Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ciencias y Sistemas

DISEÑO DE INVESTIGACIÓN PARA UNA HERRAMIENTA PÚBLICA Y DE CÓDIGO ABIERTA CON *MACHINE LEARNING* PARA LA DETECCIÓN, CLASIFICACIÓN Y PRONÓSTICO DEL CÁNCER DE MAMA

Ruth Nohemy Ardón Lechuga
Asesorado por la Msc. Inga. Mildred Caballeros

Guatemala, julio de 2023

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

DISEÑO DE INVESTIGACIÓN PARA UNA HERRAMIENTA PÚBLICA Y DE CÓDIGO ABIERTA CON *MACHINE LEARNING* PARA LA DETECCIÓN, CLASIFICACIÓN Y PRONÓSTICO DEL CÁNCER DE MAMA

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

RUTH NOHEMY ARDÓN LECHUGA
ASESORADO POR LA MSC. INGA. MILDRED CABALLEROS

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO EN CIENCIAS Y SISTEMAS

GUATEMALA, JULIO DE 2023

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANO A.I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Ing. Kevin Vladimir Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

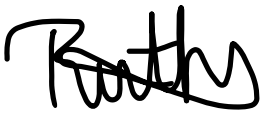
DECANA	Inga. Aurelia Anabela Cordova Estrada
EXAMINADOR	Inga. Devora Emperatris Meza Orella
EXAMINADOR	Ing. Pedro Pablo Hernández Ramírez
EXAMINADOR	Ing. Oscar Alejandro Paz Campos
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

DISEÑO DE INVESTIGACIÓN PARA UNA HERRAMIENTA PÚBLICA Y DE CÓDIGO ABIERTA CON *MACHINE LEARNING* PARA LA DETECCIÓN, CLASIFICACIÓN Y PRONÓSTICO DEL CÁNCER DE MAMA

Tema que me fuera asignado por la Dirección de Escuela de Estudios de Postgrado con fecha abril del 2023.



Ruth Nohemy Ardón Lechuga

EEPFI-PP-0355-2023
Guatemala, 13 de abril de 2023

Director
Carlos Gustavo Alonzo
Escuela De Ingenieria En Sistemas
Presente.

Estimado Ing. Alonzo

Reciba un cordial saludo de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería.

El propósito de la presente es para informarle que se ha revisado y aprobado el Diseño de Investigación titulado: **HERRAMIENTA PÚBLICA Y DE CÓDIGO ABIERTA CON MACHINE LEARNING PARA LA DETECCIÓN, CLASIFICACIÓN Y PRONÓSTICO DEL CANCER DE MAMA**, el cual se enmarca en la línea de investigación: **Mineria de datos - Mineria de datos**, presentado por la estudiante **Ruth Nohemy Ardón Lechuga** carné número **201602975**, quien optó por la modalidad del "PROCESO DE GRADUACIÓN DE LOS ESTUDIANTES DE LA FACULTAD DE INGENIERÍA OPCIÓN ESTUDIOS DE POSTGRADO". Previo a culminar sus estudios en la Maestría en ARTES en Ingeniería Para La Industria Con Especialidad En Ciencias De La Computación.

Y habiendo cumplido y aprobado con los requisitos establecidos en el normativo de este Proceso de Graduación en el Punto 6.2, aprobado por la Junta Directiva de la Facultad de Ingeniería en el Punto Décimo, Inciso 10.2 del Acta 28-2011 de fecha 19 de septiembre de 2011, firmo y sello la presente para el trámite correspondiente de graduación de Pregrado.

Atentamente,

"Id y Enseñad a Todos"

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
ESCUELA DE POSTGRADO
FACULTAD DE INGENIERÍA

Mildred Madai Caballeros M.
MAESTRÍA EN CIENCIAS Y SISTEMAS
COORDINADOR(A)

Mtra. Mildred Madai Caballeros Morales De Agreda
Asesor(a)

null Carlos Gustavo Alonzo
Coordinador(a) de Maestría

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA
DIRECCIÓN
ESCUELA DE POSTGRADO

Mtro. Edgar Dario Alvaréz Cotí
Director
Escuela de Estudios de Postgrado
Facultad de Ingeniería

UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA

FACULTAD DE INGENIERÍA

EEP-EICS-0354-2023

El Director de la Escuela De Ingenieria En Sistemas de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del Asesor, el visto bueno del Coordinador y Director de la Escuela de Estudios de Postgrado, del Diseño de Investigación en la modalidad Estudios de Pregrado y Postgrado titulado: **HERRAMIENTA PÚBLICA Y DE CÓDIGO ABIERTA CON MACHINE LEARNING PARA LA DETECCIÓN, CLASIFICACIÓN Y PRONÓSTICO DEL CANCER DE MAMA**, presentado por el estudiante universitario **Ruth Nohemy Ardón Lechuga**, procedo con el Aval del mismo, ya que cumple con los requisitos normados por la Facultad de Ingeniería en esta modalidad.

ID Y ENSEÑAD A TODOS

Ing. Carlos Gustavo Alonzo
Director
Escuela De Ingenieria En Sistemas

Guatemala, abril de 2023

USAC
TRICENTENARIA
Universidad de San Carlos de Guatemala

Decanato
Facultad e Ingeniería
24189101- 24189102

LNG.DECANATO.OIE.36.2023

El Decano de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Ingeniería En Ciencias Y Sistemas, al Trabajo de Graduación titulado: **HERRAMIENTA PÚBLICA Y DE CÓDIGO ABIERTA CON MACHINE LEARNING PARA LA DETECCIÓN, CLASIFICACIÓN Y PRONÓSTICO DEL CANCER DE MAMA**, presentado por: **Ruth Nohemy Ardón Lechuga** después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

Firmado electrónicamente por: José Francisco Gomez Rivera
Motivo: Orden de impresión
Fecha: 18/07/2023 10:43:13
Lugar: Facultad de Ingeniería, USAC.

Ing. José Francisco Gómez Rivera
Decano a.i.

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANO a.i.
Facultad de Ingeniería

Guatemala, julio de 2023

ACTO QUE DEDICO A:

Dios	Por ser mi guía, soporte y ancla en cada momento de mi vida. Por llenar mi vida de bendiciones y brindarme la sabiduría para poder afrontar cada desafío.
Mis padres	Yiznarda Lechuga, por su amor incondicional y apoyo infinito en cada paso de mi vida, porque sin ti nada sería posible y las palabras nunca alcanzarán para agradecerte. Juan Guillermo por su amor, por ser un ejemplo de trabajo y esfuerzo.
Mis hermanos	Joshua y Daniel Guerra, por su ejemplo de nobleza, perseverancia y trabajo constante. Por estar siempre para mí, sus consejos y risas.
Mis abuelos	Ruth Lechuga, por ser un modelo de servicio a los demás y gratitud. Luciano Patzan, por ser un segundo padre, por velar en que fuera mi mejor versión.

Mis sobrinos	Andrés y Camila Guerra, por recordarme que la felicidad y el amor está tras las cosas simples de la vida.
Mi novio	Víctor Reyes, por ser un gran compañero de vida; por siempre escucharme, apoyarme y motivarme; por ser un ejemplo de disciplina y resiliencia

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por la oportunidad de formarme como profesional.
Mis cuñadas	Lizbet Pérez y Kimberly Arriaga, por ser confidentes, por escucharme y aconsejarme.
Víctor Reyes	Mi novio, por apoyarme y motivarme a dar lo mejor de mí.
Mis amigos de la Facultad	Herbert Reyes, Mindi Aguilar, Andrés Carvajal, Fernando Pensamiento, Jeannira Sic, Carlos Campaneros, por hacer de la carrera una mejor experiencia, brindarme su apoyo y aprendizaje en cada curso.
Mis amigos del colegio	Camilla Rodas, Michelle Flores, Jennipher Molina, Belén Gomez, Kellen Castañeda, Alexander Rodriguez, por mostrarme el significado de una amistad verdadera; aunque nuestras vidas han tomado caminos distintos, siempre han estado ahí.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
API	Interfaz de programación de aplicaciones
CAD	Herramienta de diagnóstico asistido
CART	Árbol de clasificación y regresión
CDI	Carcinoma ductal invasivo
k-NN	k Nearest Neighbors
NB	Naive Bayes
PAAF	Biopsia con aspiración de aguja fina
SVM	Support Vector Machine
$	Porcentaje
±	Más-menos
<	Menor a
>	Mayor a
Q.	Quetzal

GLOSARIO

API	Conjunto de subrutinas, funciones y procedimientos ya creados que ofrecen las bibliotecas de servicios.
Endpoint	Herramienta de control de versiones.
Github	Herramienta de control de versiones.
Histopatológicas	Examen microscópico de tejido para estudiar las manifestaciones de la enfermedad.
Laravel	*Framework* de código abierto para desarrollar aplicaciones y servicios web con PHP.
Matlab	Sistema de cómputo numérico que ofrece un entorno de desarrollo integrado con un lenguaje de programación propio.
Microcalcificaciones	Depósito de fosfato de calcio, oxalato de calcio o carbonato de calcio que se precipitan en sitios de actividad anormal en el tejido mamario.
MySQL	Sistema de gestión de bases de datos relacional.
PHP	Lenguaje de programación.

Python	Lenguaje de programación.

RESUMEN

El cáncer es una enfermedad grave que afecta a cualquier órgano del cuerpo, esta es la principal causa de muerte a nivel mundial; entre ellos el cáncer de mama es uno de los más comunes.

El diagnóstico temprano es vital para aumentar las posibilidades de supervivencia y evitar tratamientos innecesarios. Sin embargo, el diagnóstico puede ser complejo debido a mutaciones y errores en los resultados.

Se han desarrollado herramientas de diagnóstico asistido basadas en minería de datos y aprendizaje automático, pero la mayoría requieren de conocimientos técnicos para su uso.

Por ello, es necesario crear una herramienta accesible, precisa y gratuita que facilite el análisis de mamografías y biopsias, reduciendo tiempo y recursos, y acercando el diagnóstico del cáncer de mama entre los sectores público y privado.

Esta investigación propone una herramienta pública y *open source* para la detección, diagnóstico y clasificación del cáncer de mamá mediante el uso de *Machine Learning* e Inteligencia artificial.

1. INTRODUCCIÓN

El cáncer es la principal causa de muerte alrededor del mundo, siendo el cáncer de mama el quinto con mayor número de fallecimientos alrededor del mundo, atribuyéndole 685 mil defunciones durante el 2021.

Por ello, el cáncer de mama ha sido el foco de múltiples investigaciones y proyectos que buscan brindar maneras confiables y novedosas de detectar anomalías en etapas tempranas; sin embargo, a pesar de los muchos trabajos existentes, no existe una herramienta al alcance de todos los profesionales.

Es así como el presente trabajo busca construir una herramienta publica y de código abierto para el diagnóstico y clasificación del cáncer de mama, mediante el uso de modelos de *Machine Learning* para la lectura de mamografías, pruebas de biopsia con aspiración de aguja fina y clasificación de los subtipos de cáncer de mama.

El cual se aborda como una investigación cuantitativa con carácter experimental, en esta se estudia la forma en que se relacionan diversas variables con el fin de construir modelos de *Machine Learning* para realizar inferencias.

La investiga comienza abordando la problemática del cáncer de mama y plantea las preguntas orientadas que se buscan responder durante la investigación, con ellos se procede a plantear los objetivos a conseguir mediante la investigación.

Posteriormente, se explica la metodología que se usará para desarrollar el proyecto, las técnicas de análisis de información a usar en el proyecto y el cronograma que abarca su realización.

Una vez hecho esto se procede con el estudio del estado del arte, es decir los antecedentes investigativos sobre herramientas de este carácter; así en el capítulo dos se aborda la justificación de la investigación y posteriormente, en el capítulo tres los alcances desde una perspectiva investigativa, técnica y de resultados.

El siguiente capítulo abarca todo el marco teórico que compone la investigación, se detallan los conceptos y técnicas necesarias para desarrollar el proyecto. En el capítulo cinco se presentan y detallan los resultados de la investigación.

En el capítulo seis se realiza una discusión de los resultados y se procede, basándose en los resultados, a plantear las conclusiones y recomendaciones de la investigación.

2. ANTECEDENTES

Dentro de los múltiples estudios y herramientas que se han hecho acerca del diagnóstico de cáncer de mama utilizando *Machine Learning*, el que se considera el pionero y que sentó las bases para trabajos posteriores es la tesis doctoral realizada en 1994 por W. Nick Sreet en la Universidad de Wisconsin.

W Nick Street buscó diagnosticar y pronosticar el cáncer de mama a través de métodos de *machine learning* basados en programación lineal, creo un sistema al que denomino XCYT; el sistema incluía un análisis citológico automatizado y la segmentación de imágenes que permitía calcular características de las muestras de imágenes digitales tomadas a partir de pruebas de aspiración de aguja fina (PAAF).

Para segmentar las imágenes se utilizó un procedimiento semi automático conocido como “Snakes”, en este procedimiento el usuario establece un punto inicial, con este el “Snake” localizará el límite real del núcleo celular, esto ayuda a definir el contorno de las células para obtener información de sus características; una desventaja de este proceso es que el punto inicial del Snake se debía ajustar múltiples veces para lograr definir los contornos correctamente.

Partiendo de la segmentación, W. Nick Street obtuvo información celular relevante como: Radio, perímetro, área, suavidad, concavidad, etcétera. Y con esta información creó un proceso de clasificación para determinar si la muestra de células era benigna o maligna, este proceso utiliza el algoritmo de *Multisurface Method-Tree* (MSM-Tree) que funciona a través de programación lineal.

El sistema fue utilizado durante nueve meses por el Dr. W. H. Wolberg y otras instituciones colaborativas, concluyendo que contaba con una mejor precisión al determinar la forma de los núcleos celulares y clasificarlas.

Sin embargo, W. Nick Street, apunto puntos de mejora, dentro de los cuales los más importantes fueron: Automatizar el punto inicial sobre el cual empieza a trabajar el Snake (ya que requiere de intervención humana y un experto), mejorar el algoritmo para segmentar imágenes (este es limitado por la aproximación inicial) y propone implementar una base de datos para almacenar las imágenes.

La información celular que se obtuvo al segmentar la imagen fue almacenada en una base de datos denominada *Breast Cancer Wisconsin Dataset*, esta contiene 569 muestras en donde cada muestra contiene 32 atributos.

La investigación del Dr. W. Nick Street sentó las bases para investigaciones posteriores y son utilizados hasta hoy en día; por ejemplo, en el año 2011, el Indian Journal of Computer Science and Engineering (IJCSE) público una investigación realiza por D. Lavanya y Dr. K. Usha Rani en la cual utilizaba tres *dataset* de información (Breast Cancer Wisconsin Original, Breast Cancer Wisconsin Diagnostic y Breast Cancer).

El estudio de Layanya y Usha, buscaba determinar la importancia de la selección de datos para los algoritmos de clasificación, por lo que se aplicó el algoritmo Classification and Regression Trees (CART) utilizando una herramienta para aprendizaje automático y minería de datos conocida como Weka para comparar la veracidad de los resultados en base al set de datos.

En el estudio de Layanya y Usha, se concluyó que el dataset que brindaba mayor precisión era el Breast Cancer Wisconsin Original, además se concluyó que no solo era importante considerar el número de muestras y cantidad de atributos si no la calidad de los atributos que se escogían para construir el modelo.

Por otro lado, también se ha utilizado Machine Learning para el estudio directo de las mamografías, durante el año 2015 en la universidad de los Andes, se realizó un estudio para la implementación de técnicas de machine learning que buscaba identificar microcalcificaciones en mamografías.

El estudio utilizaba como dataset la Base de Datos Digital de Mamografía de Tamizaje (DDSM) desarrollada por el Hospital General de Massachusetts, incluyendo 2620 estudios de mamografías, cada muestra con dos imágenes de cada mama, información del paciente, información de la imagen y hallazgos de cada imagen (Pedroza, 2015).

Para la implementación de la solución, se utilizó MATLAB con las librerías Neuronal Networks Toolbox y Statistics and Machine Learning Toolbox. Partiendo del dataset, se extrajeron recuadros que incluían microcalcificaciones y tejido normal para entrenar al modelo; con esta información recaudada se analizó la diferencia en niveles de gris de cada muestra.

En el diseño del modelo, se utilizaron diversas técnicas de aprendizaje supervisado, con ello se predijo si era una microcalficación maligna o un tejido normal.

Con los modelos entrenados, se construyó un sistema en el cual basado en la imagen de una mamografía, se realiza una segmentación de la imagen,

posteriormente se aplican técnicas de procesamiento de imágenes y por último se ejecuta el modelo de detección de microcalcificaciones.

Al final del estudio se obtuvo un error de clasificación mínimo del 0.92 ± 6.07 % y un error máximo del 2.86 ± 6.07 %, lo cual indica bastante precisión, sin embargo, el autor indica que se puede mejorar la precisión al utilizar otros modelos de *Machine Learning*, utilizando un dataset más numeroso e implementando otras técnicas de procesamiento de imágenes (Pedroza, 2015).

En enero del año 2022 se aprobó un trabajo de investigación por parte de Medisur, la cual es una revisa científica electrónica de las Ciencias Médicas, que consistía en una herramienta para el diagnóstico de cáncer de mama en imágenes histopatológicas denominada como HistoBCAD; el proyecto buscaba detectar regiones tumorales en imágenes histopatológicas digitales (Pérez-Marrero, 2022).

Para la construcción del modelo se utilizó un clasificador de bosques aleatorios utilizando la base de datos pública de la Cruz-Roa et al del año 2014, que contiene 162 muestras.

La herramienta tomaba una imagen y la dividía en mosaicos de 50x50 pixeles para procesarlo por bloques, cada mosaico se procesa para extraer características y se procesa para convertirla a escala de grises para obtener más información, el conjunto de características de color y texturas se utilizan como entradas para clasificar la presencia del carcinoma.

El modelo se construyó utilizando la librería OpenSlide de Python para la lectura de imágenes, mientras que el entorno gráfico se realizó utilizando PHP con Laravel y MySQL como gestor de bases de datos.

A su vez, el Massachusetts Institute of Technology (MIT) ha realizado múltiples investigaciones orientadas a desarrollar algoritmos que ayuden a mejorar los modelos para prevenir que la enfermedad progrese y se apliquen tratamientos innecesarios a causa de malos diagnósticos.

El MIT ha experimentado con *Deep Learning* para crear modelos para la lectura de mamografías, predicción del riesgo de cáncer de mama, evaluaciones mamográficas de densidad mamaria.

Su modelo para lectura de mamografías se denomina OncoNet, el cual fue desarrollado con Python y Matlab; implementaron una red neuronal convolucional utilizando una biblioteca de aprendizaje automático de código abierto de Python conocida como PyTorch.

El MIT para inicios del año 2021 empezaron a trabajar en un estudio muy prometedor, en el cual buscan predecir el cáncer de mama con antelación denominada Mirai MIT, con esta herramienta buscan que con base a una mamografía se pueda predecir si una persona tendrá cáncer, hasta con cinco años de antelación.

Como puede observarse, se han realizo múltiples herramientas e investigaciones sobre la detección y diagnóstico del cáncer de mama, estas herramientas facilitan los procesos y brindan un gran apoyo a profesionales de la salud que tienen acceso a ellas; sin embargo, es necesario poner al alcance público estas herramientas para que independientemente del lugar y clase económica, se pueda brindar el mejor servicio a los pacientes.

3. PLANTEAMIENTO DEL PROBLEMA

El cáncer es una enfermedad en la cual algunas células normales padecen de alguna anomalía y se transforman en células tumorales, estas crecen sin control y se esparcen alrededor del cuerpo humano. Esto puede ocurrir en casi cualquier órgano del cuerpo.

Según la Organización Mundial de la Salud -OMS- el cáncer es la principal causa de muerte alrededor del mundo, en 2020 se diagnosticaron 18 millones de casos nuevos y 10 millones de muertes a causa de esta enfermedad; dentro de ello el cáncer más común es el de mama diagnosticando 2,26 millones de casos nuevos y el quinto con mayor número de fallecimientos, atribuyéndole 685 mil defunciones (OMS, 2021).

El diagnóstico temprano del cáncer de mama aumenta significativamente la probabilidad de sobrevivir, además que su correcta clasificación asegura que el paciente no pasará por tratamientos innecesarios y que será tratado bajo un esquema adaptado a su caso.

Las variadas mutaciones del cáncer y la complejidad de esta enfermedad hacen que en ocasiones su diagnóstico sea complejo, según indica el UC David Helath, los falsos positivos y falsos negativos son comunes.

Según indica Breast Cancer Organization, en un estudio realizado en una organización de salud de Chicago con 741 150 casos, el 12.3 % fueron falsos positivos (Breast Cancer Organization, 2017). Así como la American Cancer

Society señala que una de cada ocho mamografías, tendrá un falso negativo (American Cancer Society, 2019)

Un falso negativo o un falso positivo, es perjudicial para los pacientes, en caso de un falso negativo esto representa tiempo crucial en el cual el cáncer seguirá creciendo sin ser tratado; mientras un falso positivo representa un tratamiento costoso e innecesario que representa en daños nocivos a la salud.

A partir de ello, podemos concluir que la detección de las células tumorales en las muestras de tejido y el análisis de este es importante para su diagnóstico. Por ello, en muchos países las herramientas de diagnóstico asistido -CAD- automatizan muchas de las tareas de los patólogos, brindando información complementaria al diagnóstico médico.

Gracias a los múltiples avances tecnológicos, se han buscado y desarrollado métodos que sean menos invasivos y a su vez proporcionen diagnósticos más acertados.

Uno de los proyectos pioneros más reconocidos es el sistema XCYT desarrollado en la Universidad de Wisconsin - Madison, este sistema a través del uso de *machine learning* proporciona el diagnóstico y pronostico del cáncer de mama basado en una muestra de biopsia con aspiración de aguja fina -PAAF-; sin embargo, este no se encuentra disponible en la web.

Los datos y resultados recopilados en el sistema XCYT han servido de base para el desarrollo de múltiples investigaciones, sin embargo, se necesitan de conocimientos técnicos computaciones para hacer uso de dichos sistemas e información. Un ejemplo de ello es el desarrollo de una librería en R basado en

el sistema XCYT que requiere tener conocimientos en programación para hacer uso de ella.

El Massachusetts Institute of Technology -MIT- desarrollo durante el año 2019 un modelo de aprendizaje profundo para la lectura de mamografías denominado OncoNet, a su vez, a inicios del año 2021, empezaron a desarrollar una inteligencia artificial para predecir el cáncer de mama denominada MIT Mirai.

Ambas tecnologías se encuentran en un repositorio público en Github, sin embargo, se necesita de conocimientos técnicos para desplegarlo en un ambiente local y utilizarlo.

Por ello, se plantea la siguiente pregunta central de investigación:

- ¿Cómo puede ponerse al alcance de cualquier persona las herramientas de *machine learning* para el diagnóstico y clasificación del cáncer de mama?

Y partiendo de la pregunta central, nos planteamos las preguntas auxiliares de investigación:

- ¿Cómo podemos extraer mediante *machine learning* información de una mamografía?

- ¿Qué herramientas y algoritmos de *machine learning* brindan una buena aproximación para determinar si un tumor es maligno o benigno en las pruebas de biopsia con aspiración de aguja fina?

- ¿Qué algoritmos de *machine learning* son más precisos para la clasificación de subtipos de cáncer de mama?

- ¿Cómo lograr una experiencia de usuario más acertada y funcional para los usuarios?

4. OBJETIVOS

4.1. General

Construir una herramienta publica y de código abierto para el diagnóstico y clasificación del cáncer de mama.

4.2. Específicos

1. Crear un modelo de *machine learning* para la lectura y extracción de información en mamografías.

2. Crear un modelo de aprendizaje supervisado para determinar si un tumor es maligno o benigno, basado en los resultados de una prueba de biopsia con aspiración de aguja fina comparando el uso de los modelos: Support Vector Machine (SVM), Decision Tree (c4.5), Naive Bayes (NB) y k Nearest Neighbors (k-NN).

3. Crear un modelo de aprendizaje no supervisado para la clasificación de los subtipos de cáncer de mama comparando el resultado de los algoritmos: K-means Clustering y Hierarchical Clustering.

4. Crear un repositorio público y entorno web intuitivo para el uso de la herramienta para el diagnóstico, clasificación y pronóstico del cáncer de mama.

5. JUSTIFICACIÓN

La línea de investigación en la cual se enfocará el trabajo es la minería de datos; la minería de datos es un proceso por el cual se busca extraer información de un almacén de datos con el objetivo de examinarla, identificar tendencias, patrones y relaciones; basándose en la información encontrada se busca tomar decisiones más precisas.

La minería de datos tiene aplicación en muchos sectores, desde marketing digital hasta medicina; dentro del campo médico, la minería de datos ha destacado al facilitar obtener diagnósticos y pronósticos más precisos a través de modelos predictivos.

La minería de datos ha ayudado a los médicos a identificar mejores tratamientos, por ejemplo: análisis de imágenes digitalizadas para detectar anomalías, análisis de las características de muestras, evaluación de la gravedad del paciente y brindar información para un plan de tratamiento más adecuado para el paciente (Dávila, 2012).

Dentro de los múltiples usos que se ha dado a la minería de datos dentro del campo médico se encuentra el diagnóstico del cáncer de mama, en donde se han utilizado algoritmos de *machine learning* para encontrar patrones, correlaciones y así obtener información de utilidad.

Sin embargo, pese a las múltiples investigaciones y trabajos que se han realizado, aún existen dos grandes problemas: Las herramientas son especialmente predominantes para el sector privado, por lo que solo personas

con un estatus económico alto pueden acceder a ellas; y a su vez los trabajos que se han desarrollado de forma pública requieren de conocimientos técnicos para poder desplegar y hacer uso de las herramientas.

A partir de esto podemos establecer que no todos los médicos tienen la oportunidad de utilizar las mismas herramientas ni de brindar el mismo servicio a los pacientes, esto repercute en que se utilicen más esfuerzos, tiempo y recursos en establecer un diagnóstico y plan de tratamiento para un paciente; que a su vez siendo el cáncer de mama una enfermedad tan compleja y múltiples variables, puede no darse el mejor diagnóstico ni tratamiento.

Por ello, se debe buscar, partiendo de las múltiples investigaciones y estudios, poner a disposición una herramienta óptima que esté al alcance de cualquier médico alrededor del mundo.

Entre los beneficios de esta herramienta está: el tener en una misma herramienta la capacidad de analizar mamografías y el analizar muestras de biopsias con aspiración de aguja fina, ambas funcionando con métodos eficientes y que muestren una mayor precisión; esto de forma gratuita

Además, se busca que la herramienta sea accesible por cualquier persona e intuitiva para el usuario, de forma que personas que no tengan conocimientos técnicos puedan utilizarla; así como albergar todo el proceso en un repositorio público para que personas que si cuenten con conocimientos técnicos puedan clonar el repositorio y puedan acoplarlo a sus necesidades.

Es importante contribuir con herramientas de este tipo para así lograr disminuir el tiempo y recursos que se utilizan para realizar estudios; a su vez

disminuir la brecha entre el sector público y privado, así cualquier médico podrá brindar diagnósticos más personalizados y certeros.

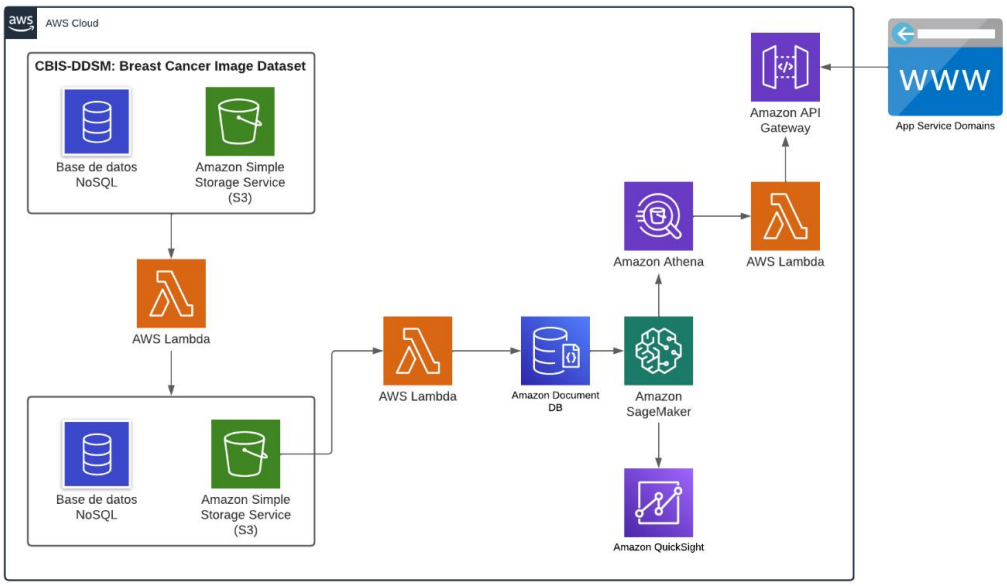
6. NECESIDADES A CUBRIR Y ESQUEMA DE SOLUCIÓN

La herramienta estará enfocada en dos necesidades principalmente: Optimizar el tiempo y recursos para el diagnóstico y clasificación del cáncer de mama; y brindar una herramienta pública e intuitiva para que cualquier persona alrededor del mundo pueda utilizar y brindar un mejor servicio a los pacientes.

La herramienta contará con tres módulos: Detección del cáncer de mama por la lectura de mamografías, detección del cáncer de mama por una muestra de aspiración con agua fina (PAAF) y la clasificación del cáncer de mama por una muestra de tejido tumoral.

En la figura 1, se presenta el esquema para la detección del cáncer de mama por la lectura de mamografías:

Figura 1. **Esquema para mamografías**



Fuente: elaboración propia mediante Lucidchart.

Para implementar la solución se construirá un modelo de *machine learning* utilizando como base el dataset Base de Datos Digital de Mamografía de Tamizaje (DDSM) que cuenta con 2620 estudios de mamografía; contiene casos normales, benignos y malignos con información patológica verificada.

Con el dataset se aplicará una función lambda que buscará disminuir el ruido, aplicando el filtro de la media que basado en estudios permite disminuir el ruido de una imagen, a su vez se aumentará el contraste para extraer variables más significativas; esto dará como salida un nuevo dataset con imágenes e información más confiable.

Con el nuevo dataset, se aplicará otra función lambda que buscará segmentar la imagen y clasificar si cada segmento es un tejido con microcalcificación maligna o es un tejido normal; partiendo de la clasificación, se determinarán los parámetros esenciales para determinar el diagnóstico.

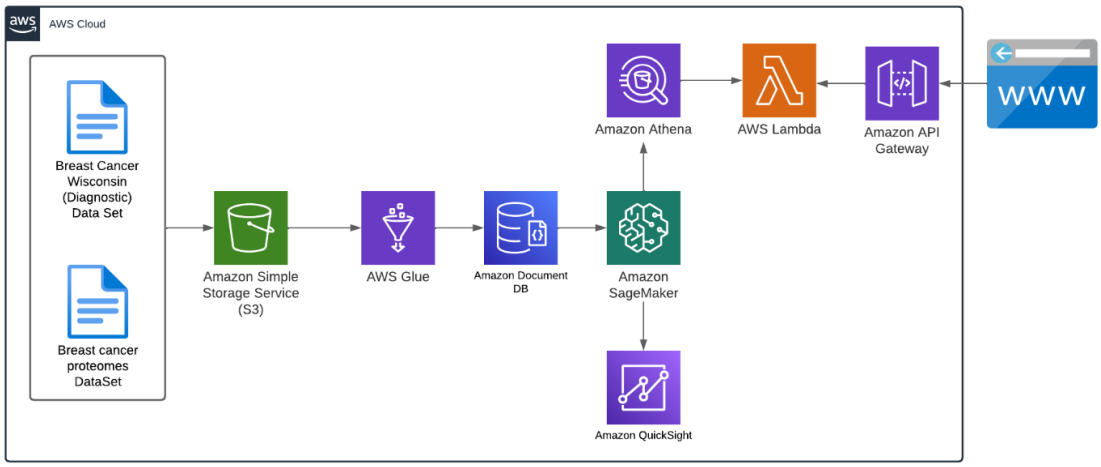
Los parámetros con la información esencial se guardarán en un documento en Amazon Document DB, con el cual se construirá el modelo utilizando Amazon SageMaker; para la construcción del modelo se utiliza el algoritmo Support Vector Machine (SVM).

Con el modelo construido, se utilizará Amazon Athena para acceder al modelo de SageMaker y obtener la inferencia; esto se expondrá mediante Amazon API Gateway que estará enlazado con una lambda que es la que hará uso de Amazon Athena.

Además, se implementaron Amazon QuickSight que brindará un dashboard personalizable con información sobre el modelo y su uso, para así obtener información relevante y más adelante hacer ajustes al modelo, buscando resultados más precisos.

Para la implementación de detección de cáncer de mama por muestras PAAF y la clasificación del cáncer de mama, se utilizará el esquema de la figura 2.

Figura 2. **Esquemas muestras PAAF y clasificación**



Fuente: elaboración propia mediante Lucidchart.

Para las muestras PAAF se utilizará como dataset de entrada el Breast Cancer Wisconsin Diagnostic y para la clasificación el Breast Cancer Proteones Dataset, ambos serán cargadas a un Amazon S3.

Se limpiarán ambos conjuntos de datos, utilizando AWS Glue y se guardará la información lista para procesar en un documento de Amazon Document DB, a partir de ellos se crearán los modelos con Amazon SageMaker.

Para las muestras PAAF, se creará un modelo supervisado, comparando el funcionamiento de los algoritmos: Support Vector Machine (SVM), Decision Tree (c4.5), Naive Bayes (NB) y k Nearest Neighbors (k-NN). Mientras que para la clasificación se utilizará un modelo no supervisado comparando el funcionamiento de los algoritmos: K-means Clustering y Hierarchical Clustering

Con ambos modelos, se utilizará Amazon Athena para acceder al modelo de SageMaker y obtener la inferencia; esto se expondrá mediante Amazon API

Gateway que estará enlazado con una lambda que es la que hará uso de Amazon Athena, cada modelo tendrá un endpoint distinto por el cual acceder.

Además, se implementará Amazon QuickSight para ambos modelos, brindando un dashboard personalizable con información sobre los modelos y su uso.

Para la construcción de la página web, se almacenará en un Amazon S3 como una web estática; el entorno web se construirá haciendo uso de Angular.

La solución hará uso de tecnología novedosa para su implementación, buscando ser óptima, rápida y sobre todo precisa.

7. ALCANCES

En este capítulo se especifica el alcance de la investigación en cuanto a su carácter descriptivo explicativo; desde la perspectiva técnica se presentan los métodos y procedimientos para la creación de los modelos y por último se detallan los productos tecnológicos que se brindarán.

7.1. Perspectiva investigativa

- Definir el proceso, metodología y variables para la lectura y extracción de información de mamografías.

- Definir las variables necesarias de una prueba de biopsia con aspiración de aguja fina para determinar si un tumor es maligno o benigno.

- Definir las variables necesarias para clasificar los subtipos de cáncer de mama.

7.2. Perspectiva técnica

- Diseñar un modelo de *machine learning* para la lectura y extracción de información en mamografías.

- Diseñar un modelo de aprendizaje supervisado óptimo para determinar si un tumor es maligno o benigno, tomando como base un conjunto de datos de una prueba de biopsia con aspiración de aguja fina.

- Diseñar un modelo de aprendizaje no supervisado óptimo para clasificar los subtipos de cáncer de mama.

- Diseñar y desarrollar el proceso de extracción, transformación y carga (ETL) para el análisis y predicción de la información.

- Implementar y comparar el uso de los modelos de aprendizaje supervisado: Support Vector Machine (SVM), Decision Tree (c4.5), Naive Bayes (NB) y k Nearest Neighbors (k-NN).

- Implementar y comparar el uso de los modelos de aprendizaje no supervisado: K-means Clustering y Hierarchical Clustering.

- Implementar un dashboard dinámico e intuitivo para el análisis, estudio y comprensión de los resultados de los modelos.

- Crear funciones serverless para interactuar y obtener las respuestas de los modelos.

- Desarrollar un entorno web público e intuitivo para el uso de la herramienta para el diagnóstico, clasificación y pronóstico del cáncer de mama.

- Crear un repositorio público con el código fuente para que otras personas puedan hacer uso de este.

7.3. Perspectiva de resultados

- *Dashboard* para el estudio y seguimiento de los modelos predictivos implementados utilizando el servicio AWS Quicksight.

- Entorno web para la interacción de los usuarios con los modelos predictivos mediante funciones serverless, la aplicación tendrá los siguientes módulos:

    - Módulo para la ejecución de la lectura y extracción de información de mamografías.
    - Módulo para la ejecución del modelo predictivo para determinar si un tumor es benigno o maligno.
    - Módulo para la ejecución del modelo predictivo para la clasificación del cáncer de mama.

- Repositorio público con el código fuente.

- Documento técnico con las especificaciones de la investigación, así como la metodología y procesos para la creación e implementación de los modelos.

8. MARCO TEÓRICO

8.1. Cáncer de mama

El cáncer de mama es una enfermedad maligna que se origina en las células de la mama, formando tumores. Es importante estudiarlo debido a su alta incidencia en mujeres y a su potencial para propagarse a otras partes del cuerpo, lo que puede tener graves consecuencias para la salud. La investigación y el estudio del cáncer de mama son fundamentales para mejorar su detección temprana, tratamiento y pronóstico, y para desarrollar estrategias de prevención más efectivas.

8.1.1. Definición, causas y diagnostico

Según indica el Instituto Nacional del Cáncer, el cáncer es una enfermedad en la cual células anormales o dañadas se forman, se multiplican sin control y se expanden a través del cuerpo; estas células anormales pueden formar tumores cancerosos o no cancerosos, también conocidos como malignos y benignos.

El cáncer de mama es el cáncer que comienza en los tejidos mamarios. Según indica la American Cancer Society, el cáncer de mama puede originarse en diferentes partes del seno, es más común que se origine en los conductos que llevan la leche hacia el pezón, algunos otros se originan en las glándulas que producen leche o en tejidos del seno.

El Centers for Disease Control and Prevention (CDC) y la American Cancer Society, dividen los factores de riesgo del cáncer de mama en dos categorías: Aquellas sobre las cuales las personas tienen influencia y aquellas que no.

Los factores de riesgo sobre los cuales no se tiene influencia son: el género (las mujeres tienen mayor probabilidad de padecerlo), antecedentes familiares, envejecimiento, hereditarios, raza y origen étnico, menopausia después de los 55 años y comienzo de los periodos menstruales a una edad temprana.

Por otro lado, algunos factores de riesgo sobre los que tenemos influencia son: consumo de bebidas alcohólicas (American Control Society señala que hay una correlación entre la cantidad de alcohol consumido y la probabilidad de padecer cáncer de seno), sobrepeso u obesidad, inactividad física, mujeres que no han tenido hijos.

Según la Organización Mundial de la Salud (OMS, 2021) durante el 2021 se diagnosticaron 2,26 millones de casos, a su vez es el quinto con mayor número de fallecimientos, atribuyéndole 685 mil defunciones.

El diagnóstico temprano aumenta significativamente la probabilidad de sobrevivir; las pruebas de detección consisten en revisar las mamas para detectar el cáncer en sus etapas iniciales.

8.1.2. Métodos de detección y diagnóstico para el cáncer de mama

El propósito de los exámenes de detección de cáncer de mama es encontrar signos de malignidad, antes de que el paciente presente algún síntoma, para así poder tratarlo a tiempo y aumentar la probabilidad del paciente de sobrevivir.

Si el examen de detección muestra algún síntoma anormal, se procederá a realizar exámenes de diagnóstico para determinar si tiene cáncer, el tipo del cáncer y el estado (en qué etapa se encuentra).

Según indica el CDC, entre los exámenes de diagnóstico para el cáncer de mama, se encuentran los siguientes: ultrasonido mamario, mamografías, imagen por resonancia magnética y biopsias.

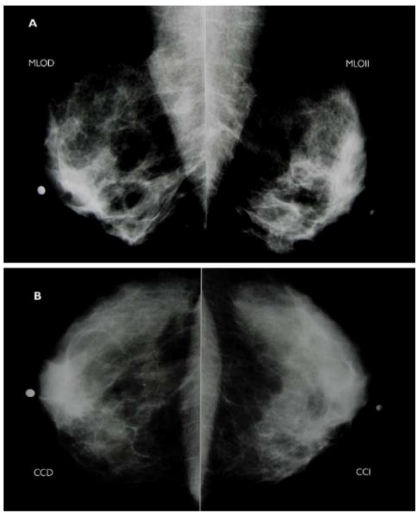
8.1.2.1. Mamografías

El CDC define a la mamografía como una radiografía de las mamas, esta es una imagen del interior de la mama tomada con rayos X. Las mamografías sirven para detectar signos de cáncer de mama principalmente en etapas iniciales.

La CDC indica que las mamografías son las mejores pruebas con las que cuentan los médicos para detectar anomalías en las mamas en etapas iniciales. En las mamografías de detección, se tomarán radiografías de cada seno desde dos ángulos diferentes (también conocida como mamografía bilateral) mientras en las de diagnóstico incluyen imágenes adicionales desde distintos ángulos (American Cancer Society, 2021).

En las mamografías de detección, las imágenes que se toman son una proyección craneocaudal (CC) y una medio-lateral-oblicua (MLO), la proyección CC permite evaluar el tejido mamario desde la parte inferior y superior, esta recoge la mayor información de la mama; por su parte, la MLO consigue mostrar desde la región axilar hasta el límite inframamario (Instituto Nacional del Cáncer de Argentina, 2012).

Figura 3. **Proyecciones medio-lateral-oblicua y craneocaudales**



Fuente: Brandan, M. E., & Villaseñor, Y. (2006). Detección del cáncer de mama: estado de la mamografía en México. Cancerología, 1(3), 147-62.

En una mamografía se buscan ciertas anomalías como:

- Microcalcificaciones: Depósitos de calcio en el tejido mamario, en la mamografía se observan como pequeñas manchas blancas; dependiendo de su tamaño, patrón de agrupación y apariencia pueden ser un indicador de malignidad (American Cancer Society, 2019).

- Masas. Área de tejidos mamario denso con una forma diferente al resto de la mama (American Cancer Society, 2019). Una masa puede ser con o sin calcificaciones, estas masas pueden ser causadas por quistes, afecciones mamarias benignas o cáncer de mama.

8.1.2.2. Pruebas de biopsia con aspiración de aguja fina

Una biopsia recolecta tejido de un posible tumor para que pueda examinarse con un microscopio; la prueba de biopsia con aspiración de aguja fina (PAAF) es un método mínimamente invasivo, en la cual se introduce una aguja hueca a través de la piel para succionar una muestra de células y líquidos (American Society of Clinical Oncology, 2022).

La muestra obtenida se envía a patología para ser examinada y determinar si el tumor es maligno o benigno, definición de grado histológico y determinar un pronóstico; esto se determina en base a características de la muestra como son el tamaño, si poseen calcificaciones, nódulos circunscritos, nódulos de contornos irregulares, entre otros (Hospital de Galdakao, 1997).

8.1.3. Clasificación del cáncer de mama

Actualmente, con los múltiples avances tecnológicos se pueden analizar los genes que tiene el cáncer de mama y definir una clasificación molecular para determinar el grupo del que forma parte, actualmente se dividen en cinco grupos: Luminal A, Luminal B (HER 2 negativo y HER 2 positivo), HER2 y Basal Like (Sociedad Española de Oncología Médica, 2023).

Esta clasificación ayuda a determinar características específicas del cáncer, riesgo de recaída de la enfermedad, tipo de tratamiento más beneficioso y severidad del cáncer. La siguiente tabla presenta una recopilación de datos por cada tipo:

Tabla I. **Clasificación molecular del cáncer de mama**

Subtipo	Comportamiento	Tratamiento
Luminal A	Positivo en receptores de estrógeno y progesterona, negativo para receptor de crecimiento epidérmico. Subtipo más común y menos agresivo. Asociado al incremento de edad.	Tratamientos hormonales y quimioterapia.
Luminal B	Positivo en receptor de estrógeno y de crecimiento epidérmico, negativo para receptor de progesterona. Similar a Luminal A pero es más agresivo.	Quimioterapia, tratamientos y terapia hormonales dirigido al receptor de crecimiento.
Basal	También se denomina cáncer de mama triple negativo. Negativos para receptor de estrógeno, progesterona y crecimiento epidérmico. Agresivo. Riesgo en edades menores a 40 años. Más frecuente en mujeres premenopáusicas afroamericanas.	Quimioterapia.
HER2	Negativo para receptor de estrógeno y progesterona, positivo para receptor de crecimiento epidérmico. Menos común y sumamente agresivo. Riesgo en mujeres menores a 40 años. Más frecuente en mujeres afroamericanas.	Quimioterapias y tratamiento dirigido al receptor de crecimiento epidérmico.

Fuente: Mayo Clinic, 2022 y Universidad Austral de Chile, 2011.

8.2. Procesamiento de imágenes digitales

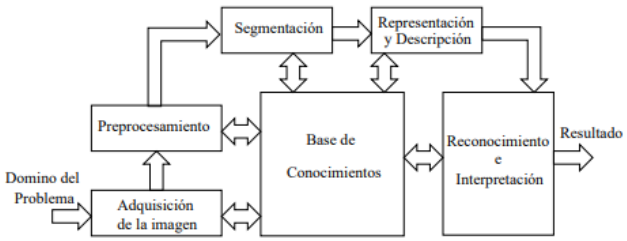
Una imagen digital se define como una función bidimensional, dada por f(x,y), en donde X y Y corresponden a coordenadas espaciales; y la amplitud o valor de la función f en un punto dado, se denomina como la intensidad de la imagen en ese punto (Woods & González, 2008).

El procesamiento de imágenes digitales consiste en una metodología en la cual, dada una imagen digital de entrada, se aplicarán diversas técnicas que darán como salida otras imágenes digitales o información de atributos extraídos de la imagen.

El procesamiento de imágenes digitales es usado en múltiples disciplinas, por ejemplo, en la medicina es utilizada para mejorar las imágenes de rayos X, dentro de la astronomía es utilizada para mejorar las imágenes tomadas por los satélites, también para reconocimiento de objetos dentro de una imagen, entre otros.

El procesamiento de imágenes sigue una serie de pasos que abarca desde obtener la imagen, el procesamiento, segmentación, representación y descripción, base de conocimiento, reconocimiento e interpretación.

Figura 4. **Pasos de procesamiento de imágenes**



Fuente: Fraga, L.G. (2011). Procesamiento digital de imágenes. Cinvestav. http://delta.cs.cinvestav.mx/~fraga/Charlas/procImagen.pdf

A continuación, se definen distintos elementos y técnicas utilizas para el procesamiento de imágenes.

8.2.1. Histograma de una imagen

Según define Richard E. Woods y Rafael C. González, el histograma de una imagen digital es el histograma construido por una función discreta; dada una imagen digital con intensidades de amplitud de [0, L − 1], definimos la función como:

$$h(r_k) = n_k$$

En donde:

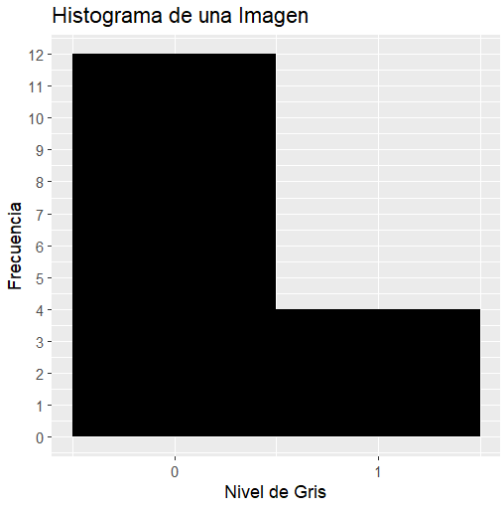
$k = nivel\ de\ gris, tal\ que: 0 < k < L - 1$
$r_k = k - ésimo\ nivel\ de\ gris$
$n_k = número\ de\ pixeles\ en\ la\ imagen\ con\ ese\ nivel\ de\ gris$

Por ejemplo, supongamos que la imagen está dada por la siguiente matriz (cada celda corresponde a un píxel, el valor dentro de la celda corresponde al nivel de amplitud):

0	0	0	1
0	0	1	0
0	1	0	0
1	0	0	0

El histograma estará dado por:

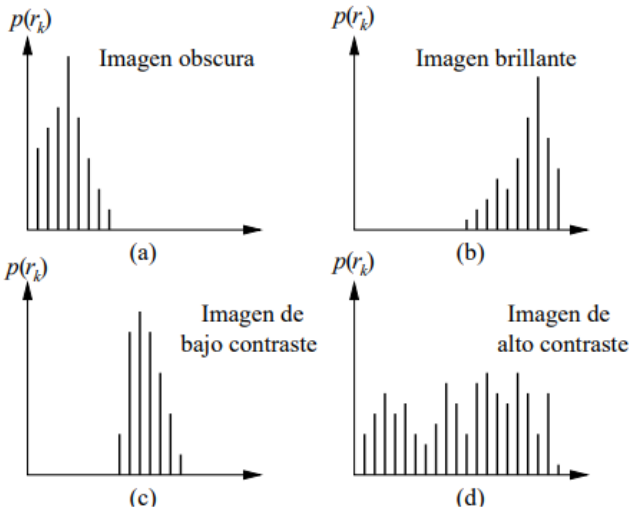
Figura 5. **Histograma de una imagen**



Fuente: elaboración propia mediante R studio.

El histograma ayuda a definir el tipo de imagen y facilita la aplicación de las diversas técnicas de procesamiento de imagen.

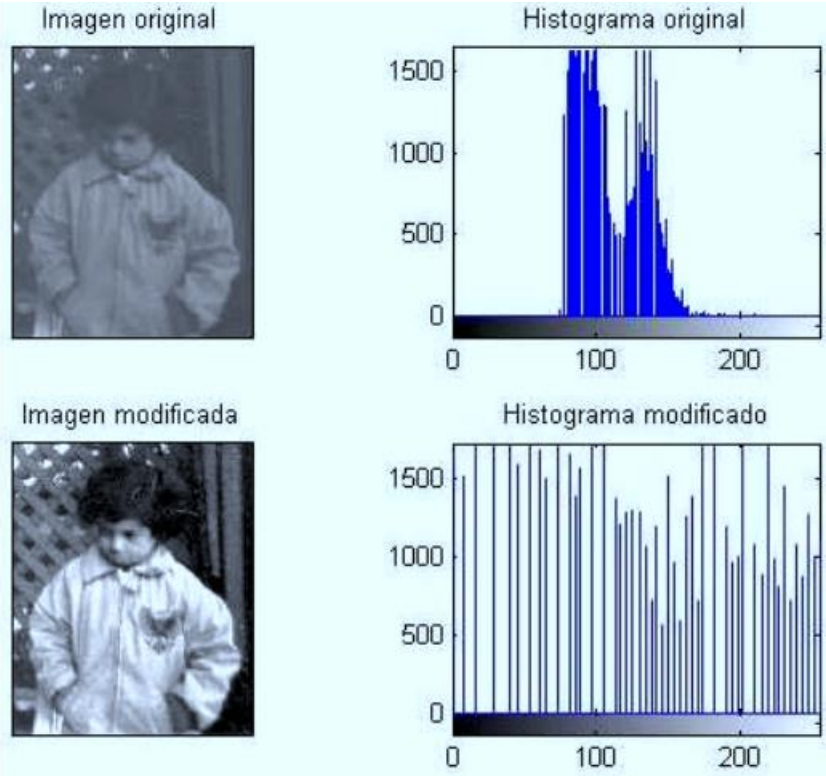
Figura 6. **Tipo de imagen según histograma**



Fuente: Fraga, L.G. (2011). Procesamiento digital de imágenes. Cinvestav. http://delta.cs.cinvestav.mx/~fraga/Charlas/procImagen.pdf

En base al histograma podemos definir las operaciones a realizar mediante funciones por partes para cada rango de intensidad, esto se ve ejemplificado en la Figura 7, en la cual la imagen original es una imagen de bajo contraste, por lo que para mejorarla se modifica el histograma de manera que sea una imagen de alto contaste.

Figura 7. **Histograma original e Histograma modificado**



Fuente: Esqueda, J. (2002). Fundamentos de Procesamiento de Imágenes. Instituto tecnológico de ciudad madero.

8.2.2. Aumento y reducción de contraste

En Fundamentos de procesamiento de imágenes de M. C. José Esqueda define que la modificación de contraste de una imagen consiste en aplicar una

función a cada píxel que compone la imagen para así obtener una nueva imagen, la función está dada de la siguiente manera:

$$p_{x,y} = f(po_{x,y})$$

En donde:

$x, y = coordenadas\ espaciales\ de\ la\ imagen$
$p_{x,y} = Nuevo\ nivel\ de\ gris\ en\ una\ coordenada\ espacial\ dada$
$po_{x,y} = Nivel\ de\ gris\ original\ en\ una\ coordenada\ espacial\ dada$
$f = Función\ aplicada\ al\ pixel\ original\ para\ obtener\ el\ nuevo\ nivel\ de\ gris$

Las transformaciones se realizan en base al número de bits de la imagen, para una imagen de 8 bits se toma como base los valores entre 0 a 255. Entonces, por ejemplo, la función inversa de una imagen estaría dada por

$$f(po_{x,y}) = 255 - po_{x,y}$$

En un artículo publicado de procesamiento de imágenes aplicadas a mamografías realizado en 2006 en la universidad tecnológica de Pereira, se resaltan algunas de las siguientes técnicas:

- Ampliación de contraste que consiste en incrementar la separación de nivel de grises del fondo y objetos.

- Ecualización del histograma en donde se modifica el histograma de la imagen de manera que el histograma tenga la forma deseada, se reasignan los rangos de nivel de gris de la imagen original de forma que la imagen resultante contenga una distribución uniforme de intensidades.

- Ecualización adaptativa de histograma: Es similar a la ecualización de histograma, pero se modifica la intensidad en base a cada rango de valores y calcula la nueva intensidad tomando el histograma de cada rango.

8.2.3. Eliminación de ruido

El ruido es información no deseada que contamina una imagen, el ruido puede clasificarse en dos tipos: Gaussiano (produce pequeñas variaciones en la imagen), sal y pimienta (el valor del píxel toma valores muy altos o bajos, es decir son negros o blancos).

Figura 8. **Tipos de ruidos**

Ruido Gaussiano

Ruido sal y pimienta.

Fuente: González, R.C., Wintz, P. (1996), Procesamiento digital de imágenes.

En el procesamiento de imágenes existen diversas técnicas para manejar el ruido sin eliminar características importantes de la imagen. En el artículo de Procesamiento de imágenes aplicadas a mamografías mencionado anteriormente, se resaltan algunas técnicas:

- Filtrado de mediana: Está técnica consiste en cambiar la intensidad de cada píxel con base a los pixeles vecinos que lo rodean, así los pixeles que tengan una intensidad muy distinta a comparación de los pixeles restantes, adquiriendo una intensidad similar.

- Filtrado espacial pasa bajo: Es similar al filtrado de mediana, pero en este se utiliza la media de los pixeles vecinos.

- Filtrado Gaussiano pasa bajo: En esta técnica se busca convertir la imagen original en una versión aproximada de la función Gaussiana, a la imagen se le aplica una máscara con una distribución Gaussiana y se calcula su desviación estándar, mientras más grande sea la desviación, se necesitará de una máscara más grande.

- Filtrado de Weinner: Esta técnica se basa en que cada píxel es el resultado de su valor original más el ruido, por lo que trata de minimizar ese ruido a cero. Para ello, se utilizan métodos estadísticos para realizar una estimación del error y así poder eliminarlo.

8.2.4. Filtros de obtención de contornos

Los filtros de obtención de contornos consisten en localizar los border de los objetos dentro de una imagen, esto ayuda a delimitar las secciones importantes y así reducir la cantidad de datos que se deben procesar (Universidad de Jaén, 2005).

Los filtros de obtención de contornos consisten en aplicar una máscara a la imagen, estas máscaras se conocen como operadores basados en el gradiente. Los operadores más utilizados son el operador de Robert (recomendado para

bordes diagonales, pero es muy sensible al ruido), operador de Prewitt (utiliza los píxeles vecinos para disminuir el efecto que puede causar el ruido), operador de Sobel (es más sensible a bordes diagonales, aunque no hay mayor diferencia con el de Prewitt), operador de Frei-Chen que aplica múltiples técnicas.

Sin embargo, el método más exitoso y utilizado, se denominada el método de Canny, que se divide en tres etapas: filtrado, decisión inicial y decisión final; durante la etapa de filtrado se busca los valores máximos de gradientes utilizando el filtro Gaussiano, posteriormente utiliza cuatro filtros para detectar bordes horizontales, verticales y diagonales; pasa a aplicar un operador de detección para obtener las estimaciones de los gradientes y se realiza una búsqueda para encontrar los pixeles que corresponden con esa magnitud de gradiente ((Quintana & Ojeda, 2011).

8.2.5. Segmentación

La segmentación es un proceso en el cual se divide una imagen en un conjunto de regiones según más convenga o con base en los elementos más representativos; algunas técnicas se basan en los filtros de obtención de contornos como son los modelos deformables o métodos basados en contornos activos, mientras otras técnicas utilizadas son:

- Umbralización: Este tipo de técnicas buscan obtener un valor de umbral que permite binarizar la imagen de forma que se pueda separar el objeto de interés del fondo. Los métodos de umbralización se construyen partiendo del histograma para distinguir la escala de gris del fondo con la del objeto (Cattaneo, 2011).

- Segmentación orientada a regiones: la imagen se divida en regiones en dependencia de las propiedades de cada píxel y su localización. Este tipo de segmentación tiene varios subtipos, entre ellos se encuentra el crecimiento de regiones en el cual se agrupan los píxeles adyacentes que presentan características similares al píxel examinado, de forma que una región va creciendo conforme se examinen los píxeles adyacentes; otro subtipo es la división y fusión de regiones en la cual se divide una imagen aleatoriamente en un conjunto de regiones, y luego estas regiones se agrupan con las regiones adyacentes si poseen propiedades similares (Palomino & Concha, 2009).

8.3. Machine Learning

Machine Learning es una rama de la inteligencia artificial en la cual se extraen patrones de un conjunto de datos, para así hacer que las máquinas aprendan y tomen decisiones sin estar explícitamente programadas para ello. (Raschka, 2019).

Machine Learning tiene múltiples utilidades, por ejemplo: para realizar análisis de riesgos, construir predicciones, clasificación de secuencias de ADN, comprensión de textos, análisis de imágenes, predicción de rutas, en marketing para análisis del mercado y comportamiento de consumo, entre otros.

8.3.1. Aprendizaje supervisado

En el aprendizaje supervisado se construyen modelos tomando como base un set de datos etiquetados que permiten hacer predicciones sobre datos futuros o no vistos. Es decir, dado un set de datos de entrenamiento, se crea un modelo el cual logrará predecir la respuesta (o salida del sistema).

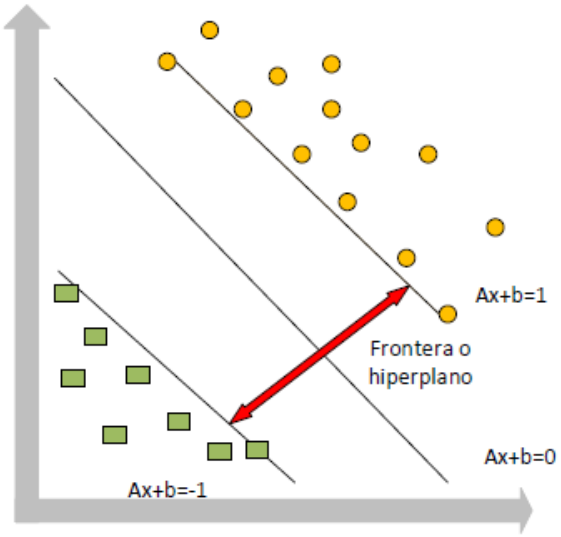
Este tipo se denomina “supervisado” porque ya se conocen las posibles salidas deseadas.

8.3.1.1. Support Vector Machine (SVM)

La máquina de soporte vectorial (SVM), es un algoritmo de clasificación múltiple de gran desempeño. SVM trabaja partiendo de dos o más set de datos etiquetados definido por un hiperplano óptimo que separa las clases (Huang, 2018)

En geometría un hiperplano es la generalización de un plano en varias dimensiones (Kuhn, M., & Johnson, K, 2013). El hiperplano ayuda a dividir el conjunto de datos según su etiqueta e intenta maximizar el margen entre su ubicación y el conjunto de datos.

Figura 9. **Hiperplano óptimo.**



Fuente: Betancour, 2005.

Cada punto se clasifica mediante una función de clasificación, en la cual se determina en qué categoría o conjunto de datos será agrupado ese punto en específico.

8.3.1.2. Decision Tree (c4.5)

Un árbol de decisión es un clasificador que busca predecir a qué clase o etiqueta pertenece un conjunto de datos partiendo de construcciones lógicas; el árbol se construye realizando varias iteraciones recursivas de partición binaria, en cada iteración se realiza una división en dos subconjuntos de datos a partir de decisiones asociadas a una variable; hasta un punto en que el proceso se detiene.

En el árbol que se forma cada nodo del árbol representa una pregunta acerca del atributo que se está examinando, cada posible respuesta a esta pregunta es uno de los nodos hijos del nodo examinado.

8.3.1.3. Naive Bayes (NB)

El modelo Naive Bayes se basa en el Teorema de Bayes en donde la probabilidad de un evento depende de la información existente sobre el mismo. En Naive Bayes se parte en que las variables son independientes entre sí (Raschka, 2019).

El Teorema de Bayes calcula el grado de probabilidad de que un evento A suceda, dado que ha ocurrido un evento B; partiendo que en un conjunto de sucesos mutuamente excluyentes (A, B), el teorema de Bayes se expresa mediante la siguiente ecuación:

$$P(A|B) = \frac{P(A)xP(B|A)}{P(B)}$$

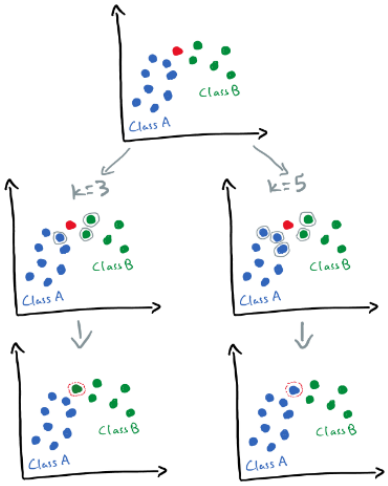
Para construir un modelo de Naive Bayes, se debe convertir el conjunto de datos en una tabla de frecuencias y utilizando el Teorema de Bayes se debe crear una tabla de probabilidades de que ocurran los diversos eventos; la clase que tenga la probabilidad más alta es el resultado de la predicción. (Dr. C. Regalado, 2009).

8.3.1.4. K-Nearest Neighbors (k-NN)

El método de k-Nearest Neighbors (k-NN) toma el set de datos como un montón de puntos marcados y los utiliza para etiquetas otro conjunto de datos no marcados; es decir, este método se basa en buscar similitudes entre casos conocidos para clasificar datos no conocidos (Raschka, 2019).

Los puntos que poseen similitudes entre sí se denominan “vecinos”; entre mayor sea la distancia entre los puntos, quiere decir que hay más diferencia entre ellos; existen distintos métodos para obtener esta distancia entre puntos, la más utilizada es la distancia euclidiana.

Figura 10. **k-Nearest Neighbors**



Fuente: Alizabeth, E. (2022). What K is in KNN and K-means.

La distancia euclidiana puede obtenerse mediante la siguiente formula:

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

En donde:

$$P_1 = Punto\ examinado\ con\ coordenadas\ x_1\ y\ y_1$$
$$P_2 = Punto\ vecino\ con\ coordenadas\ x_2\ y\ y_2$$

El punto que se está procesando o examinando, se colocará en el grupo que tenga más similitudes con este, es decir el grupo en el cual sus vecinos tengan una distancia menor.

8.3.2. Aprendizaje no supervisado

En el aprendizaje no supervisado, a diferencia del supervisado, no conocemos la salida de los datos o los datos no tienen una estructura conocida; así que se espera brindar al modelo una gran cantidad de datos de entrada y así

lograr que el modelo logre aprender y ajustar los resultados y agrupaciones mediante se utilice el modelo.

8.3.2.1. K-means clustering

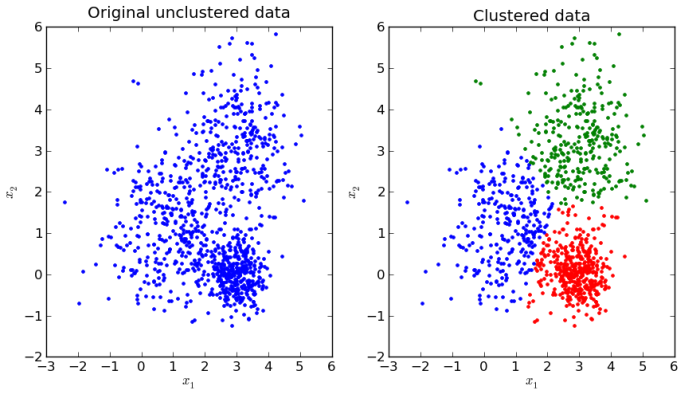
El método k-means clustering es un método de agrupamiento en el cual se agrupan objetos con base en sus características.

En el artículo “k-means++: The adventage of Careful Seeding” el autor Dadvid Author explica que para construir el modelo primero se debe especificar la cantidad de clústeres (grupos de objetos) que se quieren utilizar, la cantidad de clústeres se denomina mediante la variable k; cada clúster estará representado por una media aritmética que se denomina centro, está media se obtiene mediante los puntos que componen el grupo.

Una vez se conoce la cantidad de clústeres que se requieren implementar, se selección k puntos, cada uno de ellos representa uno de los clústeres (estos se toman como base para ir realizando iteraciones).

Posterior a ello se procesará cada punto restante y se asignará a uno de los clústeres, tomando como referencia el centro del clúster que coincida o esté más cerca al punto que se procesa; como se muestra en la Figura 11.

Figura 11. **K-means clustering**



Fuente; Arthir, D (2007). k-means++: The Advantages of Careful Seeding.

Una vez se han asignado todos los puntos a un clúster, se recalcula el centro ahora como la media aritmética de todos los puntos que componen el clúster. Y se repite el proceso nuevamente hasta que no haya cambios significativos entre los clústeres de cada iteración o bien hasta que se alcance un número máximo de iteraciones.

8.3.2.2. Hierarchical clustering

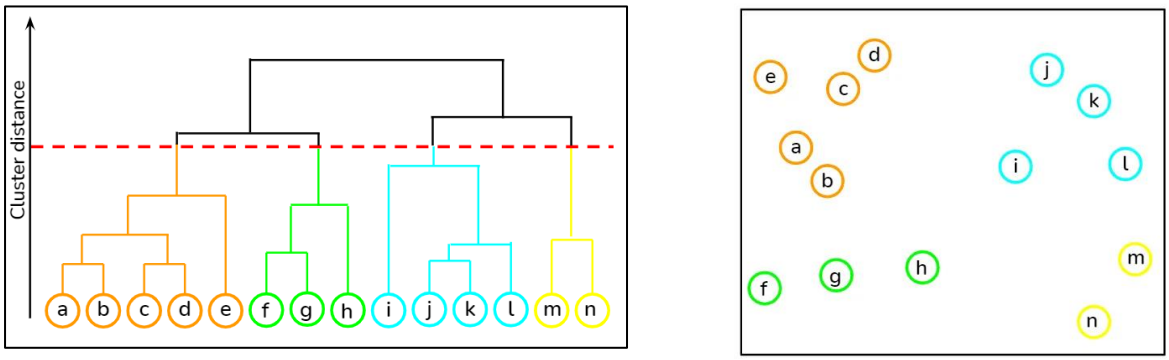
El método hierarchical clustering es un método de agrupamiento en el cual se busca construir jerarquías de grupos; este modelo no es muy utilizado en el mundo real.

Está técnica utiliza una representación gráfica denominada Dendrograma, Minitab la define como un diagrama de árbol que muestra los grupos que se forman al crear conglomerados de observaciones en cada paso y sus similitudes. (Raschka, 2019); se puede observar la estructura de un dendrograma en la figura 12.

Este algoritmo, en vez de dividir el conjunto de datos en clústeres, se toma el conjunto de datos como un solo grupo y por iteración se separan los puntos

diferentes del grupo principal; cada punto que se separa es considerado como un nuevo grupo, al final del proceso se obtienen N clústeres.

Figura 12. **Dendrograma, Hierarchical clustering**



Fuente: Pai Prasad (2021). Hierarchical clustering explained.

8.3.3. Métodos para validar modelos

Un paso importante en la creación de modelos de Machine Learning, es garantizar la exactitud de las predicciones del modelo; para ello se deben utilizar métodos para su validación

8.3.3.1. Cross-Validation

El método Cross-validation es un método para probar la exactitud y efectividad de un modelo de Machine Learning; a este método también se le denomina remuestreo y consiste en apartar una serie de datos de entrenamiento para posteriormente probarlo y ver si la predicción coincide con el resultado esperado.

8.4. Deep Learning

Deep Learning o aprendizaje profundo se considera como una rama de Machine Learning que consiste en entrenar redes neuronales artificiales con varias capas (Raschka, 2019).

8.4.1. Redes neuronales

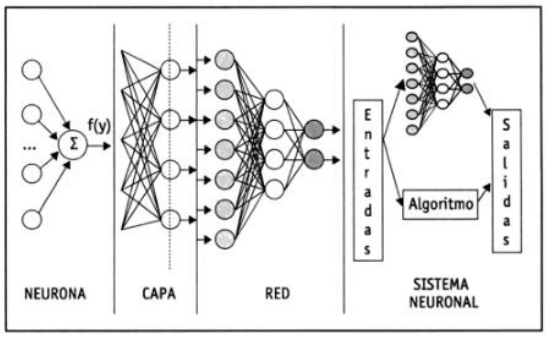
Deep learning combina muchas técnicas y procedimientos basados en Machine Learning; busca simular el comportamiento que realiza el cerebro humano a través de las neuronas.

La red neuronal se define como “Redes interconectadas masivamente en paralelo de elementos simples y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo que lo hace el sistema nervioso” (Kohonen, 1988).

8.4.1.1. Componentes de una red neuronal

La estructura de una red neuronal se muestra en la Figura 13.

Figura 13. **Estructura de una red neuronal**

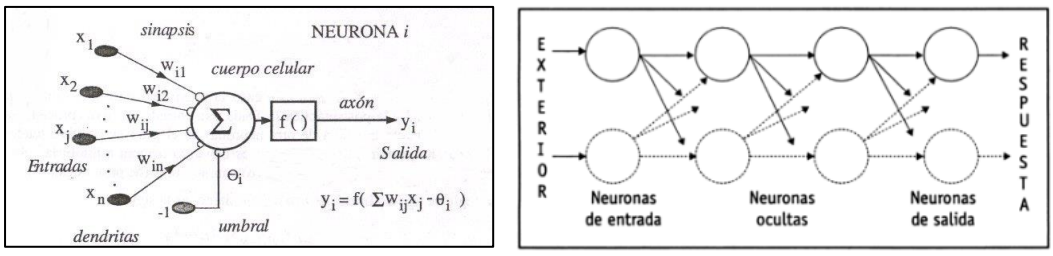


Fuente: Flores, R (2008). Las redes neuronales artificiales: Fundamentos teóricos y aplicaciones prácticas.

El elemento básico de las redes neuronales es la neurona artificial; una neurona artificial es un elemento de procesamiento que dado una entrada proveniente del exterior o de un estímulo recibido de otras neuronas, proporcionará una salida.

Las neuronas artificiales se dividen en tres tipos: Las de entrada que reciben señales provenientes del exterior, las de salida que envían la respuesta y las ocultas que reciben estímulos de otras neuronas y emiten una respuesta a otras neuronas. La forma en que se distribuyen estas neuronas se puede observar en la Figura 14:

Figura 14. **Neurona y tipos de neuronas artificiales**



Fuente: Flores, R (2008). Las redes neuronales artificiales: Fundamentos teóricos y aplicaciones prácticas.

La red neuronal a su vez tiene algunos elementos claves, como son: La entrada que reciben, el peso sináptico (favor de importancia asignado a la entrada que vienen de otras neuronas) una regla de propagación (la entrada, peso sináptico y operación de la célula ayuda a terminar el nuevo peso sináptico) y una función de activación (el valor de la regla de propagación pasa por una función que ayuda a encontrar la salida de la neurona) (McCulloch, 1943)

8.4.1.2. Tipos de redes neuronales

Existen diversos tipos de redes neuronales, se pueden clasificar con base a las características que poseen (Raschka, 2019); así se pueden clasificar en distintos grupos.

Por ejemplo, se puede clasificar por la cantidad de capas en monocapa en donde la red neuronal solo tiene la capa de entrada y salida, o multicapas en donde entre la capa de entrada y capa de salida, hay más capas ocultas.

También se pueden clasificar por el tipo de conexiones en no recurrentes, este no es tan utilizado porque no existe retroalimentación ni tienen memoria y las recurrentes en donde las neuronas en una misma capa o entre capas se pueden brindar retroalimentación, esto agrega memoria.

Se pueden clasificar por el grado de las conexiones que pueden estar totalmente conectadas, en estas todas las neuronas están conectadas entre sí y las parcialmente conectadas en donde no todas las neuronas están conectadas.

Existen otro tipo de redes, denominada como redes neuronales convolucionales (CNN) que se caracterizan porque cada capa de la red se puede entrenar para realizar diversas tareas. Es decir, cada capa tiene tareas

especializadas y cumplen con una jerarquía, así las primeras capas se dedican a realizar ciertas tareas y las siguientes se van especializando con la información ya recolectada.

8.4.1.3. Redes neuronales convolucionales para procesamiento de imagen

Las redes neuronales convolucionales son muy utilizadas para el procesamiento de imagen; en este tipo de procesamiento y siguiendo una jerarquía, las primeras capas se centran en detectar propiedades y formas básicas, y cada capa se va especializa hasta que las capas más profundas son capaces de reconocer formas complejas como rostros (Raschka, 2019).

9. PROPUESTA DE ÍNDICE DE CONTENIDOS

4.1.2.1. Mamografías

4.1.2.2. Pruebas de biopsia con aspiración de aguja fina

4.1.3. Clasificación del cáncer de mama

4.2. Procesamiento de imágenes digitales

4.2.1. Histograma de una imagen

4.2.2. Aumento y reducción de contraste

4.2.3. Eliminación de ruido

4.2.4. Filtro de obtención de contornos

4.2.5. Segmentación

4.3. Machine Learning

4.3.1. Aprendizaje supervisado

4.3.1.1. Support Vector Machine (SVM)

4.3.1.2. Decision Tree (c4.5)

4.3.1.3. Naive Bayes (NB)

4.3.1.4. K-nearest Neighbors (k-NN)

4.3.2. Aprendizaje no supervisado

4.3.2.1. K-means clustering

4.3.2.2. Hierarchical clustering

4.3.3. Métodos para validar modelos

4.3.3.1. Cross-validation

4.4. Deep Learning

4.4.1. Redes neuronales

4.4.1.1. Componentes de una red neuronal

4.4.1.2. Tipo de redes neuronales

4.4.1.3. Redes neuronales convolucionales para procesamiento de imagen

5. Presentación de resultados

5.1.1. Extracción, transformación y carga de datos (ETL)

10. METODOLOGÍA

10.1. Tipo de estudio

Se realizará una investigación cuantitativa, se analizarán grandes volúmenes de datos y se buscarán relaciones entre las variables que componen.

10.2. Diseño

La investigación será de caracteres experimental, se estudiará la forma en que se relacionan las variables con el resultado de los análisis con el fin de crear un modelo confiable.

10.3. Alcance

El alcance de la investigación es descriptivo explicativo porque busca describir y estudiar el cáncer de mama a través de sus propiedades y características; mientras que busca explicar las causas tomando en cuenta diversas variables.

10.4. Variables

Variables	Definición	Sub-variables	Indicadores
Estado de en una mamografía	Determina si una mamografía contiene microcalficaciones.	Métodos de procesamiento de imágenes utilizados. Precisión media. Porcentaje de aciertos en cross-validation. Tiempo de ejecución.	Histograma de la imagen. Niveles de gris (media local, desviación estándar, diferencia en intensidades). Simetría. Texturas.
Estado de benignidad en un tumor	Determina si un tumor es maligno o benigno basado en las características extraídas de una prueba de biopsia con aspiración de aguja fina.	Precisión media. Porcentaje de aciertos en cross-validation. Tiempo de ejecución	Radio, área y perímetro Textura Suavidad Compacidad Concavidad Simetría
Subtipo de cáncer de mama	Determina el subtipo de cáncer de mama al que pertenece un tumor con base en las proteínas presentes en un tejido tumoral.	Precisión media. Porcentaje de aciertos en cross-validation. Tiempo de ejecución. Estado para Luminal A Estado para Luminal B Estado para HER2	Tipos de proteínas presentes en las muestras tumorales. Cantidad de presencia de ciertas proteínas en las muestras tumorales.

Fuente: elaboración propia.

10.5. Fases del estudio

A continuación, se propone una serie de hitos que son cruciales para la realización del proyecto.

10.5.1. Extracción, transformación y carga de datos (ETL)

Esta fase consistirá en realizar un proceso extracción, transformación y carga de datos (ETL). En esta fase, se realizará una recolección de los datos tomando como fuente el Breast Cancer Wisconsin Diagnostic y el Breast Cancer Proteomes DataSet, posteriormente se limpiarán y transformarán los datos y por último se almacenarán para ser utilizados posteriormente.

10.5.2. Procesamiento de imágenes

En esta fase se procesarán las mamografías del dataset Breast Cancer Image Dataset (CBIS-DDSM); se realizará un procesamiento de imágenes aplicando diversos filtros como son la mediana, ecualización del histograma y el aumento de contraste para mejorar la calidad de esta, posteriormente se realizará una segmentación de la imagen de la cual se extraerán las características más importantes partiendo del histograma de nivel de gris, se obtendrá la intensidad, media de la intensidad promedio del histograma, varianza, sesgo, curtosis, etc. esta información se almacenará para ser utilizada posteriormente.

10.5.3. Creación y validación de los modelos

Con la información obtenida en los pasos anteriores, se procederá a crear los diversos modelos de Machine Learning, utilizando los modelos: SVM, Decision tree, NB, k-NN, K-means clustering y Hierarchical Clustering; se analizará la efectividad de los algoritmos de entrenamiento a través del método cross-validation para determinar el modelo óptimo para cada sección.

10.5.4. Construcción y pruebas de los servicios

Una vez construidos los modelos, se procederá a construir los servicios necesarios para hacer uso de los modelos de Machine Learning construidos. Se realizarán pruebas de los servicios para validar que la respuesta corresponda con la información esperada del modelo, además se medirá el tiempo promedio de respuesta de los servicios para medir su rendimiento.

10.5.5. Construcción de la interfaz web

Se construirá y pondrá a disposición un entorno WEB para que los usuarios puedan hacer uso de los modelos mediante una aplicación intuitiva.

10.5.6. Documentación

Por último, mediante un repositorio público de GIT se actualizará y compartirá todo el proceso realizado, con documentación sobre la construcción y validación de los modelos; así como el código fuente.

10.6. Técnicas de recolección de la información

Para la creación de los diversos modelos, se utilizarán dataset disponibles en internet.

- Breast Cancer Image Dataset (CBIS-DDSM): Base de Datos Digital de Mamografía de Tamizaje (DDSM) creada por el Hospital General de Massachusetts en la Universidad del Sur de Florida y los Laboratorios Nacionales Sandia, cuenta con 2620 estudios de mamografía; contiene

casos normales, benignos y malignos con información patológica verificada.

- Breast Cancer Wisconsin Diagnostic: Base de Datos Digital creada por el Departamento de informática y cirugía general de la Universidad de Wisconsin-Madison con información de muestras con aspiración de aguja fine de 569 pacientes, esta contiene 32 atributos relacionados con el núcleo celular.

- Breast Cancer Proteomes Dataset: Base de Datos Digital creada por Clinical Proteomic Tumor Analysis Consortium (CPTAC) con información de muestras de tejido tumoral de 77 pacientes; dichas muestras contienen información de la presencia de 12553 proteínas.

11. TÉCNICAS DE ANÁLISIS DE LA INFORMACIÓN

Diagrama de componentes: Se realizará un diagrama de componentes para plantear la arquitectura del diseño de la solución, así como la tecnología que se usará para construirla, en el diagrama se detallaran las relaciones entre los diversos componentes que forman el sistema.

Media: Para el procesamiento de las mamografías se utiliza la intensidad promedio del histograma de color y además se utilizará como una variable descriptiva como media local de intensidad de gris.

Desviación estándar: Se utilizará como variable descriptiva en el procesamiento de imágenes para obtener la desviación estándar local de la intensidad de gris.

Varianza: Se utilizará como variable descriptiva al estudiar el histograma de color de las mamografías, la varianza determinará el ancho del histograma.

Sesgo: Se utilizará como variable descriptiva al estudiar el histograma de color, servirá para determinar la simetría de la distribución alrededor de la media del histograma.

Curtosis: Se utilizará como variable descriptiva al determinar el grado de concentración de la distribución alrededor de la media en el histograma de color, esto ayudará a determinar la uniformidad en secciones de la mamografía.

Entropía: Se utilizará como variable descriptiva para determinar la medida de dispersión del histograma y determinar los niveles de contraste de la imagen.

Matriz de co-ocurrencia: Se utilizará como un método estadístico para determinar la ocurrencia de pixeles con ciertos niveles de gris, partiendo de ella se estimarán otros valores estadísticos.

Contraste: Se utilizará como una variable para estudiar un segmento de la imagen, el contraste se obtendrá obteniendo la media de contraste entre un píxel y sus vecinos.

Correlación: Se utilizará para medir la correlación entre un píxel y los pixeles vecinos tomando como base la matriz de co-ocurrencia.

Energía: Se utilizará como variable para medir la uniformidad de la imagen a través de la suma de los elementos al cuadrado de la matriz de co-ocurrencia.

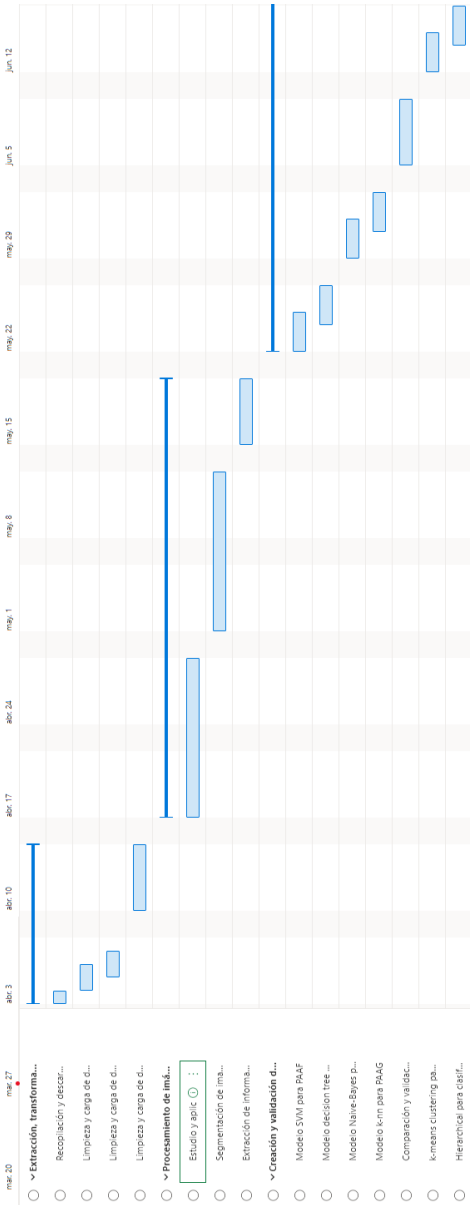
Histograma de nivel de gris: Se construirá un histograma de nivel de gris para cada mamografía este sé utilizará para el procesamiento de las mamografías, para determinar las demás variables y aplicar los filtros para poder eliminar el ruido.

Diagrama de correlaciones-dispersión: Se construirán diagramas de correlación-dispersión para estudiar la relación entre las variables de los sets de datos y los resultados, para así determinar las variables que tendrán más repercusión en el modelo.

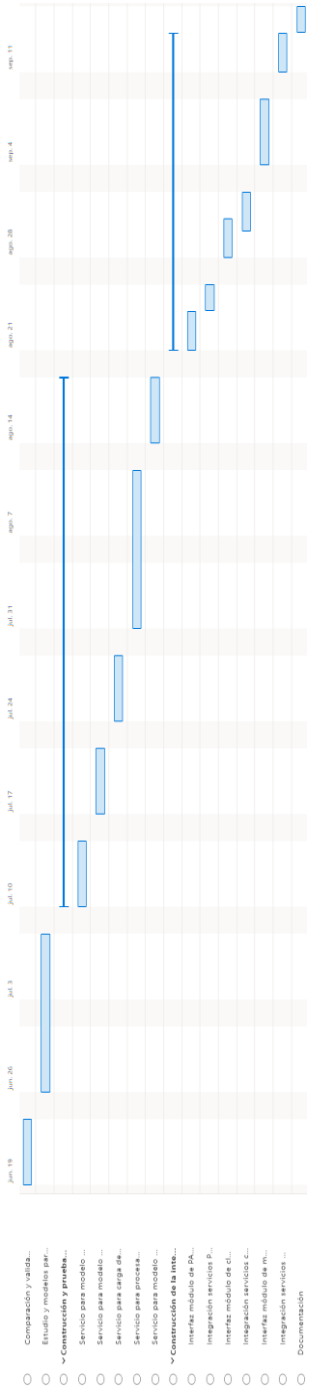
Mapa de calor: Se construirá un mapa de calor para estudiar la presencia de proteínas en las muestras de tejido tumoral.

12. CRONOGRAMA

Figura 15. **Cronograma**



Continuación Figura 15.



Fuente: elaboración propia mediante MS Project.

13. FACTIBILIDAD DEL ESTUDIO

A continuación, se detallan los distintos recursos necesarios para lograr el cumplimiento de los objetivos planteados, clasificándolos desde el punto de vista operativo, técnico y económico.

13.1. Factibilidad operativa

La implementación del sistema, así como su uso y operación, requiere de los siguientes componentes:

- Set de datos confiable que contenga los resultados de los análisis, para así crear los diversos modelos planteados y comprobar los resultados brindados por el modelo.
- Documentación técnica de investigaciones realizadas previamente.
- Disponibilidad de equipo de computación con acceso a internet.
- Disponibilidad a SET de datos de PAAF, muestras mamarias y mamografías.
- Documentación técnica y entorno web intuitivo para los usuarios.

Conclusión: Con base a los requerimientos operativos, establecemos que hay datasets gratuitos de instituciones confiables en internet para crear y probar los diversos modelos y existe basta documentación técnica sobre los diversos hitos que componente el proyecto. Con base a lo anterior, se puede concluir que el estudio desde un punto de vista operativo es factible de ser realizado.

13.2. Factibilidad técnica

Para la ejecución de todos los procesos técnicos que permiten la implementación del sistema, se requiere:

- Cuenta de AWS con capa gratuita disponible.
- Cuenta en Azure devops.
- Equipo de computación con características similares o superiores a:
  - Procesador i5, 10ma generación
  - 8GB de memoria RAM
  - Disco de estado sólido de 500GB o disco duro de 1TB
- Equipo de computación con los siguientes programas:
  - IDE o editor de código fuente.
  - NodeJS
  - Python
  - Angular
  - Postman
  - Swagger
- Herramienta de diagramación.
- Herramienta para crear prototipos.
- Dominio WEB para el despliegue de la solución.

Conclusión: Con base en los requerimientos técnicos, detallamos que: AWS brinda una capa gratuita para hacer uso de las herramientas con un costo flexible; a su vez algunas de las herramientas son *open source* o bien tiene disposición para usarse de forma gratuita con limitaciones; se utilizará el equipo de computación de personal para desarrollar el proyecto, solo el dominio web requerirá de inversión. Con base a lo anterior, se puede concluir que el estudio

es técnicamente factible de realizar en el tiempo asignado y se cuenta con todos los componentes técnicos necesarios para su implementación.

13.3. Factibilidad económica

Se requiere y presupuestan los siguientes recursos económicos para el desarrollo del proyecto:

Tabla II. **Presupuesto**

Recurso	Descripción	Costo total
Asesor	Por asesoría del trabajo de graduación	Q 2,500.00
Equipo de computación	1 computadora con las siguientes características • Procesador i5, 10ma generación • 8GB de memoria RAM • Disco de estado sólido de 500GB o disco duro de 1TB	Q8,000.00
AWS S3	Almacenamiento de archivos y publicación de archivos estáticos para entorno WEB. *La capa gratuita brinda 5GB de almacenamiento gratuito, se estima la cantidad restante con un S3 Estándar con precios de almacenamiento mensual de $0.023 por GB.	Q30.00
AWS Lambda	Funciones Severless para procesar información y hacer uso de los modelos. *La capa gratuita brinda hasta 1 millón de peticiones sin cargo	Q0.00

Continuación Tabla II.

AWS Document DB	Servicio para almacenar la información limpia para construir los modelos *Se utilizará una instancia t3.medium en la cual la capa gratuita brinda 750 horas al mes sin cargo	Q0.00
Amazon SageMaker	Servicio para construir los modelos de Machine Learning. *Se utilizará una instancia ml.t3.mediun que brinda 50 horas al mez de entrenamiento y 125 horas de inferencias durante los primeros dos, se hace una estimación con los demás meses para la instancia con precio de $0,05.	Q150.00
Amazon QuickSlight	Servicio para crear un dashboard personalizado para analizar el uso de los modelos. *Se utiliza AWS Pricing Calculator para realizar una aproximación teniendo solo un autor y calculando solo a mes y medio ya que se utilizará solo en la última parte del proyecto.	Q200.00
Amazon Athena	Servicio para acceder al modelo en SageMaker y obtener la inferencia de los datos. *Se utiliza AWS Pricing Calculator para efectuar una aproximación, estableciendo 25 consultas al día y escaneando 25MB de datos.	Q10.00
Amazon API Gateway	Servicio para exponer los servicios. *La capa gratuita brinda 1 millón de llamadas a la API al mes gratuitas.	Q0.00

Continuación Tabla II.

AWS Glue	Servicio para limpiar los datos a ser procesados. *La capa gratuita brinda 1 millón de objetos almacenados y solicitudes realizadas al mes.	Q0.00
AWS CloudFront	Servicio de CDN para entorno WEB. *La capa gratuita brinda 1TB de transferencia saliente de datos.	Q0.00
AWS Route 53	Servicio de control de dominio. *Se utiliza AWS Pricing Calculator para estimar el alojamiento en una zona con un 1 millón de consultas estándar.	Q50.00
AWS Certificate Manager	Servicio para crear certificados SSL. *Los certificados SSL/TLS son gratuitos.	Q0.00
Dominio WEB	Dominio WEB en donde se publicará la aplicación. *Se estima mediante los precios de GoDaddy.com	Q40.00
	Total	Q10,980.00

Fuente: elaboración propia

Conclusión**:** Con base en los requerimientos económicos, se estima que el costo total del proyecto ronda los Q11,000.00. Sin embargo, se hará uso del equipo de computación personal por lo cual no habrá que invertir en ello y el resto del costo (Q3000.00) será cubierto por el encargado de la investigación. Con base a lo anterior, se concluye que el proyecto es factible desde un punto de vista económico.

Con la información descrita anteriormente, se concluye que el proyecto es factible de forma operativa, técnica y económica.

14. REFERENCIAS

Álvarez G., Damián A.; Guevara G., (agosto, 2006) Preprocesamiento de imágenes aplicadas a mamografías digitales. Scientia Et Technica, vol. XII, núm. 31, pp. 1-6 Universidad Tecnológica de Pereira. Pereira, Colombia.

Artola Moreno, Á. (2019). Clasificación de imágenes usando redes neuronales convolucionales en Python.

Canales, J. C., Zhang, X. L., & Liu, W. Y. (2009). Clasificación de grandes conjuntos de datos vía Máquinas de Vectores Soporte y aplicaciones en sistemas biológicos. Instituto Politécnico Nacional, México.

Cáncer de mama. (2021, 26 marzo). Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer

Cattaneo, C. A., Larcher, L. I., Ruggeri, A. I., Herrera, A. C., & Biasoni, E. M. (2011). Métodos de umbralización de imágenes digitales basados en entropia de shannon y otros. Mecánica Computacional, 30(36), 2785-2805.

Elizondo, J. E., & Maestre, L. P. (2005). Fundamentos de procesamiento de imágenes. Mexicali: Universidad Autónoma de Baja California.

Fernández Regalado, R. (2009). El teorema de Bayes y su utilización en la interpretación de las pruebas diagnósticas en el laboratorio clínico. Revista cubana de investigaciones biomédicas, 28(3), 158-165.

Fine Needle Aspiration (FNA) of the Breast. (s. f.). Disponible en: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-biopsy/fine-needle-aspiration-biopsy-of-the-breast.html

Flores, R (2008). Las redes neuronales artificiales: Fundamentos teóricos y aplicaciones prácticas.

Gonzalez, R. C., & Woods, R. E. (2008). Digital Image Processing. Prentice Hall.

Hernández, J. G. (2017). Reducción de dimensionalidad en Machine Learning.: Diagnóstico de cáncer de mama basado en datos genómicos y de imagen. Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universitat Politècnica de València. Disponible en: https://riunet.upv.es/bitstream/handle/10251/92565/GALARZA%20-%20Reducci%c3%b3n%20de%20dimensionalidad%20en%20Machine%20Learning.%20Diagn%c3%b3stico%20de%20c%c3%a1ncer%20de%20mama%20basado%20e.%20.%20..pdf?sequence=1&isAllowed=y

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. Cancer genomics & proteomics, 15(1), 41-51.

Lavanya, D. & Usha Rani, K. (2011). Analysis of feature selection with classfication: Breast Cancer Datasets (Vol. 2). Indian Journal of Computer Science and Engineering. Disponible en: http://ijcse.com/docs/INDJCSE11-02-05-167.pdf

Lopez, R. F., & Fernandez, J. M. (2008). Las Redes Neuronales Artificiales. Netbiblo Sl.

Instituto Nacional del Cáncer de Argentina. (2012). Manual operativo para el uso de Mamografía en tamizaje. Ministerio de Salud de la Nación.

Milan, J. A. & Robles, J. B. (2020). Modelo en machine learning para el diagnóstico del cáncer de mama. Disponible en: https://repository.udistrital.edu.co/bitstream/handle/11349/25070/RoblesFajardoJaimeBrandon2020.pdf?sequence=1&isAllowed=y

Nick Street, W. (1994). Cancer Diagnosis and prognosis via linear-programming-based machine learning. University of Wisconsin Madison. Disponible en: https://minds.wisconsin.edu/bitstream/handle/1793/64580/94-14.pdf?sequence=1&isAllowed=y

Palomino, N. L. S., & Concha, U. N. R. (2009). Técnicas de segmentación en procesamiento digital de imágenes. Revista de investigación de Sistemas e Informática, 6(2), 9-16.

Pedraza, C. M. (2015). Identificación asistida de microcalcificaciones malignas en mamografías de tamizaje: Implementación de técnicas de Machine Learning para la identificación asistida de lesiones tumorales en imágenes médicas. Universidad de Los Andes. Disponible en: https://repositorio.uniandes.edu.co/bitstream/handle/1992/18449/u721814.pdf?sequence=1

Pérez-Marrero C, Vázquez-Romaguera T, Mulet-De-Los-Reyes A, Vázquez-Seisdedos C, Perdigón-Romero F. HistoBCAD: herramienta de código

abierto para detección de cáncer de mama en imágenes histopatológicas. Medisur [revista en Internet]. 2022 [citado 2022 Nov 12]; 20(2):[aprox. 11 p.]. Disponible en: http://medisur.sld.cu/index.php/medisur/article/view/5371

Quintana, C., & Ojeda, S. (2011). Mammography image detection processing for automatic micro-calcification recognition. Spatial Statistics and Image Modeling, 2(2), 69-79. https://www.soche.cl/chjs/volumes/02/02/Quintana_etal(2011).pdf

Raschka, S., & Mirjalili, V. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2, 3rd Edition. Van Haren Publishing.

Women Who Receive False Positive Mammogram Results May Be More Likely to Delay Next Screening. (s. f.). Disponible en: https://www.breastcancer.org/research-news/false-positives-may-lead-to-screening-delays