor a la compañía ya sea incrementando las ventas o el número de clientes, reduciendo el riesgo o los costos de operación, entre otros.

CRISP-DM se describe en un modelo de proceso jerárquico compuesto de cuatro niveles de abstracción que van de lo general a lo específico, los cuales se visualizan en la figura 5 (Wirth & Hipp, 2000).

Figura 5. *Niveles de abstracción de la metodología CRISP-DM*



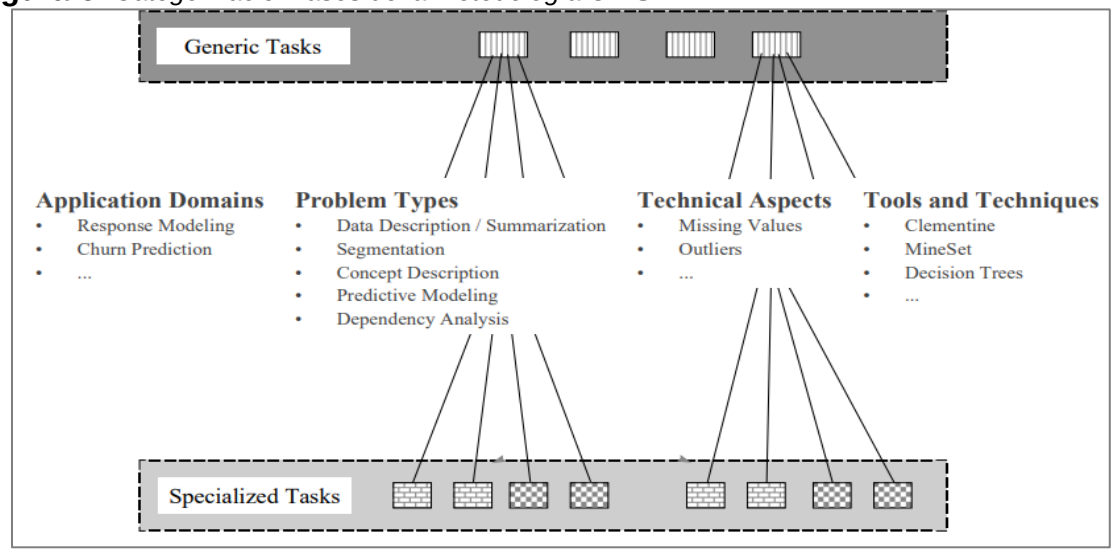
Nota: Tomado de (Wirth & Hipp, 2000).

A partir de la figura anterior, CRISP-DM se distingue por los modelos de referencia y la guía del usuario, el primero hace referencia a una descripción general de las fases, tareas y sus resultados, y responde al ¿Qué hacer en un proyecto de minería

de datos?; el segundo término es una lista detallada de sugerencias y consejos para cada una de las fases y tareas dentro de una fase, y describe ¿Cómo hacer el proyecto?

Cada una de las tareas genéricas que usa CRIPS-DM se dividen en cuatro categorías principales las cuales son: dominios de aplicación, tipos de problema, aspectos técnicos y herramientas y técnicas. Esto se presenta en la figura 6.

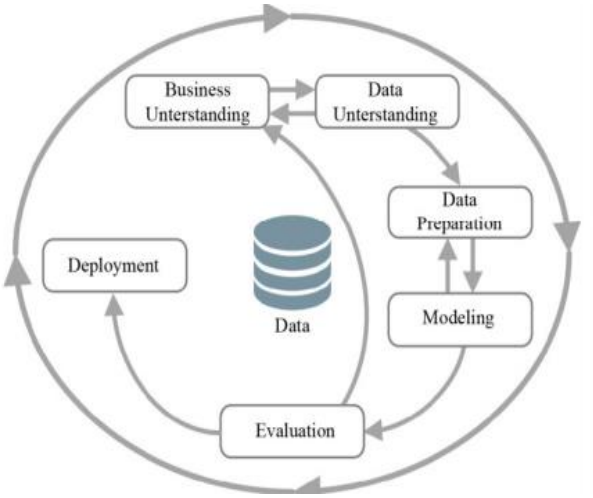
Figura 6. *Categorización fases de la metodología CRISP-DM*



Nota: Tomado de (Wirth & Hipp, 2000).

El modelo de referencia de CRISP-DM divide el proyecto en seis fases que se muestran en la figura 7. Cada fase contiene sus respectivas tareas y resultados, las cuales proporcionan una descripción general del ciclo de vida de un proyecto de analítica. Esta secuencia no es estricta, las flechas indican las dependencias más frecuentas e importantes de las fases, pero son flexibles acordes al resultado (Wirth & Hipp, 2000).

Figura 7. *Fases del proceso de CRISP-DM*



Nota: Tomado de (Wirth & Hipp, 2000).

El círculo externo que rodea todas las fases del ciclo de vida del proyecto simboliza la naturaleza cíclica propia de la minería de datos. A continuación, se describe brevemente cada fase:

Fase 1. Entendimiento del Negocio (Business Understanding): es probablemente la fase más importante dado que abarca las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva comercial, con el propósito de llevar ese entendimiento en una definición del problema desde la óptica de Data Mining y un plan de proyectos preliminar diseñado para lograr los objetivos trazados.

Principales tareas que componen esta fase: determinar los objetivos del negocio; evaluación de la situación; determinación de los objetivos del Data Mining; realizar el plan del proyecto.

Fase 2. Comprensión de los Datos (Data Understanding): se comienza con una recopilación inicial de datos para empezar un primer acercamiento con el problema, se emplean otras actividades para familiarizarse con los datos y la identificación de problemas de la calidad de estos, estableciendo subconjuntos o detectando

relaciones para formular las primeras hipótesis. De no contar con datos adecuados que permitan abordar el proyecto planteado, se deberá devolverse a la fase 1, colaborando con la empresa para determinar qué se necesita y verificar la integridad de estos.

Principales tareas que componen esta fase: recolección de datos iniciales; descripción de los datos; exploración de los datos; verificación de la calidad de los datos.

Fase 3. Preparación de los Datos (Data Preparation): una vez se efectúa la recolección de datos iniciales se procede a su preparación, la que comprende todas las actividades necesarias para construir el conjunto de datos finales que se utilizarán para las técnicas del modelado a partir de los datos sin procesar. En función de la técnica de modelado seleccionada se realiza a partir de un proceso de ETL (Extracción, Carga y Transformación) que conforma los datos que serán procesados por el algoritmo.

Principales tareas que componen esta fase: seleccionar los datos; limpiar los datos; estructurar los datos; integrar los datos; formateo de los datos.

Fase 4. Modelización (Modeling): una vez se ha logrado completar el proceso de ETL se seleccionan y prueban diferentes algoritmos de Machine Learning y Deep Learning, buscando adecuarlos a los datos recolectados. Regularmente en el proceso de modelado puede existir problemas con los datos y es necesario regresar a la fase anterior para ajustarlos. También, en esta fase, hay un proceso de hiperparametrización adaptativa que busca generar modelos auto-entrenables y escalables.

Principales tareas que componen esta fase: seleccionar la técnica del modelado; generación del plan de prueba; construcción del modelo; evaluación del modelo.

Fase 5. Evaluación (Evaluation): en este punto se han construido uno o más modelos que parecen cumplir con los criterios de éxito del problema. Antes de proceder a la implementación es necesario validar con la empresa que se estén cumpliendo las metas y objetivos trazados. Se debe evaluar nuevamente el proceso verificando cada una de las tareas realizadas con el fin de evitar errores en el proceso; así mismo, a la organización se le debe mostrar los resultados obtenidos y su potencial de forma tangible con el fin de cuantificar el impacto. Finalmente, se procede al despliegue del modelo.

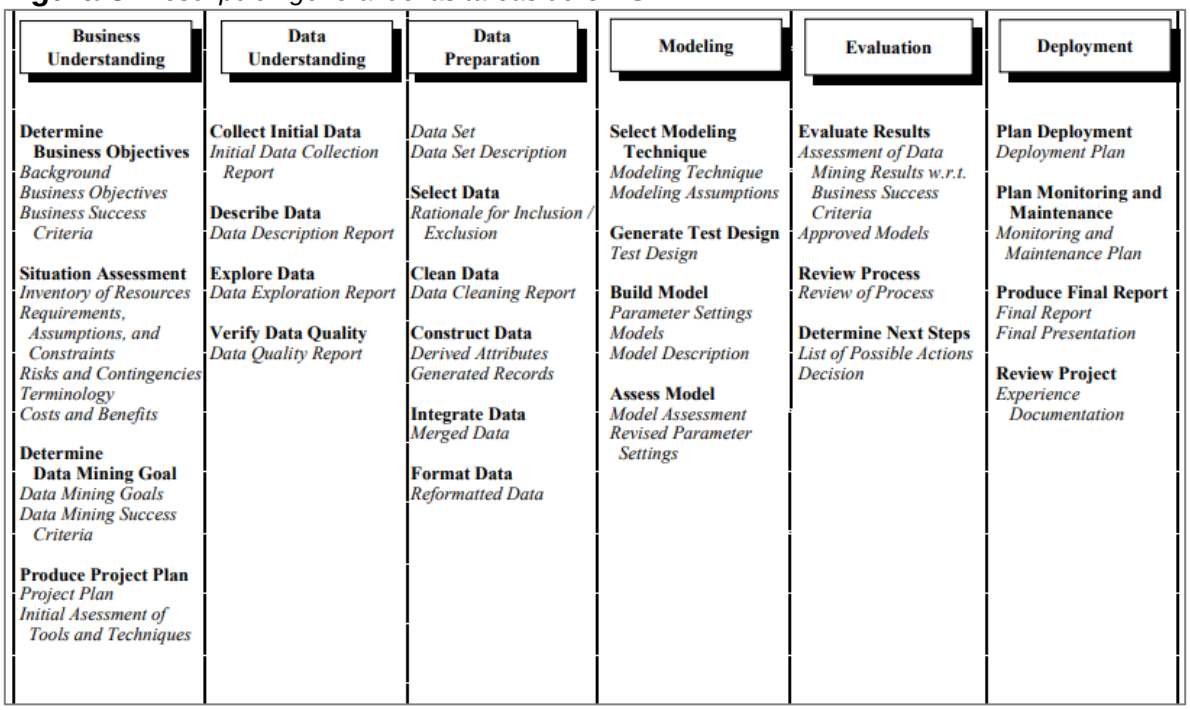
Principales tareas que componen esta fase: evaluar los resultados; revisar el proceso; determinar los próximos pasos.

Fase 6. Despliegue (Deployment): en este punto ya se ha construido y validado el modelo, para luego ser implementado en la organización. También, se debe presentar un documento y los resultados a las áreas de la compañía que intervinieron durante todo el proyecto, con el objetivo de lograr un incremento del conocimiento. La organización tendrá que conocer los pasos para aplicarlo, las acciones que deben llevarse a cabo, el mantenimiento de esta y la posible difusión de los resultados.

Principales tareas que componen esta fase: crear un plan de implementación; crear un plan de monitoreo y mantención; realizar el informe final; revisar el proyecto.

La figura 8 resume las tareas que componen las fases del CRISP-DM y los resultados que se deben obtener (Wirth & Hipp, 2000).

Figura 8. *Descripción general de las tareas de CRISP-DM*



Nota: Tomado de (Wirth & Hipp, 2000).

4. PRESENTACIÓN DE LA PROPUESTA

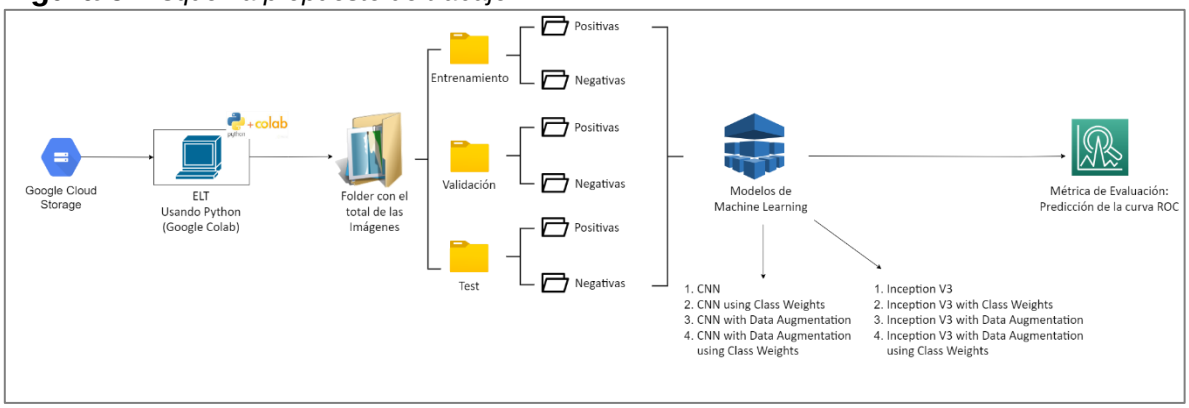
4.1 Entendimiento de los datos

Se utilizaron datos secundarios recolectados en un centro de oftalmología de la ciudad de Cali. Se contó con un archivo de datos que contenía 5.829 imágenes de pacientes del Valle del Cauca, correspondientes a 5.406 imágenes negativas y 423 imágenes positivas para glaucoma de personas que se realizaron un examen oftalmológico completo que incluyeron toma de agudeza visual, presión intraocular, gonioscopia, evaluación del segmento posterior que permite concluir la presencia o no de glaucoma; adicional, se contó con un archivo de Excel correspondiente a variables sociodemográficas e historial clínico de cada uno de los pacientes. Entre los resultados de las diferentes pruebas realizadas, se empleó como insumo principal una imagen del interior del ojo que permite evidenciar si el ángulo entre el iris y la córnea es estrecho (glaucoma de ángulo cerrado) o abierto (glaucoma de ángulo abierto).

Las imágenes estaban disponibles en un almacenamiento en nube de Google Cloud Storage, las que se cargaron mediante un script de Python usando Google Colab pro. Allí se almacenaron, categorizaron y posteriormente se dividieron en conjuntos de entrenamiento, validación y test para el preprocesamiento de los datos.

Finalmente, se analizaron los resultados de cada uno de los modelos tomando como referencia el mayor valor AUC ROC en el set de validación y test, para elegir el mejor modelo para este problema. La figura 9 muestra el esquema de la propuesta de este trabajo.

Figura 9. *Esquema propuesto de trabajo*



A continuación, se explican cada una de las actividades ejecutadas en cada una de las fases de la metodología CRISP-DM.

4.2 Preparación de los datos

Se dispone de un total de 5.829 imágenes del fondo del ojo de pacientes del Valle del Cauca, los participantes fueron seleccionados de un estudio transversal/seccional llevado a cabo por la Fundación Oftalmológica del Valle entre los años 2015 a 2018, el cual fue diseñado para evaluar la prevalencia de glaucoma y factores de riesgo vascular en la ciudad de Cali, Colombia. El reclutamiento y la metodología fueron aprobados por la Junta de Revisión Institucional de la Clínica Farallones (Cali, Colombia) que se adhirió a la Declaración de Helsinki y la Ley de Portabilidad y Responsabilidad del Seguro de Salud. Es importante aclarar que, para esta investigación se obtuvo el consentimiento informado de todos los participantes en el momento del reclutamiento.

El total de imágenes tiene un peso de almacenamiento de 28,19 Gigabytes, categorizadas en 5.406 imágenes de pacientes sin glaucoma (93%) y 423 imágenes de pacientes con glaucoma (7%). El total de imágenes se cargó en Google Colab pro, creando un directorio para cada una de las clases de las imágenes.

En este punto se planteó la siguiente pregunta: ¿cuál es la mejor manera de ajustar el desbalance entre clases? Dado que la mayor proporción es de pacientes sin glaucoma. Para dar solución, fue necesario realizar múltiples reuniones con los directores de la tesis, con el objetivo de identificar el problema y determinar cuáles serían los objetivos, alcance e impacto de la investigación.

En la revisión de la literatura se encontraron diferentes métodos de cambio en la distribución de clases, que se dividen en tres técnicas básicas: submuestreo (under-sampling) heurístico y no heurístico, sobre muestreo (oversampling) heurístico y no heurístico y muestreo avanzado (advanced sampling). Todas las estrategias fueron comparadas para aprender de conjuntos de datos desequilibrados y se concluyó que el submuestreo y el sobre muestreo son métodos muy efectivos para lidiar con el desequilibrio de clases (Japkowicz, 2000).

En el caso de este proyecto y como sucede en la mayoría de los trabajos que cuentan con datos desequilibrados, los casos minoritarios son de interés, y perder un caso minoritario cuesta más que perder un caso mayoritario; este fenómeno es evidente en el ámbito médico donde los pacientes de alto riesgo pertenecen a la clase minoritaria (Rahman & Davis, 2013). Por lo anterior, para este trabajo el submuestreo solo se aplica a la clase mayoritaria.

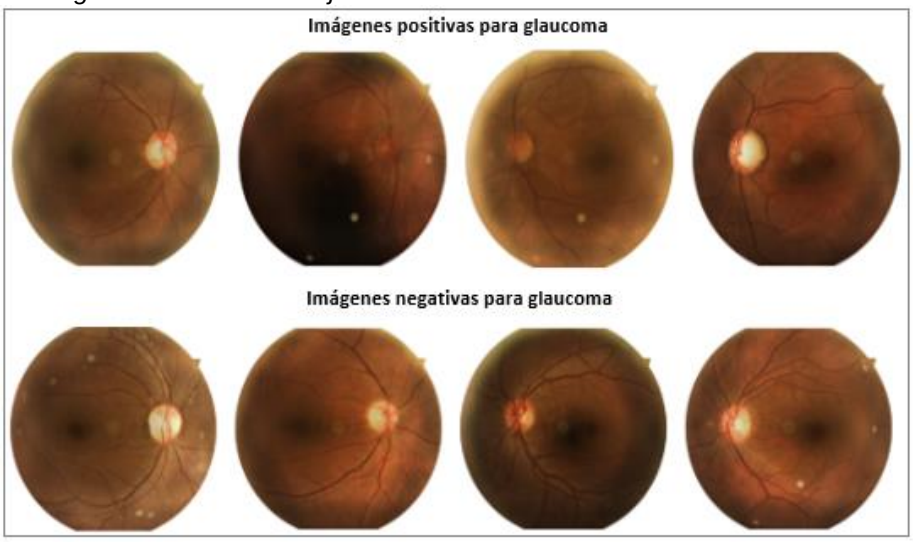
En estudios más recientes, se encontró el realizado por Lee y Bang, los cuales propusieron una técnica de clasificación binaria en las clases mayoritaria y minoritaria de datos estructurados desequilibrados. El modelo propuesto combina la técnica del submuestreo, el algoritmo de CNN y se compone de tres pasos. En el primer paso, se trata de crear un conjunto de entrenamiento equilibrado mediante submuestreo. Luego, cada ejemplo se convierte en una imagen que representa un gráfico de líneas. En el último paso, se diseña y entrena una CNN utilizando las

imágenes. Lo anterior, demuestra que la combinación de submuestreo y CNN es un enfoque viable para la clasificación de datos desequilibrados (Lee & Bang, 2021).

Una vez esto explorado y ya definido los objetivos del proyecto, en común acuerdo con los directores de tesis, se decidió que la mejor forma de ajustar el desbalance era utilizar un submuestreo no heurístico aleatorio, que intenta equilibrar las distribuciones de clase mediante la eliminación aleatoria de ejemplos de clases mayoritaria. Por lo tanto, se optó por mezclar y barajar las 5.406 imágenes de pacientes sin glaucoma y eliminar 3.714 imágenes de forma aleatoria, para un nuevo total de 1.692 imágenes (7,3 Gigabytes), y manteniendo las 423 imágenes (1,8 Gigabytes) de pacientes con glaucoma. De esta forma, la distribución pasó a ser 80% de imágenes con pacientes sin glaucoma y un 20% en pacientes con glaucoma dadas las recomendaciones de los directores de tesis y revisión de la literatura.

En la figura 10 se pueden visualizar algunas de las imágenes utilizadas para los sets de entrenamiento, validación y test.

Figura 10. *Imágenes del fondo del ojo*



Nota. Imágenes positivas: con presencia de glaucoma. Imágenes negativas: sin presencia de glaucoma.

Con un total de 2.115 imágenes de pacientes con y sin glaucoma, se procedió a dividir el conjunto de datos en imágenes para: entrenamiento, validación y test. Se usaron 1.480 imágenes (70%) para entrenar los modelos; 317 imágenes (15%) para validar los modelos; y 318 imágenes (15%) para testear los modelos en los nuevos conjuntos de imágenes.

4.3 Selección de los mejores modelos de clasificación

Una vez realizado el proceso de extracción, carga y transformación de los datos, se definieron los algoritmos para realizar la tarea de clasificación, CNN e Inception V3. Para cada uno de estos modelos se realizaron cuatro pruebas diferentes, para un total de 8 experimentos, con el objetivo de revisar distintas arquitecturas, parámetros e hiperparámetros. En la siguiente tabla se presenta los dos algoritmos empleados y los diferentes experimentos realizados.

Tabla 3. *Modelos y propuesta de experimentación*

Experimentos	Convolutional Neural Network (CNN)	Inception V3
Experimento 1	CNN tradicional	Inception V3 tradicional
Experimento 2	CNN usando Class Weights	Inception V3 usando Class Weights
Experimento 3	CNN usando Data Augmentation	Inception V3 usando Data Augmentation
Experimento 4	CNN aplicando Class Weights y Data Augmentation	Inception V3 aplicando Class Weights y Data Augmentation

El objetivo principal de usar CNN es por dos razones. La primera, es por los resultados en la literatura de diferentes modelos que se han desarrollado de aprendizaje profundo para la detección del glaucoma haciendo uso de las CNN. Algunos de los artículos de investigación que emplearon CNN para sus modelos fueron desarrollados por (Chen, Xu, Wong, Wong, & Liu, 2015) que propusieron y entrenaron desde cero una arquitectura CNN con seis capas: cuatro capas convolucionales y dos capas completamente conectadas, para clasificar automáticamente las imágenes del fondo del ojo glaucomatoso. Otra investigación fue realizada por (Alghamdi, Tang, Waheeb, & Peto, 2016) utilizando un nuevo

enfoque con dos CNN: una CNN para clasificar la región del disco óptico y la otra para clasificar una determinada región del disco óptico en tres clases: normales, sospechosas y anormales. Este algoritmo tiene la capacidad de aprender características con una marcada discriminación a partir de pixeles sin procesar y con resultados bastante buenos para la detección de esta patología.

La segunda razón, es por la arquitectura misma de los modelos de CNN, a partir de las primeras capas se extraen características generales del fondo del ojo, como bordes y ubicaciones generales de la imagen. Las capas intermedias detectan estructuras compuestas por características particulares de los bordes, y la última capa, detecta estructuras aún más complejas y específicas que son combinaciones de diferentes partes del fondo del ojo o partes de objetos familiares por el modelo. Por las anteriores razones esta investigación presenta una arquitectura para la detección del glaucoma basada en el aprendizaje profundo (Deep Learning) haciendo uso de la CNN.

Desde otra perspectiva está el hecho que entrenar una CNN desde cero implica una gran cantidad de datos etiquetados y disponer de recursos computacionales para su carga y procesamiento. En consecuencia, se hace necesario el explorar otras alternativas para no entrenar desde un principio una CNN y buscar modelos que sean más robustos contra el sobre aprendizaje (overfitting). Existen modelos para tareas de clasificación de imágenes médicas que han sido entrenados para este tipo problemas utilizando un gran conjunto de datos etiquetados de una aplicación diferente, como es el caso de ImageNet, un banco de datos para la investigación y desarrollo de software específico para el reconocimiento de imágenes.

Algunos trabajos investigativos (Dhungel, Carneiro, & Bradley, 2016) probaron que usando modelos CNN previamente entrenados con ImageNet fueron eficaces para aplicarlos en imágenes médicas, a pesar de las diferencias entre las imágenes. Entre los modelos de aprendizaje de transferencia (Transfer Learning) más

utilizados están el ResNet, VGG Family y el Inception. Este último modelo fue seleccionado por la baja complejidad de su arquitectura y rapidez computacional, siendo conveniente para el problema que se está tratando. A su vez, en un estudio realizado por Bianco et al. pretendían medir la eficiencia con la que cada modelo utiliza sus parámetros, midieron la precisión de cada modelo sobre el número de parámetros y encontraron que el Inception V3 ocupó el puesto número 14 entre 44 modelos evaluados (Bianco, Cadene, Celona, & Napoletano, 2018). Por ello se seleccionó este modelo para realizar diferentes experimentos en este estudio.

4.3.1.1 Modelos de CNN

Para las CNN se llevaron a cabo cuatro experimentos, todos se basaron en la misma arquitectura de red convolucional, pero con diferentes técnicas para combatir el sobreajuste. Las imágenes a color previamente fueron preetiquetadas y clasificadas, contaban con una dimensión de 2118 pixeles de ancho por 1944 pixeles de alto, con una profundidad en bits de 24. El tamaño de todas las imágenes se convirtió a 200 pixeles de ancho por 200 pixeles de alto, esto con el fin que sea más fácil de procesar por la red dado que las 2115 imágenes tienen un peso de 9,1 Gigabytes, de esta forma se reduce el tiempo y proceso computacional sin perder las características críticas para una buena predicción.

En la codificación de la arquitectura se usaron 3 módulos: convolution + relu + maxpooling. Las convoluciones operan en ventanas de 3x3 y las capas maxpooling operan en ventanas de 2x2. La primera convolución extrae 16 filtros, la siguiente extrae 32 filtros y la última extrae 64 filtros. En la revisión de literatura se encontraron diferentes artículos que respaldan y usan esta configuración de arquitectura porque sirven para un bajo número de imágenes de entrenamiento (alrededor de 1.000) y el uso de solo tres módulos convolucionales reducen el riesgo del sobreajuste y aseguran mantener un modelo pequeño.

En la parte final de la red se colocan dos capas completamente conectadas (two fully-connected layers), porque es un problema de clasificación binaria. Finalmente, se termina la arquitectura de la red con una activación sigmoidea, con la finalidad de tener la salida de la red con un escalar entre 0 y 1, codificando la probabilidad que la imagen es de clase 1 (a diferencia de la clase 0).

La columna "forma de salida" de la arquitectura presenta una evolución en el tamaño del mapa de características en cada capa sucesiva. Las capas de convolución reducen un poco el tamaño de los mapas de características debido al padding, y cada capa de agrupación reduce a la mitad el mapa de características.

En cuanto a la configuración de compilación para el entrenamiento del modelo, se utilizó la pérdida de entropía cruzada binaria (binary cross-entropy loss), porque como se dijo antes, es un problema de dos clases y la activación final usada es una sigmoide. El optimizador que se aplicó fue el RMSprop porque automatiza el ajuste de la tasa de aprendizaje (learning rate) que para el ejercicio fue de 0,001.

Para el preprocesamiento, se configuraron tres generadores de datos para las imágenes de entrenamiento, validación y test, cada uno con un batch size de 32, tamaño de los pixeles 200*200, etiquetado binario y con una paciencia de 20. Estos generadores son los encargados de leer las imágenes desde las carpetas de origen (train, validation y test) ya creadas y convirtiéndolas en tensores float32, siendo el insumo para la red. Los datos que ingresaron a la red han sido normalizados para un procesamiento más rápido, así todos los valores están en el rango entre 0 y 1, cuando originalmente los valores están en un rango de 0 a 255.

En los cuatro experimentos se entrenaron 1.480 imágenes (70% del total de imágenes), posterior, se validaron en 317 imágenes y se testearon en 318 imágenes.

Experimento 1: CNN tradicional

El primer experimento consistió en analizar el comportamiento de la red convolucional tradicional. Se utilizó el optimizador RMSprop, el learning rate fue de 0,001, el tamaño del lote se estableció en 32 y con una paciencia de 20. Todos estos hiperparámetros se eligieron de manera óptima para obtener el mejor rendimiento.

Experimento 2: CNN tradicional usando Class Weights

Para este segundo experimento se usó la misma arquitectura de la red convolucional (experimento 1) pero se decidió usar los pesos de las clases (class weights) dado que el conjunto de datos está desbalanceado y se requería buscar una técnica que permitiera mejorar los resultados de clasificación del modelo. Con esta técnica los class weights dan a las dos clases la misma importancia en cada actualización del gradiente, en promedio, independientemente del número de muestras que se tengan en los datos de entrenamiento. De esta forma se evita que el modelo prediga con una mayor frecuencia la clase más frecuente solo porque es más común.

La fórmula utilizada es:

$$class\_weight = \{0: 1692 \,/\, (1692 + 423), 1: 423 \,/\, (1692 + 423)\}$$

El 1692 indica el número total de imágenes etiquetadas negativas (sin glaucoma) y el 423 es el número total de imágenes etiquetadas positivas (con glaucoma).

Experimento 3: CNN tradicional usando Data Augmentation

En este tercer experimento nuevamente se empleó la misma red, pero con una leve modificación en su arquitectura, colocando entre el intermedio de las dos capas completamente conectadas una tasa de dropout de 0,5, lo que elimina aleatoriamente neuronas de la red durante el entrenamiento, por lo cual otras neuronas deberán intervenir y manejar la representación requerida para hacer predicciones para las neuronas faltantes. El efecto de aplicar esta regularización hace que la red se vuelva menos sensible a los pesos específicos de las neuronas,

capaz de aprender múltiples representaciones internas independientes, lo que se traduce en una mejor generalización y una menor probabilidad de caer en sobreajuste.

Así mismo, se aplicó la técnica de aumento de los datos (data augmentation) por medio de transformaciones que generan nuevas imágenes artificiales que permiten entrenar un modelo más robusto y menos propenso al sobre aprendizaje. En este punto se sostuvo múltiples reuniones con los directores de la tesis, dado que no todos los parámetros que permiten ajustar la imagen en rotación, ancho y alto, escala y recorte aplicaban en el problema de imágenes del fondo del ojo.

Después de analizar e investigar se seleccionaron los siguientes parámetros para aprovechar al máximo el número reducido de imágenes de entrenamiento:

- *Rotation range = 10*, es un rango dentro del cual se rota la imagen aleatoriamente, va de 0 a 180 grados.
- *horizontal_flip = True*, voltea aleatoriamente la mitad de las imágenes horizontalmente. Esto es relevante cuando no hay supuestos de asimetría horizontal (por ejemplo, imágenes del mundo real).
- *fill_mode = "nearest"*, estrategia utilizada para rellenar píxeles recién creados, que pueden aparecer después de una rotación o un cambio de ancho/alto.

Experimento 4: CNN tradicional usando Data Augmentation y Class Weights

Para el cuarto experimento se emplearon las técnicas de data augmentation y class weights en el modelo de CNN. La idea con este ejercicio era probar que también podría comportarse el rendimiento del modelo usando ambos métodos juntos, dado que se habían probado en experimentos anteriores por separado. La configuración de la red convolucional fue la misma que la del experimento tres. Se aplicó la misma

fórmula para los pesos de clase y se usaron los mismos parámetros para la técnica de data augmentation.

4.3.1.2 Modelos de Transfer Learning

Usando el modelo Inception V3 desarrollado en Google y entrenado en ImageNet, el cual es un conjunto de 1.4 millones de imágenes web para 1.000 clases diferentes, se realizaron cuatro diferentes experimentos.

Antes de desarrollar cada una de las pruebas, se configuró la arquitectura de la red y los pasos para su ejecución. Lo primero fue crear una instancia del modelo Inception V3 precargado con pesos entrenados en ImageNet y como segundo paso, era hacer que el modelo no se pudiera entrenar, dado que solo se usaría para la extracción de características, con el objetivo de no actualizar los pesos del modelo preentrenado durante el entrenamiento (weights=None).

Para la elección de la capa intermedia de Inception V3 se seleccionó “mixed7”, esta capa no es el cuello de botella de la red (bottleneck of the network) pero se eligió dado que se necesitaba mantener un mapa de características lo suficientemente grande, en este caso de 7*7. Caso contrario, de haberse usado la capa de cuello de botella de la red, se hubiera obtenido un mapa de características de 3*3, demasiado especializado para el problema. Finalmente, se colocó un clasificador totalmente conectado (fully connected classifier) encima de la última salida.

Cada uno de los cuatro experimentos que se realizaron utilizaron el mismo proceso de extracción de características mediante un modelo preentrenado, con la diferencia que para cada uno se aplicaron diferentes técnicas para reducir el sobreajuste en el entrenamiento.

En relación con la arquitectura de la red, se aplanó la capa de salida a una dimensión, se agregó una capa completamente conectada con 1.024 unidades

ocultas y activación de ReLU, una tasa de dropout de 0,2 y finalmente una capa sigmoide para la clasificación.

Referente a la compilación para el entrenamiento del modelo, se utilizó la pérdida de entropía cruzada binaria (binary cross-entropy loss). El optimizador que se aplicó fue el descenso de gradiente estocástico (SGD), con una tasa de aprendizaje (learning rate) muy baja de 0,00001. Para usar el Inception V3, se realizó una suscripción paga a Colab Pro, con el fin de contar con suficientes recursos computacionales para utilizarlo y usando como hiperparámetro un momentum de 0,9, esto aceleró las direcciones del descenso de gradiente a diferencia del SGD simple.

Para el preprocesamiento, al igual que con los modelos de CNN, se configuraron tres generadores de datos para las imágenes de entrenamiento, validación y test, cada uno configurado con un batch size de 32, tamaño de los pixeles 200*200, una paciencia de 20 y los datos que ingresaron a la red fueron normalizados.

Experimento 1: Inception V3

El primer experimento consistió en reproducir los pasos del Inception V3 previamente explicados. El objetivo era tener un modelo base para revisar su comportamiento general. Este sería el modelo Baseline para posteriormente probar las diferentes técnicas contra el sobreajuste que se usarían en los experimentos posteriores.

Experimento 2: Inception V3 usando Class Weights

Para este segundo experimento se usó la misma arquitectura de la red Inception V3 del primer experimento, pero utilizando ahora los pesos de las clases (class weights).

La fórmula utilizada es:

$$class\_weight = \{0: 1692 / (1692 + 423), 1: 423 / (1692 + 423)\}$$

El 1692 indica el número total de imágenes etiquetadas negativas (sin glaucoma) y el 423 es el número total de imágenes etiquetadas positivas (con glaucoma).

Experimento 3: Inception V3 usando Data Augmentation

En este tercer experimento nuevamente se empleó la misma red y se aplicó la técnica de aumento de los datos (data augmentation). Tal como se mencionó anteriormente, para definir los parámetros del nuevo conjunto de imágenes se realizaron reuniones con los directores de la tesis.

Los parámetros utilizados fueron:

- *Rotation range = 10*, es un rango dentro del cual se rota la imagen aleatoriamente, va de 0 a 180 grados.
- *horizontal_flip = True*, voltea aleatoriamente la mitad de las imágenes horizontalmente. Esto es relevante cuando no hay supuestos de asimetría horizontal (por ejemplo, imágenes del mundo real).
- *fill_mode = "nearest"*, estrategia utilizada para rellenar píxeles recién creados, que pueden aparecer después de una rotación o un cambio de ancho/alto.

Experimento 4: Inception V3 usando Data Augmentation y Class Weights

Para el cuarto y último experimento se emplearon las técnicas de data augmentation y class weights en el modelo de Inception V3. Se pretendía con esto aprovechar las ventajas de entrenar un conjunto de datos grande y aplicar al mismo tiempo estas técnicas para comprobar si el rendimiento del modelo podría mejorar usando ambos métodos juntos y no por separado. Para ambas técnicas se usaron los mismos parámetros e hiperparámetros descritos en el modelo de Inception V3.

4.4 Métricas de clasificación

La curva ROC se usa con frecuencia para evaluar el desempeño de algoritmos de clasificación binaria. También, las medidas intrínsecas como la sensibilidad (recall o sensitivity), la especificidad (specificity), la correctitud (accuracy) y el F1-score son indispensables para dar un resultado de la capacidad de un modelo para clasificar a un individuo en uno de los grupos (poblaciones).

La sensibilidad es una estimación de la eficacia del experimento para predecir una enfermedad. La especificidad evalúa la probabilidad de que los pacientes sin enfermedad puedan identificarse correctamente. La curva ROC es una representación gráfica de *1 – especificidad* (eje X) y *sensibilidad* (eje Y). Un valor de AUC = 0.5 es un clasificador aleatorio (sin poder predictivo) y un AUC = 1.0 es un clasificador perfecto, esto significa una tasa de verdaderos positivos del 100 % y una tasa de falsos positivos del 0%. El objetivo por lo general de los problemas de clasificación binarios es encontrar un valor del AUC que esté entre 0.5 y 1.0.

La curva ROC y las puntuaciones AUC también permiten comparar el rendimiento de diferentes clasificadores para el mismo problema, lo cual es útil para los ocho experimentos que se ejecutaron. Por lo tanto, el criterio usado como métrica para este problema y ampliamente utilizado en la literatura para medir el desempeño de un modelo es el área bajo la curva ROC (AUC).

Para calcular la curva ROC, se necesita variar el umbral de probabilidad (threshold probability) que usa el clasificador para predecir si un paciente tiene glaucoma (objetivo = 1) o no (objetivo = 0). El hallar este umbral, no es asignar una etiqueta a una determinada clase, sino determinar la probabilidad que una observación pertenece a una clase en específico, esto es una ventaja dado que permite adaptar el modelo al problema de clasificación.

Al evaluar un clasificador de pacientes con glaucoma se debe prestar especial atención al error tipo II, es decir, cuando la persona está enferma pero la prueba predice de forma incorrecta que no presenta la enfermedad (falso negativo). En el contexto del problema un error de este tipo sería mucho más grave que un error tipo I, donde la persona no está enferma pero la prueba predice de manera incorrecta que si presenta la enfermedad (falso positivo). En otras palabras, la tarea es encontrar una tasa de falsos negativos lo más baja posible.

Complementado la métrica del ROC AUC se utiliza el indicador F1-score porque resume la precisión y sensibilidad en una sola. Esto es ideal cuando la distribución de las clases es desbalanceada y a su vez es un indicador muy utilizado en el campo de la salud (Provost & Fawcett, 1997).

5. DISEÑO DE EXPERIMENTO DE VALIDACIÓN

Para el modelo de CNN e Inception V3 y los 4 diferentes experimentos entrenados, se ejecutaron en Google Colab Pro, cuyo entorno provee una GPU NVIDIA-SMI 460.32.03.

Para realizar un seguimiento de las métricas: AUC, validation AUC (val_auc), loss y validation loss (val_loss) y examinar su comportamiento, tiempo de duración y número de épocas, se utilizó la plataforma de Machine Learning Weights and Biases, esta permite administrar el ciclo de vida de los modelos, datos y experimentos de Machine Learning.

El criterio de selección de modelos se basa en obtener el resultado más alto en los sets de validation y test, buscando a su vez el menor sobreajuste posible.

6. RESULTADOS OBTENIDOS

6.1 Análisis descriptivo

En total se seleccionaron 2.115 imágenes, 423 (20%) con glaucoma (positivo) y 1.692 (80%) sin glaucoma (negativo). El 67% fueron de sexo femenino, con edades entre los 18 y 101 años y la mayor concentración se encontraba en el rango de 61 a 71 años (29.5%). El 77.4% de los pacientes reportaron estar casados y tan solo el 2.9% eran solteros al momento del estudio. La etnia predominante fue mestizo con el 62%. Para el nivel educativo de los pacientes, el 23.9% reportó haber cursado la educación primaria completa y el 21.7% el nivel educativo de primaria incompleta; la mayor cantidad de pacientes pertenecían a estrato socioeconómico nivel 2 con un 49.7%, seguido del nivel 3 con un 32%. Con relación a la zona de ubicación de la vivienda, el 99.97% residían en zona urbana y el 0.03% en zona rural. Esta información se puede visualizar en las figuras 11 y 12.

Figura 11. *Características sociodemográficas de la población (Sexo, Zona, Estado Civil y Etnia/Raza)*

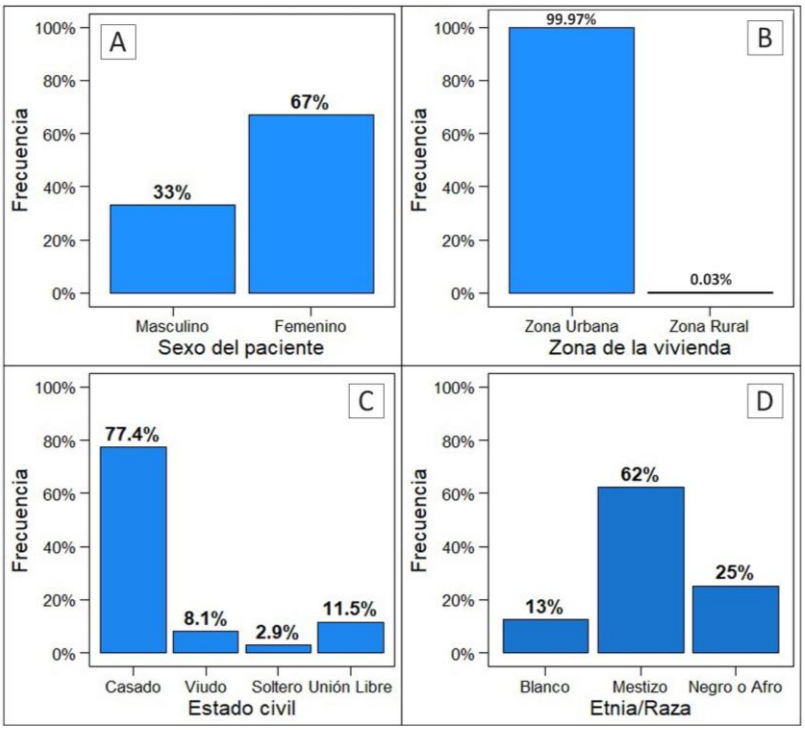
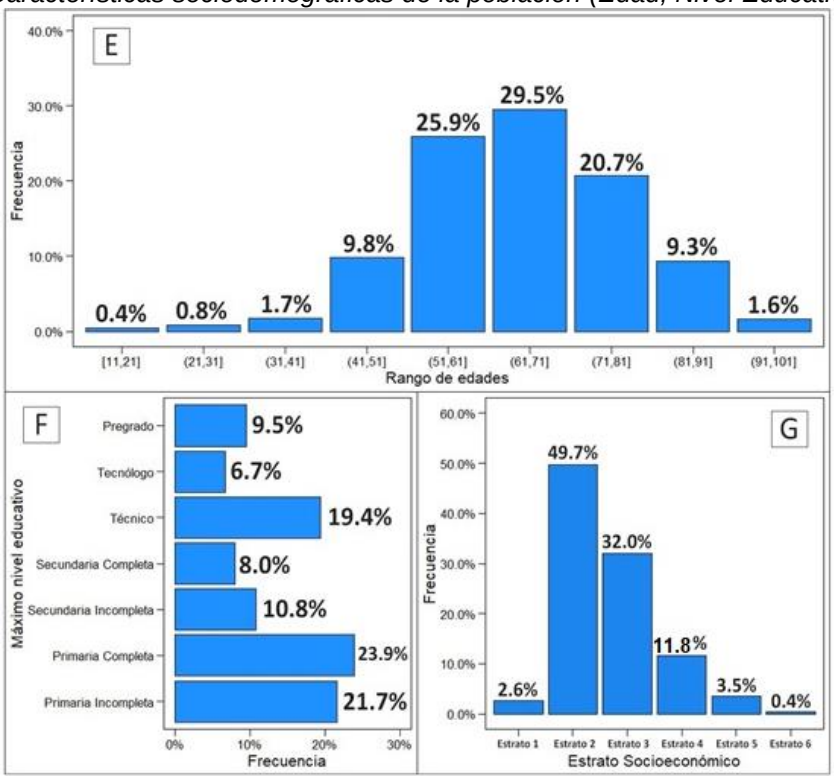


Figura 12. *Características sociodemográficas de la población (Edad, Nivel Educativo y Estrato)*



Antecedentes clínicos de los pacientes

En la tabla 4 se presentan los antecedentes clínicos de los pacientes, donde se aprecia que las enfermedades más frecuentes fueron hipertensión y dislipidemia, con valores de 31.2% y 25% respectivamente; por el contrario, las menos frecuentes son cáncer y enfermedades coronarias.

Tabla 4. *Antecedentes clínicos de los pacientes*

Variable	*Si*		*No*		*No sabe*	
	n	*%*	*n*	*%*	*n*	*%*
¿Tiene dislipidemia?	*528*	*25*	*1492*	*70,5*	*95*	*4,5*
¿Sufre de la tiroides?	*272*	*12,9*	*1764*	*83,4*	*79*	*3,7*
¿Sufre de migraña?	*256*	*12,1*	*1783*	*84,3*	*76*	*3,6*
¿Sufre de alguna enfermedad coronaria?	*156*	*7,4*	*1880*	*88,9*	*79*	*3,7*
¿Tiene o ha sufrido de cáncer?	*62*	*2,9*	*1977*	*93,5*	*76*	*3,6*
¿Sufre de hipertensión?	*660*	*31,2*	*1380*	*65,2*	*75*	*3,5*
¿Sufre de diabetes?	*211*	*10*	*1829*	*86,5*	*75*	*3,5*

Nota: n = 2.115

En la tabla 5 se presentan las variables que se tomaron en cuenta para el diagnóstico del glaucoma. A la pregunta si el paciente tenía algún familiar que hubiera tenido glaucoma (abuelos, padres, hermanos o tíos), el 26.6% respondió afirmativamente. A la pregunta, si el paciente sospechaba que tenía glaucoma, el 16.3% reportó que sí sospechaba que tenía la enfermedad mientras que el 83.7% respondió que no. Finalmente, para definir si un paciente tenía o no glaucoma, se seleccionaron aquellas personas que marcaron "Sí" al menos a una de las características que definen a este, ya sea "Glaucoma primario de ángulo abierto o cerrado" encontrando confirmación de pacientes con la afección en un 20% positivo y 80% negativo.

Tabla 5. *Variables relacionadas con el glaucoma en los pacientes*

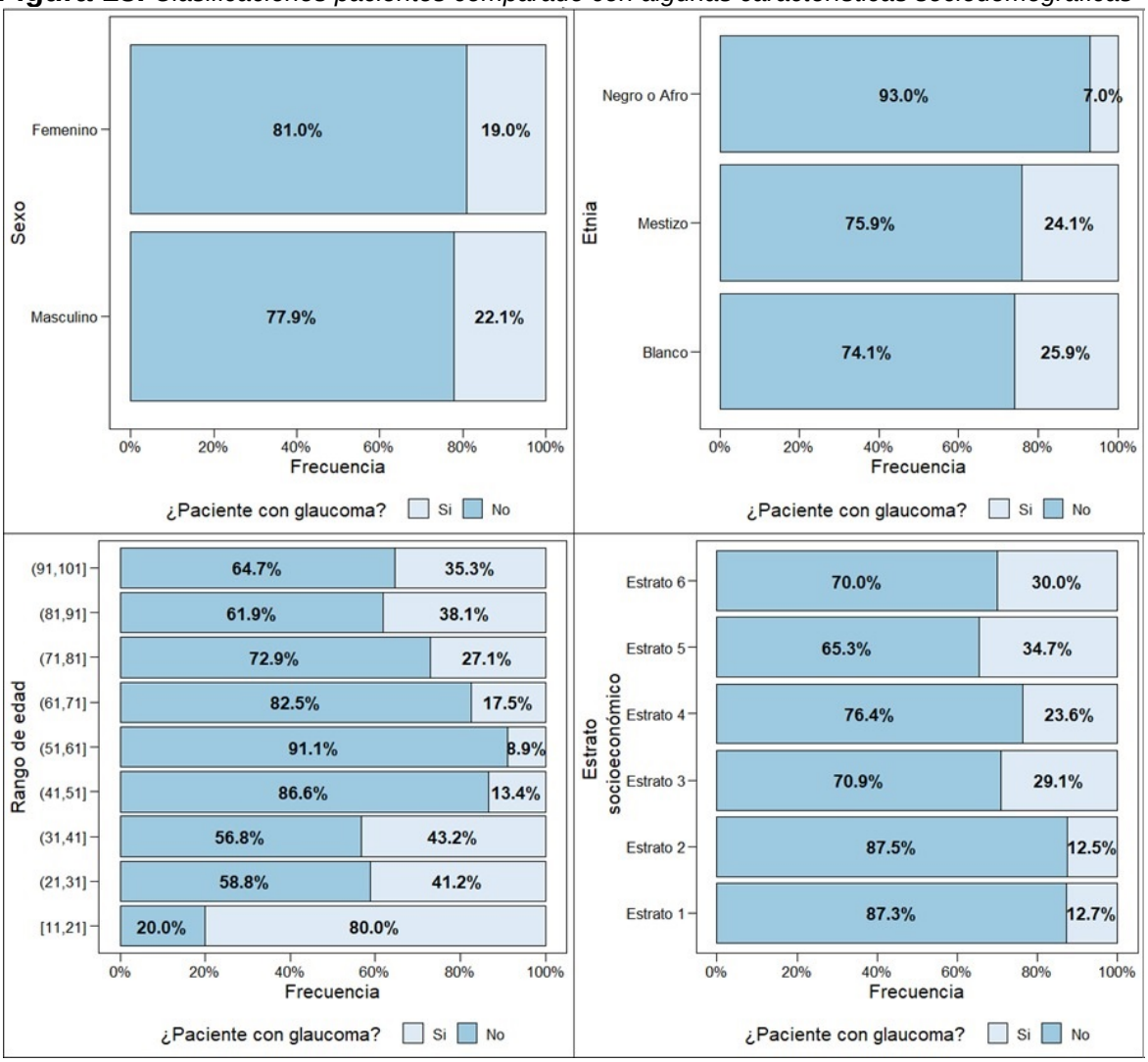
Variable	Si		No	
	n	%	n	%
¿Sabe si hay un antecedente familiar de glaucoma?	*563*	*26,6*	*1552*	*73,4*
¿Sospecha que tiene glaucoma?	*344*	*16,3*	*1771*	*83,7*
Tiene Glaucoma Primario de Ángulo Abierto	*363*	*17,2*	*1752*	*82,8*
Tiene Glaucoma Primario de Ángulo Cerrado	*60*	*2,8*	*2055*	*97,2*
Confirmado si tiene o no Glaucoma	*423*	*20*	*1692*	*80*

Nota: n = 2.115

Se observaron algunas características de los pacientes con la variable glaucoma (positivos y negativos) comparados con las variables sexo, etnia, edad y estrato socioeconómico. Se encontró que, del total de personas de sexo femenino (1.417) el 19% resultó positivo para glaucoma (269) comparado con el sexo masculino (698), con un porcentaje de positivos del 22.1% (154). Referente a la variable etnia, los mestizos tienen la mayor proporción de esta enfermedad, con un 24.1% equivalente a 317 personas; la etnia blanca tiene un 25,9% de casos positivos para glaucoma, equivalente a 69 personas; y la etnia negra o afro cuenta con un 7% de casos positivos, equivalente a 37 personas. Para el rango de edad, los pacientes en edades entre 71 a 81 años son quienes más tienen glaucoma con 27.1% (119) seguido del rango de edades entre 61 a 71 años con un 17.5% (109 pacientes). Por

último, la mayor concentración de pacientes positivos se encuentra en los estratos 2 y 3 con un 12.5% y 29.1% respectivamente, equivalentes a 132 y 197 pacientes glaucomatosos, esta relación se puede observar en la figura 13.

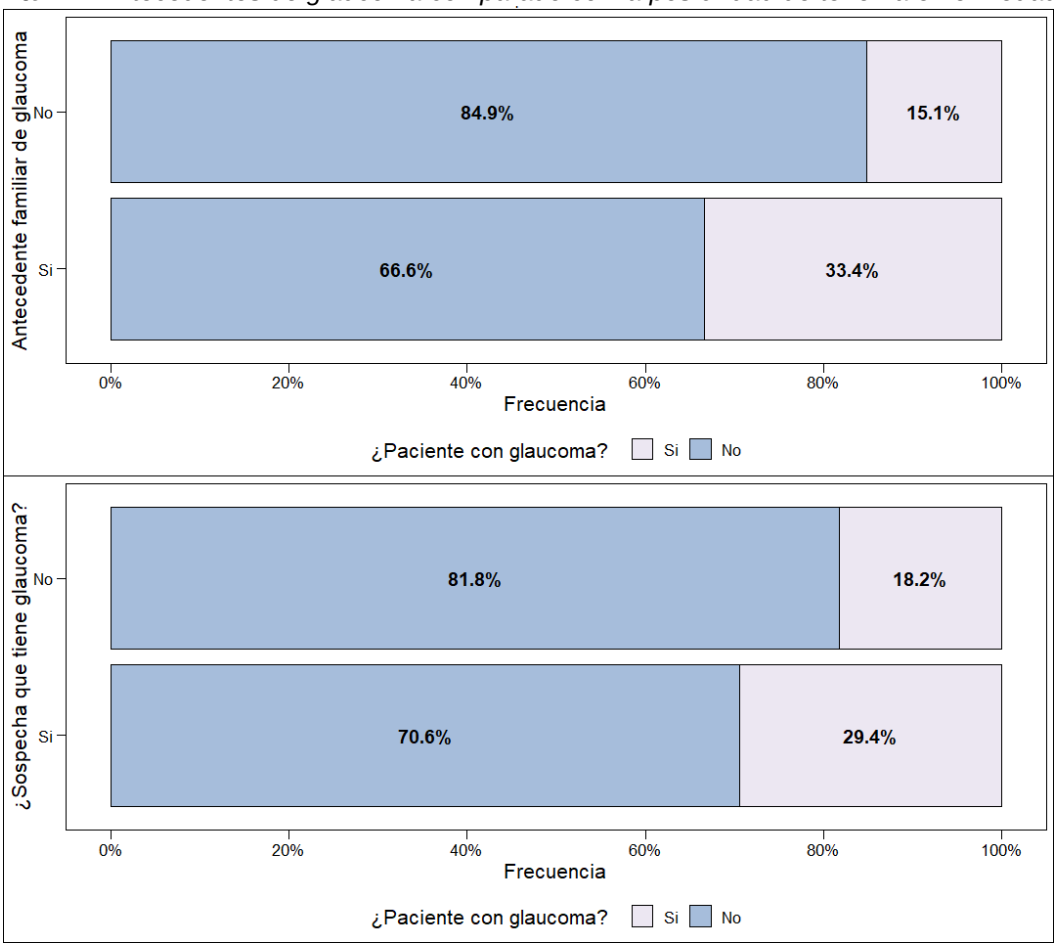
Figura 13. *Clasificaciones pacientes comparado con algunas características sociodemográficas*



Para evaluar el antecedente familiar de glaucoma a cada uno de los pacientes, se les preguntó si uno o más familiares, sea abuelos, padres, hermanos y/o tíos habían sufrido o sufren de esta enfermedad, el 26.6% del total de personas evaluadas afirmó tener uno o más familiares con antecedentes, de estos el 33.4% (189) fueron

positivos para la enfermedad, en comparación con el 73.4% que aseveró no tener un antecedente familiar de este tipo, sin embargo el 15.1% fue positivo para el glaucoma (234). Por otra parte, los pacientes que sospecharon tener glaucoma (344) solo el 29.4% tienen la enfermedad, equivalente a 101 pacientes, mientras que la cantidad de pacientes que sospechan no tener la enfermedad (83.7%), realmente de estos, el 18.2% son positivos para la enfermedad, con un total de 322 pacientes. Ver figura 14.

Figura 14. *Antecedentes de glaucoma comparado con la posibilidad de tener la enfermedad*



6.2 Comparación de experimentos

Para el modelo de CNN y los 4 diferentes experimentos entrenados se esperaba que los resultados en training estuvieran muy sobreajustados dado el bajo número de imágenes para entrenamiento (1.480 imágenes). Así mismo, se presupuestaba que la técnica del data augmentation serviría para reducir el sobreajuste en los resultados de entrenamiento, esto se evidenció en los experimentos tres y cuatro que arrojaron las menores tasas en el set de training. Por otra parte, a diferencia de lo que se esperaba al usar la técnica de class weights, su resultado marco un claro overfitting del 100%, lo que deja demostrado para este ejercicio que el algoritmo de CNN junto con este método no da buenos resultados. Finalmente, el algoritmo de CNN aplicando ambas técnicas permite controlar el sobreajuste y dar los mejores resultados en validation y test. Ver tabla 6.

Tabla 6. *Métricas de evaluación AUC para modelos CNN*

Convolutional Neural Network (CNN)		*Resultados*		
		Training	*Validation*	*Test*
Experimento 1	*CNN tradicional*	*0.9993*	*0.7528*	*0.8163*
Experimento 2	*CNN usando Class Weights*	*1.0*	*0.7920*	*0.7459*
Experimento 3	*CNN usando Data Augmentation*	*0.9305*	*0.8008*	*0.8299*
Experimento 4	*CNN aplicando Class Weights y Data Augmentation*	*0.9581*	*0.8442*	*0.8597*

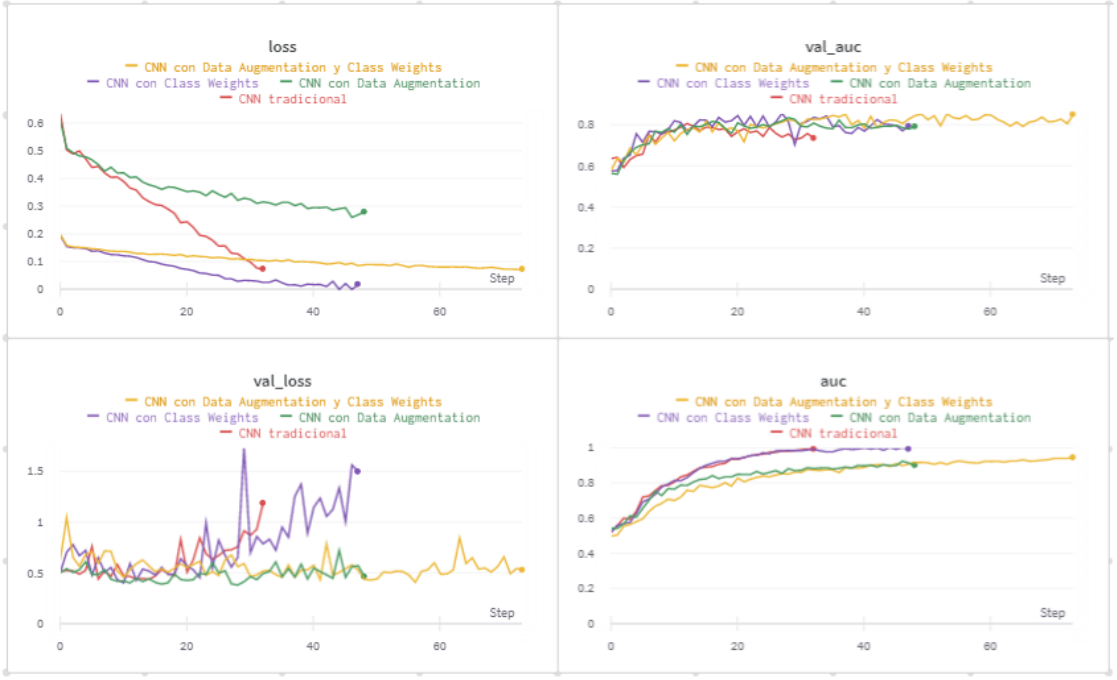
A continuación, en la tabla 7 se presentan los resultados de los cuatro experimentos de CNN para las métricas de evaluación, el tiempo de duración, el número de épocas, el AUC y validation AUC (val_auc). Es claro como las técnicas de data augmentation y class weights incrementan el tiempo de duración de los modelos, así como el número de épocas.

Tabla 7. *Características de los experimentos de CNN*

	Experimento 1	*Experimento 2*	*Experimento 3*	*Experimento 4*
Duración	*1h 32m 33s*	*2h 12m 4s*	*2h 23m 3s*	*4h 32m 42s*
Tipo de GPU	*Tesla P100-PCIE-16GB*	*Tesla P100-PCIE-16GB*	*Tesla P100-PCIE-16GB*	*Tesla T4*
Épocas	*33*	*48*	*49*	*74*
auc	*0.9993*	*1,0*	*0,9305*	*0,9581*
val_auc	*0,7528*	*0.7920*	*0,8008*	*0,8442*

Para realizar un seguimiento de las métricas y examinar su comportamiento se utilizó la plataforma de Machine Learning, Weights and Biases. En la figura 15 se presenta un resumen de las métricas: loss, val_auc, val_loss, y auc.

Figura 15. *Resultados de las métricas de evaluación para los cuatro experimentos de CNN*



A partir de la figura 15 se muestra como el loss (pérdida de entrenamiento) que es el promedio de los losses sobre cada batch de entrenamiento, disminuye abruptamente en el modelo de CNN tradicional y CNN con Data Augmentation. Esto es un claro indicio de la baja precisión del modelo para modelar la relación de los datos de entrada y los objetivos de salida. Por el contrario, el modelo de CNN con

Data Augmentation y Class Weights tiene una reducción del loss muy baja y constante a lo largo de las épocas entrenadas, esto en gran medida puede deberse a la técnica de regularización del dropout, que ayuda al modelo a obtener una mayor precisión en la validación y test sacrificando la precisión del entrenamiento.

Revisando el val_loss (perdida en validación) para el modelo CNN tradicional y CNN con Class Weights este crece exponencialmente a partir de la época 20, esto significa que el modelo está acumulando valores y no está aprendiendo, un claro ejemplo de sobre aprendizaje. A diferencia del modelo de CNN con Data Augmentation y Class Weights, donde el val_loss comienza a disminuir y mantener un comportamiento con bajas variaciones entre la época 20 y 60 y el val_auc comienza a aumentar y mantener un desempeño casi que constante entre la época 40 y 74. Lo anterior, significa que el modelo construido está aprendiendo y funcionando bien.

A diferencia de los experimentos con el algoritmo de CNN, para los cuatro experimentos usando el modelo Inception V3 se esperaba encontrar un menor sobreajuste debido al método de Transfer Learning que permite tener un modelo de clasificación de imágenes sin un proceso de aprendizaje desde cero. La tabla 8 resume los resultados obtenidos.

Tabla 8. *Métricas de evaluación AUC para modelos Inception V3*

Inception V3		*Resultados*		
		Training	*Validation*	*Test*
Experimento 1	*Inception V3*	*0.9236*	*0.8706*	*0.9084*
Experimento 2	*Inception V3 usando Class Weights*	*0.8639*	*0.8228*	*0.8654*
Experimento 3	*Inception V3 usando Data Augmentation*	*0.8899*	*0.8630*	*0.8612*
Experimento 4	*Inception V3 aplicando Class Weights y Data Augmentation*	*0.8179*	*0.8728*	*0.8586*

Después de haber entrenado, validado y testeado los cuatro modelos, los resultados estuvieron acorde a lo que se esperaba en un principio. Al usar juntas las técnicas de data augmentation y class weights los datos arrojados en training fueron los más bajos, sinónimo de una mejoría en cuanto al sobreajuste; obtuvo el validation más

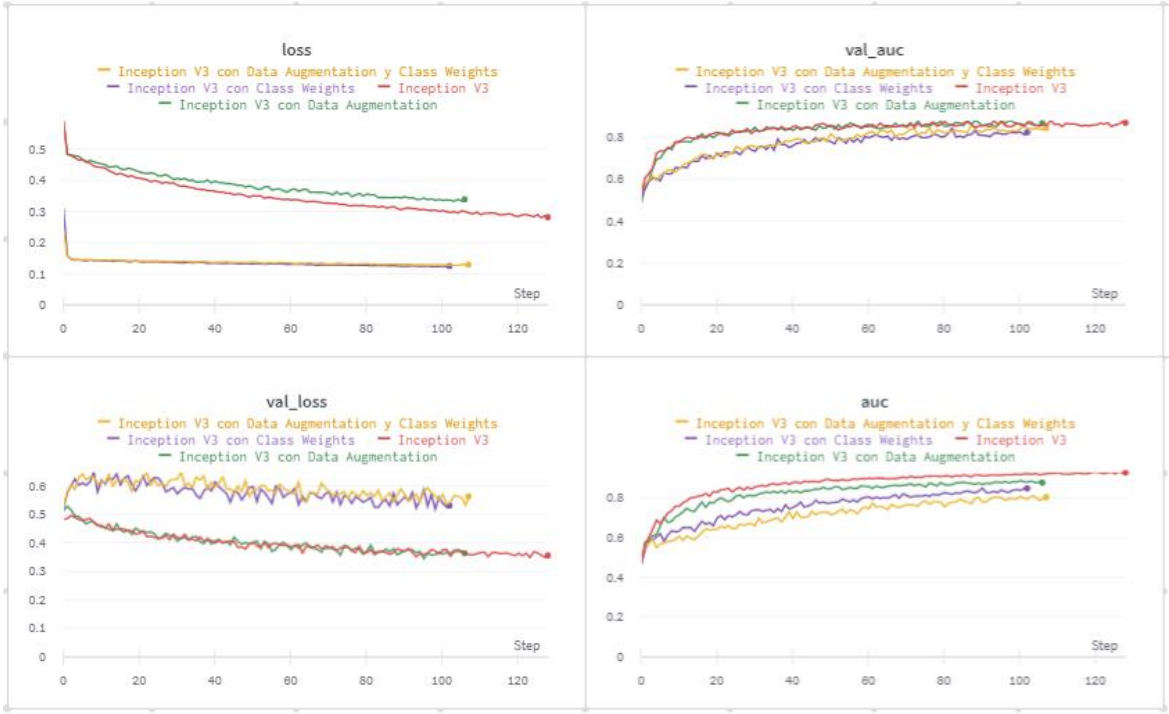
alto pero el test más bajo de los cuatro experimentos. El Inception V3 por su parte, alcanzó el valor más alto en training, el segundo valor más alto en validation y el mayor valor en test. En la tabla 9 se presentan los resultados obtenidos para las métricas de evaluación, el tiempo de duración, el número de épocas, el AUC y validation AUC (val_auc).

Tabla 9. *Características de los experimentos de Inception V3*

	Experimento 1	*Experimento 2*	*Experimento 3*	*Experimento 4*
Duración	*8h 45m 3s*	*5h 22m 32s*	*5h 59m 8s*	*7h 56m 28s*
Tipo de GPU	*Tesla T4*	*Tesla T4*	*Tesla T4*	*Tesla T4*
Épocas	*129*	*103*	*107*	*108*
auc	*0.9236*	*0.8639*	*0.8899*	*0.8179*
val_auc	*0.8706*	*0.8228*	*0.8630*	*0.8728*

De igual manera que con los modelos de CNN, se usó Weights and Biases para realizar un seguimiento y medición de las métricas evaluadas. Ver figura 16.

Figura 16. *Resultados de las métricas de evaluación para los cuatro experimentos de Inception V3*

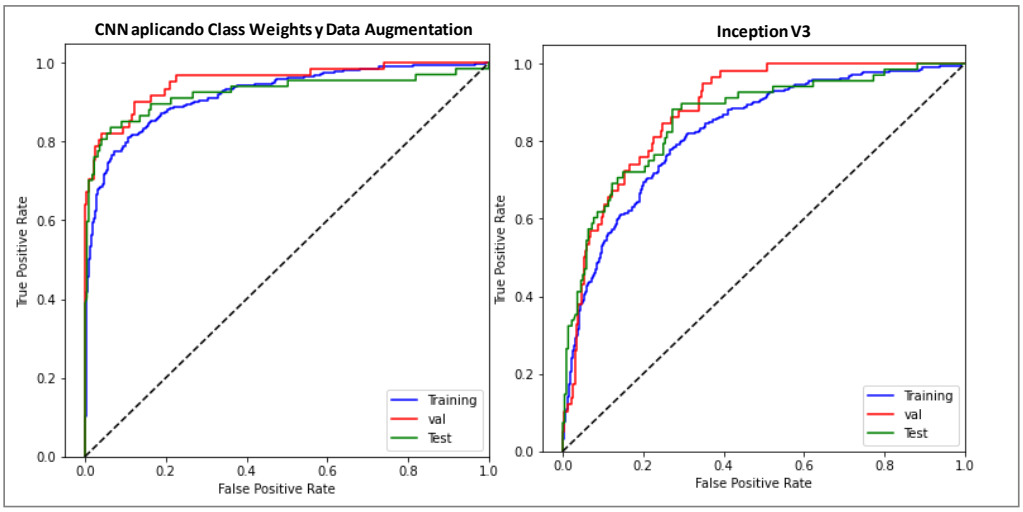


En términos generales los cuatro modelos probados tienen comportamientos bastantes similares y se nota una mejoría si se compara con los modelos de CNN, en este caso, el val_loss decrece a tasas muy bajas a medida que el número de épocas aumenta y el val_auc por el contrario aumenta paulatinamente al igual que el auc. Todos los modelos tienden a estabilizarse en la época 50, con un aumento paulatino y lento del val_auc y una disminución constante y lenta del val_loss.

6.3 Discusión

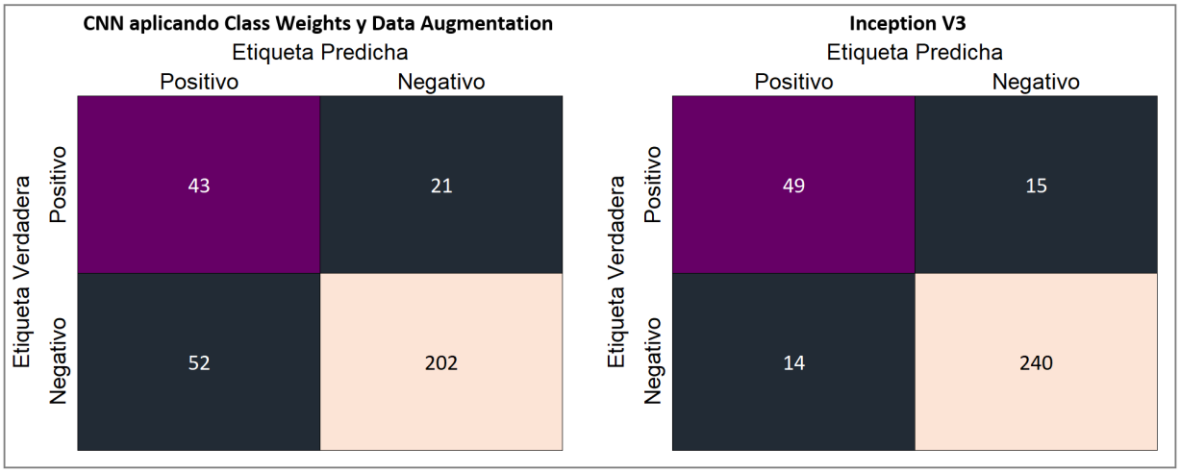
Luego de validar los ocho experimentos y haber revisado diferentes métricas y evaluado su comportamiento en lo datos de training, validation y test, es claro que para el modelo de CNN e Inception V3 los mejores resultados se lograron con los experimentos números 4 y 1 respectivamente. En ambos casos se llegó a esta decisión por contar con resultados en validation y test más altos. En la figura 17 se observa la comparación del ROC para el experimento 4 usando el modelo CNN (figura izquierda) y experimento 1 usando Inception V3 (figura derecha).

Figura 17. *Evaluación de la curva ROC en modelos de CNN e Inception V3*



Para medir la efectividad de ambos modelos se utilizó la matriz de confusión (confusion matrix), medida del rendimiento para la clasificación de problemas de Machine Learning. En la figura 18 se presenta los resultados para el experimento 4 usando CNN (figura izquierda) y experimento 1 usando Inception V3 (figura derecha).

Figura 18. *Resultados matrices de confusión para los modelos CNN e Inception V3*



Para el modelo CNN, los verdaderos negativos son 202 y los verdaderos positivos 43, es decir, de los 245 casos (diagonal) de un total de 318 imágenes tomadas del test set, predijo correctamente las categorías de cada clase. Por su parte, el Inception V3 tiene 240 verdaderos negativos y 49 verdaderos positivos, lo que significa que en 289 casos (diagonal) se clasificó correctamente las categorías de cada clase.

El modelo CNN, cuenta con 52 casos de error tipo I y 21 casos de error tipo II. Por su parte, el modelo Inception V3 cuenta con 14 casos de error tipo I y tan solo 15 casos de error tipo II. Como se dijo anteriormente, esto para el problema que se está tratando es el error más grave que puede ocurrir.

Evaluando los resultados de ambos modelos en la matriz de confusión se calcula las métricas de precisión, sensibilidad y F1-score, esta última será a partir de la cual

se tomó la decisión de elegir el mejor modelo para clasificación, dado que ha sido diseñada para funcionar bien cuando se tienen distribuciones de clases desiguales, con un alto valor de precisión en la clase mayoritaria (negativos) y un bajo recall en la clase minoritaria (positivos). Ver tabla 10.

Tabla 10. *Métricas de la matriz de confusión*

Modelos seleccionados		Resultados		
		Precision	Recall	F1 Score
Experimento 4	CNN aplicando Class Weights y Data Augmentation	45,3%	67,2%	54,1%
Experimento 1	Inception V3	77,8%	76,6%	77,2%

Para el modelo de CNN, si bien el recall presenta un resultado aceptable, la precisión es cercana al 0.5, esto significa que el modelo de ML detecta bien la clase, pero también incluye muestras de la otra clase. Por su parte, el Inception V3 cuenta con resultados altos en precisión y recall, indicando que el modelo la mayoría de las veces maneja bien la clasificación de clases y un F1-score del 77.2% que respalda un buen rendimiento del modelo.

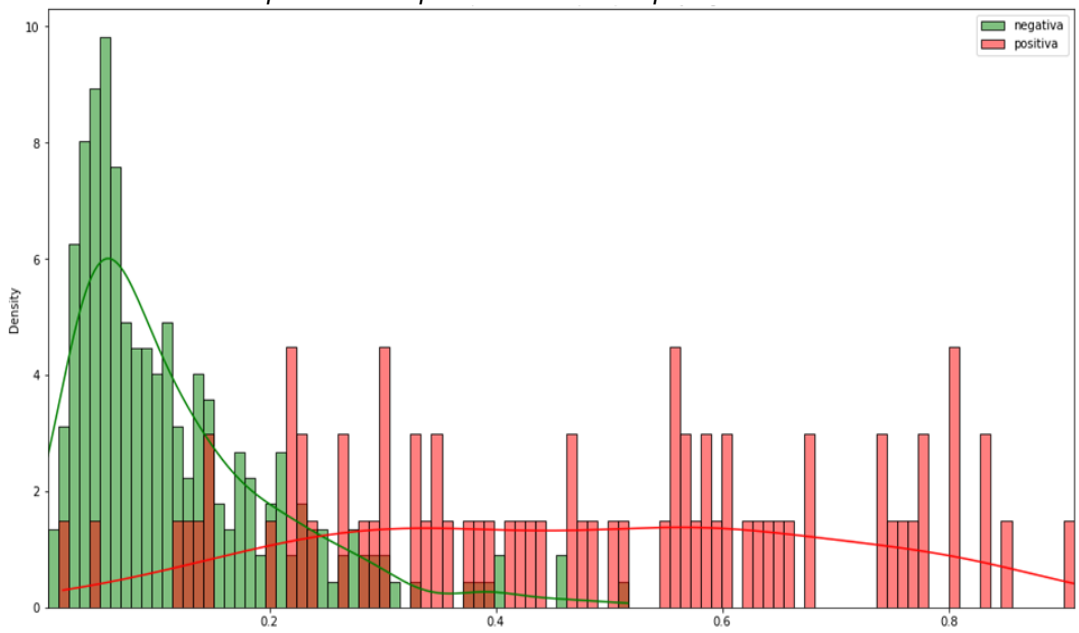
Teniendo en cuenta los resultados del training, validation, test y la métrica del F1-score de ambos modelos y analizando la matriz de confusión, se decide seleccionar como mejor modelo el Inception V3. Las razones para tomar esta decisión fueron las siguientes:

- Evaluando los resultados del AUC en la curva ROC, este modelo obtuvo el segundo resultado más alto de los ocho experimentos en validation y el resultado más alto en test, esto es sinónimo de robustez del modelo, al contar con un gran número de imágenes y un modelo previamente preentrenado.
- Disminuye los efectos adversos de contar con una baja cantidad de datos y clases desbalanceadas.

- La métrica del F1-score fue del 77.2% un buen resultado y aproximadamente 23 puntos porcentuales más alto si se compara con el modelo de CNN.

Con base en el modelo seleccionado, se calculó el umbral de probabilidad en el set de validación para saber desde que punto el clasificador asigna una observación a una clase en específico y posteriormente se realizó un gráfico de distribución de probabilidades para observar el comportamiento en el set de test. El umbral obtenido fue de 0.22, este es el punto que máxima la diferencia entre la tasa de verdaderos positivos y falsos negativos en el set de validación. Esto significa que si la probabilidad del modelo es mayor o igual el modelo clasifica la imagen del glaucoma del paciente como positivo y si es menor la clasifica como negativo. Ver figura 19.

Figura 19. *Distribución de probabilidad para el modelo Inception V3*

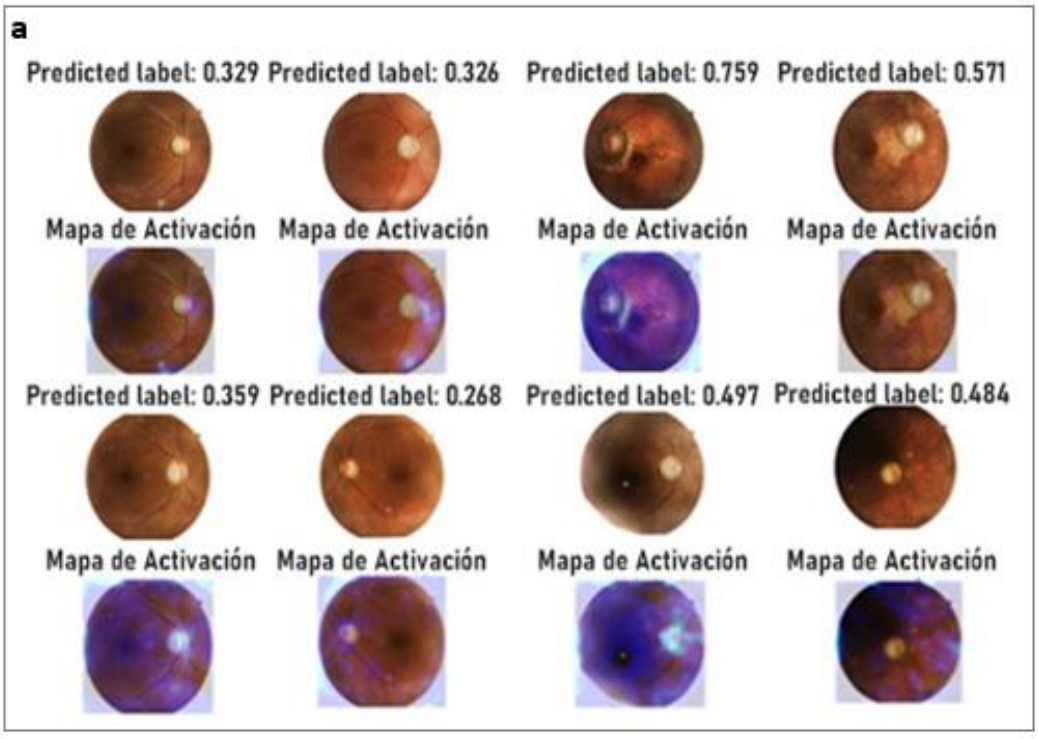


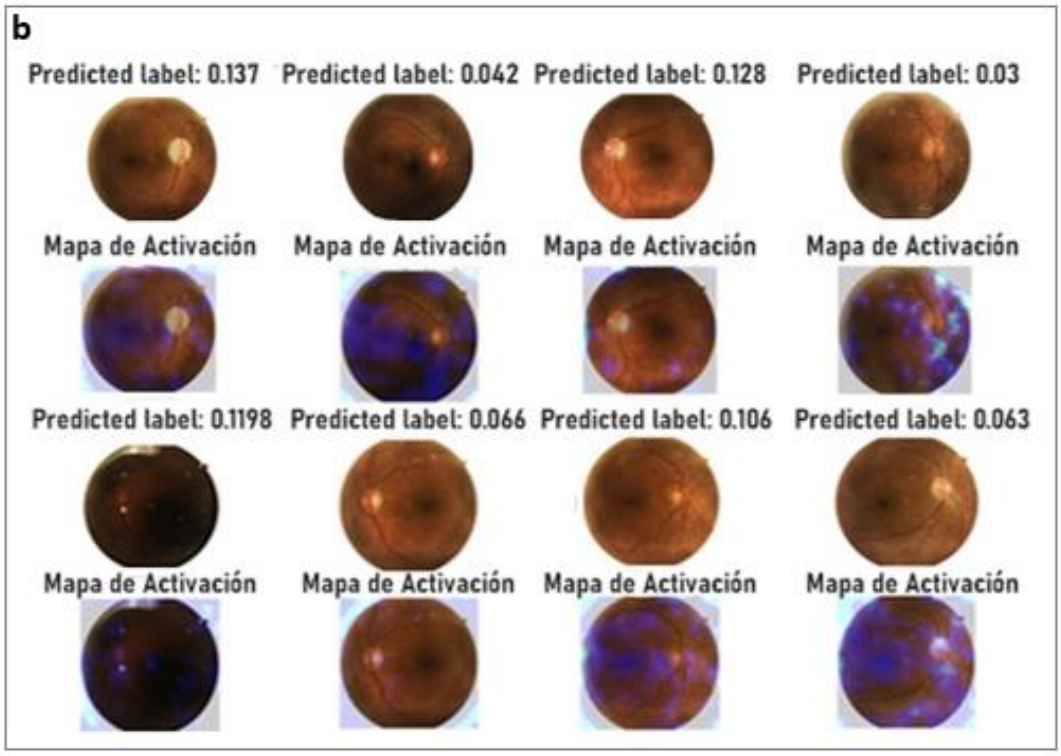
Lo anterior, se complementó usando Grad-Cam, técnica ampliamente utilizada en modelos de Machine Learning para tareas de clasificación dado que no requiere una arquitectura de CNN en particular. Esta técnica es una generalización de CAM

(class activation mapping), un método que requiere usar una arquitectura en particular.

Este método permite producir mapas de calor que se aplica a una red neuronal ya entrenada una vez completado el entrenamiento y fijado los parámetros. El objetivo de usar esta técnica es comprender qué partes de una imagen de entrada fueron importantes para una decisión de clasificación por parte del algoritmo y ser usado como material de insumo para la toma de decisiones por parte del médico oftalmólogo. En la figura 20 se observa el resultado de diferentes fotografías del fondo del ojo correctamente categorizadas en positivas y negativas, visualizando las regiones de entrada que son "importantes" para las predicciones de este modelo y el cálculo del umbral de probabilidad que define a qué categoría pertenece dicha imagen.

Figura 20. *Técnica del Grad Cam en imágenes positivas y negativas en el modelo Inception V3*





Nota: el umbral de probabilidad es 0.22. Una probabilidad mayor o igual a este umbral significa que la imagen del glaucoma del paciente es positiva y si es menor la clasifica como negativa. El Primer grupo de imágenes representa la categoría positiva correctamente clasificada (a) y el segundo grupo representa la categoría negativa correctamente clasificada (b).

A partir de la figura anterior, se observa cómo el mapa de activación se concentra principalmente en la mayoría de las imágenes en la zona del nervio óptico, cercana a la copa óptica, parte brillante y central del ojo. Esto se explica porque una característica de una imagen con glaucoma es la presencia de un tamaño anormal en esta zona que usualmente el especialista examina para emitir un parte médico.

6.4 Despliegue

Luego de haber sido documentado, construido y validado el modelo, se presenta con los resultados obtenidos a los tutores de este proyecto, se explica su funcionamiento, alcances, propuestas y mejoras que se podrían realizar a futuro. El objetivo para los próximos años es construir un conjunto de prácticas (MLOps) para la comunicación entre el personal de analítica y científicos del glaucoma en la

organización que permita adquirir nuevos conjuntos de datos; tener un seguimiento y control de versiones para experimentos y ejecuciones de entrenamiento de modelos; y configurar los pipelines de implementación y monitoreo para los modelos que llegan a producción. Lo anterior, con el fin de simplificar el proceso de gestión y automatización de la implementación de modelos de Machine Learning y Deep Learning para un contexto médico. De esta manera será más fácil alinear los modelos con las necesidades médicas y comerciales, así como con los requisitos reglamentarios.

Esto servirá de insumo para ser usado por parte de la clínica donde trabaja el Doctor Carlos Rivera, uno de los directores de tesis, para la clasificación del glaucoma en nuevos pacientes, así como, ser usado de insumo para futuras investigaciones sobre modelos de Machine Learning para el tamizaje del glaucoma.

7. CONCLUSIONES Y RECOMENDACIONES

A lo largo de este proyecto se formularon y validaron modelos de Machine Learning adaptados a la población del Valle del Cauca tomando como insumo fotos a color del fondo del ojo, encontrando para este problema que el modelo que mejor se ajustó a las métricas de evaluación AUC fue Inception V3 con resultados en el set de validación y test de 0.8706 y 0.9084 respectivamente; sus métricas en la matriz de confusión arrojaron un F1-score de 77.2%, con una precisión del 77.8% y una sensibilidad del 76.6%; por último, el umbral de probabilidad a partir del cual clasifica el modelo a una persona con glaucoma fue de 0.22.

En la literatura abordada, los resultados en las métricas de la matriz de confusión por lo general están en un rango para los valores de sensibilidad y especificidad entre 90% y 99%, esto se debe en gran medida a los extensos conjuntos de datos empleados, que oscilan en promedio entre las 30.000 y 50.000 imágenes, aunado a la capacidad de contar con equipos capaz de procesar múltiples modelos de Deep Learning, con diferentes parámetros e hiperparámetros para grandes volúmenes de datos, sin restricciones de capacidad, memoria y/o tiempo.

Aunque muchos algoritmos de estudio han demostrado valores impresionantes de AUC, sensibilidad y especificidad, es difícil comparar la aplicabilidad clínica entre diferentes estudios con disímiles metodologías empleadas. Además, los algoritmos pueden variar entre los entornos clínicos a causa de la variabilidad ocular considerable dentro de las poblaciones de pacientes según factores como la edad, el género, el error de refracción, las comorbilidades médicas y el origen étnico; así como, debido a la naturaleza subjetiva y variable de los datos informados por los pacientes.

En el trabajo de grado se validó que los modelos de Transfer Learning para el uso de Deep Learning tienen una marcada ventaja sobre modelos entrenados desde cero, como fue el caso de los modelos CNN empleados. Se enfrentó a un desbalance de clases y una baja cantidad de datos, que fue contrarrestada utilizando las técnicas de Data Augmentation, Class Weights y de regularización para reducir el sobreajuste y aumentar la robustez en el modelo**.**

Se recomienda utilizar muestras balanceadas que permitan comparar los resultados en las métricas de evaluación y determinar cuánto difieren de las clases desbalanceadas.

En un trabajo futuro podría considerarse las variables como el sexo, la edad y la etnia para explorar modelos individuales sobre cada una de estas, utilizando diferentes enfoques para clasificar la región del disco óptico es una determinada clase, tales como: normales, sospechosas y anormales. Así mismo, Se podrían considerar otras técnicas de muestreo, tales como el oversampling y SMOTE. También, se pueden explorar modelos de extracción de características como es el caso del SVM y KNN. Se puede utilizar técnicas de mapas de calor diferentes al Grad-CAM, que permitan resaltar solo las regiones de la imagen que el modelo utilizó para la predicción como la técnica del Hirescam.

Respecto a las limitaciones del proyecto, fue necesario estar constantemente adaptando el cronograma de trabajo dado que los modelos fueron ejecutados en Google Colab, llegando a pagar una suscripción pro dada las limitaciones de la plataforma en tiempo, memoria y ram disponible.

En definitiva, tener un modelo que involucra la variable raza/etnia como eje central de la investigación para clasificar pacientes con glaucoma es un resultado inédito en el Valle del Cauca, con métricas de evaluación similares a las presentadas en el estado del arte. Sin lugar a duda, es claro como la inteligencia artificial permite

mejorar el papel de los médicos especialistas e inevitablemente dará forma al futuro de la atención del glaucoma a la próxima generación. Lo anterior, servirá como complemento para el diagnóstico de glaucoma sin reemplazar el juicio médico, pero sí facilitando la toma de decisiones.

BIBLIOGRAFÍA

Alemayehu, D., & Zou, K. H. (2012). Applications of ROC analysis in medical research: recent developments and future directions. *Academic radiology, 19*(12), 1457-1464.

Alghamdi, H. S., Tang, H. L., Waheeb, S. A., & Peto, T. (2016). Automatic optic disc abnormality detection in fundus images: A deep learning approach. *OMIA3*, 10-17.

Barros, D., Moura, J. C., Freire, C. R., Taleb, A. C., Valentim, R. A., & Morais, P. S. (2020). Machine learning applied to retinal image processing for glaucoma detection: review and perspective. *Biomedical engineering online, 19*(1), 1-21.

Bianco, S., Cadene, R., Celona, L., & Napoletano, P. (2018). Benchmark analysis of representative deep neural network architectures. *IEEE Access, 6*, 64270-64277.

Bock, R., Meier, J., Nyúl, L. G., Hornegger, J., & Michelson, G. (2010). Glaucoma risk index: automated glaucoma detection from color fundus images. *Medical image analysis, 14*(3), 471-481.

Boyd, K. (22 de Septiembre de 2021). *What is Glaucoma?: American Academy of Ophtalmology*. Obtenido de American Academy of Ophtalmology: https://www.aao.org/eye-health/diseases/what-is-glaucoma

Chen, X., Xu, Y., Wong, D. W., Wong, T. Y., & Liu, J. (Agosto de 2015). Glaucoma detection based on deep convolutional neural network. En *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (págs. 715-718). IEEE.

Dhungel, N., Carneiro, G., & Bradley, A. P. (Octubre de 2016). The automated learning of deep features for breast mass classification from mammograms. En *International Conference on Medical Image Computing and Computer-Assisted Intervention* (págs. 106-114). Springer, Cham.

Diaz-Pinto, A., Morales, S., Naranjo, V., Köhler, T., Mossi, J. M., & Navea, A. (2019). CNNs for automatic glaucoma assessment using fundus images: an extensive validation. *Biomedical engineering online, 18*(1), 1-19.

Efros, A. A., & Freeman, W. T. (Agosto de 2001). Image quilting for texture synthesis and transfer. En *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* (págs. 341-346).

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., & Madry, A. (2018). A rotation and a translation suffice: Fooling cnns with simple transformations.

Feng, K., Hong, H., Tang, K., & Wang, J. (2019). Decision making with machine learning and ROC curves. *Available at SSRN 3382962*.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics, 7*(2), 179-188.

Fradkov, A. L. (2020). Early history of machine learning. *IFAC-PapersOnLine, 53*(2), 1385-1390.

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., . . . Allen, N. E. (2017). Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American journal of epidemiology, 186*(9), 1026-1034.

Gobernación del Valle del Cauca. (2020). *Demografía Valle del Cauca.* Valle del Cauca.

Harasymowycz, P., Birt, C., Gooi , P., Heckler, L., Hutnik, C., Jinapriya, D., . . . Day, R. (2016). Medical management of glaucoma in the 21st century from a Canadian perspective. *Journal of ophthalmology, 2016*.

Isaac, P. D. (1976). EGAN, JP" Signal Detection Theory and ROC Analysis"(Book Review). *The Psychological Record, 26*, 567.

Japkowicz, N. (Julio de 2000). Learning from imbalanced data sets: a comparison of various strategies. En *AAAI workshop on learning from imbalanced data sets* (Vol. 68, págs. 10-15). AAAI Press Menlo Park, CA.

Khan, S. M., Liu, X., Nath, S., Korot, E., Faes, L., Wagner, S. K., . . . Denniston, A. K. (2021). A global review of publicly available datasets for ophthalmological

imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health, 3*(1), e51-e66.

Kingman, S. (2004). Glaucoma is second leading cause of blindness globally. *Bulletin of the World Health Organization, 82*, 887-888.

Križaj, D. (2019). What is glaucoma? *Webvision: The Organization of the Retina and Visual System [Internet]*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems, 25*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*(7553), 436-444.

Lee, Y. S., & Bang, C. C. (2021). Framework for the Classification of Imbalanced Structured Data Using Under-sampling and Convolutional Neural Network. *Information Systems Frontiers*, 1-15.

Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

López Rojas, C., Belalcázar Rey, S., & Dávila Ramírez, F. (2015). Prevalencia del Glaucoma y su Contribución a la Discapacidad Visual en Colombia. *Sociedad Colombiana de Oftalmología, 48*(2), 175 -181.

McMonnies, C. W. (2017). Glaucoma history and risk factors. *Journal of optometry, 10*(2), 71-78.

Mehta, P., Petersen, C. A., Wen, J. C., Banitt, M. R., Chen, P. P., Bojikian, K. D., . . . & Vision Consortium. (2021). Automated detection of glaucoma with interpretable machine learning using clinical data and multimodal retinal images. *American Journal of Ophthalmology, 231*, 154-169.

Mikołajczyk, A., & Grochowski, M. (Mayo de 2018). Data augmentation for improving deep learning in image classification problem. En *2018 international interdisciplinary PhD workshop (IIPhDW)* (págs. 117-122). IEEE.

Mursch-Edlmayr, A. S., Ng, W. S., Diniz-Filho, A., Sousa, D. C., Arnould, L., Schlenker, M. B., . . . Jayaram, H. (2020). Artificial intelligence algorithms to

diagnose glaucoma and detect glaucoma progression: translation to clinical practice. *Translational vision science & technology, 9*(2), 55-55.

Nathan, N., & Joos, K. M. (Julio de 2016). Glaucoma disparities in the Hispanic population. En *Seminars in ophthalmology* (Vol. 31, págs. 394-399). Taylor & Francis.

Nayak, J., Acharya, U. R., Bhat, P. S., Shetty, N., & Lim, T. C. (2009). Automated diagnosis of glaucoma using digital fundus images. *Journal of medical systems, 33*(5), 337-346.

Öhnell, H., Heijl, A., Anderson, H., & Bengtsson, B. (2017). Detection of glaucoma progression by perimetry and optic disc photography at different stages of the disease: results from the Early Manifest Glaucoma Trial. *Acta Ophthalmologica, 95*(3), 281-287.

Pérez Molina, E., & León Veitía, L. (2017). La fotografía de fondo de ojo como método de diagnóstico en el glaucoma. *Medicentro Electrónica, 21*(1), 3-10.

Phan, S., Satoh, S. I., Yoda, Y., Kashiwagi, K., & Oshika, T. (2019). Evaluation of deep convolutional neural networks for glaucoma detection. *Japanese journal of ophthalmology, 63*(3), 276-283.

Provost, F., & Fawcett, T. (1997). Analysis and visualization of classifier performance with nonuniform class and cost distributions. En *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management* (págs. 57-63).

Quigley, H. A., & Broman, A. T. (2006). The number of people with glaucoma worldwide in 2010 and 2020. *British journal of ophthalmology, 90*(3), 262-267.

Rahman, M. M., & Davis, D. N. (2013). Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing, 3*(2), 224.

Sekimitsu, S., & Zebardast, N. (2021). Glaucoma and machine learning: A call for increased diversity in data. *Ophthalmology Glaucoma, 4*(4), 339-342.

Smits, D. J., Elze, T., Wang, H., & Pasquale, L. R. (Mayo de 2019). Machine learning in the detection of the glaucomatous disc and visual field. En *Seminars in Ophthalmology* (Vol. 34, págs. 232-242). Taylor & Francis.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). Going deeper with convolutions. En *Proceedings of the IEEE conference on computer vision and pattern recognition* (págs. 1-9).

Tielsch, J. M., Sommer, A., Katz, J., Royall, R. M., Quigley, H. A., & Javitt, J. (1991). Racial variations in the prevalence of primary open-angle glaucoma: the Baltimore Eye Survey. *Jama, 266*(3), 369-374.

Wang, H., & Raj, B. (2017). On the origin of deep learning. *arXiv preprint arXiv:1702.07800*.

Wirth, R., & Hipp, J. (Abril de 2000). CRISP-DM: Towards a standard process model for data mining. En *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1, págs. 29-40).

Wu, C. W., Shen, H. L., Lu, C. J., Chen, S. H., & Chen, H. Y. (2021). Comparison of Different Machine Learning Classifiers for Glaucoma Diagnosis Based on Spectralis OCT. *Diagnostics, 11*(9), 1718.

Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (Abril de 2020). Random erasing data augmentation. En *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, págs. 13001-13008).