UNIVERSIDAD NACIONAL DEL ALTIPLANO

FACULTAD DE INGENIERÍA MECÁNICA ELÉCTRICA, ELECTRÓNICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

MODELO DE MACHINE LEARNING USANDO UN CLASIFICADOR DE MÁQUINAS DE SOPORTE VECTORIAL PARA LA DETECCIÓN Y CLASIFICACIÓN DEL CÁNCER DE SENO USANDO IMÁGENES MAMOGRÁFICAS

TESIS

PRESENTADA POR:

Bach. CRISTHIAN WILSSON LAUREANO YUPANQUI

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

PUNO – PERÚ

2022

DEDICATORIA

Primeramente, agradecer a Dios por darme la oportunidad de vivir cada día, por tener a mi familia y seres queridos con buena salud, por todas las bendiciones y oportunidades que me brinda para ser el profesional en el cual quiero convertirme.

A mi querida madre Olga Alicia Yupanqui Alanoca, por todo el apoyo emocional, sus consejos y esfuerzos para que pueda ser una persona de bien, y por ser ella el pilar fundamental en mi formación profesional.

A mi padre Ricardo Laureano Nina y mi hermana Diana Laureano Yupanqui, por todo su apoyo moral durante mi formación profesional.

Cristhian Wilsson Laureano Yupanqui

AGRADECIMIENTOS

Mi eterno agradecimiento a la Universidad Nacional del Altiplano por haberme acogido en sus aulas en el proceso de mi formación académica, a los docentes de la Escuela profesional de Ingeniería de Sistemas, quienes compartieron sus conocimientos y experiencias para nuestra formación profesional durante los 10 semestres académicos.

A mi asesor de tesis el Dr. Edelfré Flores Velásquez por todo su apoyo incondicional y haber compartido sus conocimientos para la elaboración de este proyecto de investigación.

A los miembros del jurado M.Sc. William Eusebio Arcaya Coaquira, D.Sc. Donia Alizandra Ruelas Acero y M.Sc. Lenin Huayta Flores por sus aportes y sugerencias en el proceso de investigación.

Agradezco al Hospital Regional "Manuel Núñez Butron" Puno y al departamento de diagnóstico por imágenes por todo su apoyo en la adquisición del material de estudio que pudo hacer posible la finalización de este proyecto.

Cristhian Wilsson Laureano Yupanqui

ÍNDICE GENERAL

Área : Inteligencia artificial

Tema : Modelo de machine learning y máquinas de soporte vectorial

FECHA DE SUSTENTACIÓN: 09/Febrero/2022

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ACRÓNIMOS

ANN : Artificial Neural Networks

BCDR : Breast Cancer Digital Repository

BI-RADS : Sistema de datos e informes de imágenes mamarias.

DDSM : Digital Database for Screening Mammography

DT : Decision Tree

HRMNB : Hospital Regional Manuel Núñez Butron

INEI : Instituto Nacional de Estadística e Informática

KNN : K-Nearest Neighbors

MIAS : Mammographic Image Analysis Society

MINSA : Ministerio de Salud

NB : Naive Bayesian

OMS : Organización Mundial de la Salud

OPS : Organización Panamericana de la Salud

RBF : Radial Basis Function

ROI : Region Of Interest

SVM : Support Vector Machine

RESUMEN

Este proyecto de investigación de tesis abordó el tema de la detección y clasificación del cáncer de seno a través de uso de un modelo de Machine Learning (ML); debido a los incrementos de los casos de cáncer de mama en los últimos años en el Perú, es necesario contemplar la búsqueda de nuevas alternativas tecnológicas para el apoyo en las tomas de decisiones. El objetivo planteado para esta investigación fue el desarrollo de un modelo de ML que utilice un clasificador de Máquinas de Soporte Vectorial (SVM) el cual clasifique el cáncer de seno por medio del uso de imágenes mamográficas. Se planteó el uso de una metodología con un enfoque de tipo cuantitativo con el diseño de investigación de tipo exploratorio y cuasi experimental. Para el desarrollo del proyecto de investigación se utilizó los dataset MiniMIAS, DDSM, BCDR e imágenes del Hospital Regional "Manuel Núñez Butron" Puno, se consideró las etapas de preprocesamiento, extracción de características y clasificación para el entrenamiento del modelo de ML. Para el clasificador SVM se seleccionó el kernel de función de base radial (RBF) con los parámetros de $C = 100$ y $\gamma = 2^{-9}$. Los resultados obtenidos a la propuesta del modelo de ML, se consiguió un 90% de exactitud (accuracy) y las métricas de precisión, sensibilidad y f1 es desarrollada por cada clase respectivamente, para benigno (88.6%, 83.8%, 86.1%), maligno (83.9%, 96.3%, 89.7%) y la clase normal (94.3%, 89.2%, 91.7%); para las pruebas del modelo entrenado a través del uso de otros dataset se obtuvieron que DDSM resulto con un 89%, BCDR obtuvo un 84% y HRMNB un 83%. Las conclusiones obtenidas para el proyecto de investigación se afirma que la propuesta del modelo de ML logró realizar la detección y clasificación del cáncer de mama de manera satisfactoria con un grado de exactitud aceptable.

Palabras Clave: Aprendizaje automático, cáncer de seno, clasificación, SVM

ABSTRACT

This thesis research project addressed the topic of breast cancer detection and classification through the use of a Machine Learning (ML) model; Due to the increases in breast cancer cases in Peru in recent years, it is necessary to consider the search for new technological alternatives to support decision-making. The objective set for this research was the development of a ML model that uses a Support Vector Machine (SVM) classifier which classifies breast cancer through the use of mammographic images. The use of a methodology with a quantitative approach with an exploratory and quasi-experimental research design was proposed. For the development of the research project, the MiniMIAS, DDSM, BCDR dataset and images of the "Manuel Núñez Butron" Puno Regional Hospital were used, considering the stages of preprocessing, feature extraction and classification for the training of the ML model. For the SVM classifier, the radial basis function (RBF) kernel was selected with the parameters of C=100 and $\gamma = 2^{-9}$. The results obtained from the ML model proposal, 90% accuracy was achieved and the precision, sensitivity and f1 metrics are developed for each class respectively, for benign (88.6%, 83.8%, 86.1%), malignant (83.9%, 96.3%, 89.7%) and the normal class (94.3%, 89.2%, 91.7%); For the tests of the model trained through the use of other datasets, it was obtained that DDSM resulted in 89%, BCDR obtained 84% and HRMNB 83%. The conclusions obtained for the research project state that the ML model proposal managed to detect and classify breast cancer satisfactorily with an acceptable degree of accuracy.

Keywords: Breast cancer, classification, machine learning, SVM.

CAPÍTULO I

INTRODUCCIÓN

El cáncer es una de las enfermedades con mayores afecciones a nivel mundial y catalogada como una de las principales causas de muerte según la OMS, el cáncer de seno es la enfermedad neoplásica con más afecciones en la población femenina, para su detección en sus primeras etapas se recurre a la aplicación de técnicas como la mamografía, esta técnica reconoce células tumorales que son pequeñas y difícilmente palpables para distinguir el cáncer de mama.

A pesar del uso de técnicas para la prevención como la mamografía, la influencia de algunos elementos como el estado de las máquinas mamográficas y la experiencia de los especialistas definen el resultado de un posible falso positivo y a la realización de una biopsia innecesaria, en los últimos años la influencia de las máquinas de aprendizaje en las áreas de salud ha venido siendo un gran apoyo a los especialistas en las tomas de decisiones.

En las últimas décadas las máquinas de aprendizaje han venido tomando una gran popularidad por la habilidad que brinda a un computador el aprendizaje por medio de la experiencia, dándonos una herramienta que pueda procesar una gran cantidad de datos para dar soluciones a problemas complejos.

La siguiente investigación busca realizar la detección del cáncer de seno utilizando imágenes mamográficas a partir del desarrollo de un modelo de machine learning enfocado en el aprendizaje supervisado, para el cual se utilizará el modelo de tipo clasificación utilizando el algoritmo de máquinas de soporte vectorial, del mismo modo, se pretende implementar el modelo seleccionado e identificar el nivel de precisión

para descartar los casos de sobre ajuste o sub ajuste ya que estos hacen referencia si se encontrase algún fallo en nuestro modelo seleccionado. Para realizar el entrenamiento del modelo se utilizará el dataset Mini-MIAS el cual contiene un conjunto de imágenes mamográficas con una resolución predeterminada y las limitantes que se pueden encontrar en este dataset radican en la existencia de ruido en algunas de las imágenes, la aparición de etiquetas medicas en las mamografías y en algunos casos las calcificaciones son distribuidas en distintas partes en lugar de concentrarse en un solo sitio.

El proyecto de investigación esta segmentado en cuatro capítulos:

En el *Capítulo I* se muestra el planteamiento del problema el cual desarrolla la problemática del cáncer de seno, así como la justificación, planteamiento de los objetivos, limitaciones e hipótesis de la investigación.

En el *Capítulo II* se presenta los antecedentes del estudio, fundamentación teórica, proporcionando referencias a investigaciones relevantes y las definiciones conceptuales.

En el Capítulo III se establece el enfoque y diseño de la investigación, la población y muestra, la adquisición del dataset, el material experimental utilizado y las métricas utilizadas para la evaluación del modelo.

En el *Capítulo IV* se presenta las conclusiones del proyecto de tesis en base a los objetivos, así también las recomendaciones y trabajos futuros.

1.1. Planteamiento del Problema

El cáncer hace referencia a un conjunto de enfermedades que puede originarse en diferentes partes del cuerpo a causa del crecimiento anormal de las células. Así mismo, en 2018 se tuvo una tasa de mortalidad a nivel mundial de aproximadamente de 9.6 millones de personas que fallecieron a causa de esta enfermedad (OMS, s.f.). La Organización Mundial de la Salud (OMS) declara a esta enfermedad como un problema de salud pública (Cazap et al., 2010) y se estima que para el 2040 el incremento de los casos de cáncer será aproximadamente de 30 millones de nuevos casos, donde el mayor crecimiento se verá en los países de ingresos bajos y medios (OPS, s.f.).

La aparición de esta enfermedad como es el cáncer de seno puede tomar un tiempo prolongado y no lograr presentar algún síntoma, esta enfermedad suele originarse en el tejido glandular de los senos (ACS, 2019). El cáncer de seno es considerada mortal por la baja tasa de supervivencia a causa de un diagnostico tardío, la realización de un diagnostico suele ser complicado por los recursos que este procedimiento implica (Anvesh, 2014).

A nivel mundial, la OMS en 2020 indicó que el cáncer de seno es la principal enfermedad neoplásica que afecta a las mujeres, así mismo, se registran cerca de 2.2 millones de casos donde 1 de cada 12 mujeres padece de esta enfermedad, además se registró un total de 685 mil fallecimientos a causa de esta enfermedad (OMS, 2021).

En América Latina se refleja lo que acontece a nivel mundial, aproximadamente 152 mil mujeres padecen esta enfermedad y 43 mil fallecen a causa de ella (OPS, 2016), en 2020 se registró que alrededor de 37 mujeres de cada 100 mil fallecen por el cáncer de seno (IARC, 2021).

La Agencia Internacional de Investigación en Cáncer (IARC) indica que la incidencia de los casos de mujeres en las edades de 45 y 69 años es de 89 por cada 100 mil, mientras que la tasa de mortalidad es de 22 por cada 100 mil. El Ministerio de Salud (MINSA) indica que el número de mujeres entre las edades de 50 a 64 años que se realizaron pruebas para la detección del cáncer de seno es bajo, resultando que la enfermedad llegue a detectarse en etapas muy avanzadas (MINSA, 2017). Según los informes presentados por el Instituto Nacional de Estadística e Informática (INEI), las mujeres que se realizaron algún examen físico de mama, el 23,9% vinieron de zonas urbanas y solo un 11,3% provienen de zonas rurales (INEI, 2018).

Debido al incremento de la demanda a nivel mundial y local para la detección temprana del cáncer de mama en los centros médicos ha ocasionado la búsqueda de nuevas rutas para nuevas investigaciones (Krithiga & Geetha, 2021). El método más eficaz para la detección del cáncer es por medio de una mamografía (Orozco et al., 2020), los riesgos que conlleva un diagnóstico incorrecto podría ocasionar la realización de una biopsia innecesaria (Hepsağ et al., 2017); el impacto de una biopsia innecesaria trae consigo un conjunto de factores que logra incrementar los costos de una atención médica, por esta razón es necesario la búsqueda de nuevas alternativas para la reducción de resultados de falsos positivos en los diagnósticos (Hamidinekoo et al., 2018).

Las evaluaciones médicas para la detección del cáncer de mama que utilicen imágenes mamográficas pueden variar según el especialista que lo realice por ser una apreciación cualitativa (Krithiga & Geetha, 2021), para evitar las biopsias innecesarias es importante la incorporación de métodos que utilicen el diagnóstico asistido por computadora para que los especialistas de salud tomen decisiones más precisas (Hepsağ et al., 2017). En los últimos años, los algoritmos de machine learning han venido

revolucionando como herramientas de enorme potencial para el desarrollo de nuevos métodos para los diagnósticos asistidos por computador, así también por la gran capacidad que esta herramienta presenta para el procesamiento de una gran cantidad de datos para dar solución a problemas complejos. Los métodos de aprendizaje automático utilizando Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés) están logrando obtener grandes resultados en los temas de clasificación, ampliamente utilizados en los campos de procesamiento de imágenes y visión artificial por el nivel de exactitud que se pueden obtener.

1.2. Justificación del Problema

Una adecuada detección y tratamiento en las etapas iniciales de esta enfermedad son factores importantes para la recuperación de los pacientes (INEI, 2018; MINSA, 2017; Pinillo, 2016). La técnica más confiable para el descarte del cáncer de mama es mediante la mamografía, con este método el especialista puede observar las distintas anormalidades que se puedan presentarse (OPS, 2016), los resultados de los diagnósticos que son considerados como falsos positivos es en consecuencia a la diferenciación del especialista al analizar las zonas con lesiones y las zonas sanas (Orozco et al., 2020).

Por el permanente incremento de los casos de cáncer y la necesidad de acoger nuevas tecnologías para el apoyo en las tomas de decisiones en las áreas de salud, es imprescindible el desarrollo de sistemas asistidos por computadores y las técnicas que son relacionadas a estas son objetos de estudio (Orozco et al., 2020), este tipo de sistemas apoyara en la toma de decisiones a los especialistas de salud, disminuyendo los casos de falsos positivos y posibles biopsias innecesarias.

SVM el cual es uno de los algoritmos ampliamente utilizados para temas de clasificación; es aplicado tanto en ámbitos académicos como en las áreas de

bioinformática, marketing, industrial, etc. La importancia de un modelo que utilice este tipo de clasificadores para el apoyo en la toma de decisiones de los especialistas encargados, podrá disminuir los recursos médicos y económicos hacia los pacientes.

1.3. Objetivos de la investigación

1.3.1. Objetivo General

Desarrollar un modelo de machine learning utilizando un clasificador de máquinas de soporte vectorial para detectar y clasificar el cáncer de mama a partir de imágenes mamográficas.

1.3.2. Objetivo Específicos

- Analizar los modelos de machine learning y el clasificador de máquinas de soporte vectorial.
- Implementar el modelo de machine learning utilizando el clasificador de máquinas de soporte vectorial.
- Identificar el nivel de exactitud del modelo de machine learning con el clasificador SVM para la detección y clasificación del cáncer de mama.

1.4. Alcances y limitantes

1.4.1. Alcances

El alcance de este proyecto de investigación es el desarrollo del modelo de machine learning para el diagnóstico del cáncer de mama de una manera confiable y aceptable, analizando y comprendiendo el algoritmo SVM para su aplicación, esto modelo de ML propuesto involucra dataset proporcionados por bibliotecas digitales, comunidades científicas, así también el uso de imágenes de un entorno real.

1.4.2. Limitantes

A pesar de la gran efectividad que puede otorgar la utilización de las máquinas de soporte vectorial a la hora de realizar la clasificación, existe un riesgo de la complejidad en la dimensionalidad de los datos por la cantidad de imágenes que se pueda procesar. El tamaño de estas imágenes permitiría tener un mayor entrenamiento y validación para su clasificación como benigno o maligno.

Las posibles restricciones para la adquisición de imágenes de un centro médico local, no se prevé el estado y funcionamiento de las máquinas para la recolección de las imágenes de manera local para su extracción en una buena calidad.

Por otra parte, no se cuenta con un equipo de grandes capacidades para acelerar el entrenamiento del algoritmo, debido al tamaño de las imágenes que se tenga que procesar así mismo será el tamaño de los datos que se extraigan para su análisis, esto conlleva a que tenga largos periodos de entrenamientos para lograr obtener un modelo con un nivel de exactitud aceptable.

1.5. Hipótesis

El desarrollo del modelo de machine learning utilizando el clasificador de máquinas de soporte vectorial, es posible la detección y clasificación del cáncer de mama a través de imágenes mamográficas.

CAPÍTULO II

REVISIÓN DE LITERATURA

2.1. Antecedentes de la investigación

A continuación, se presenta las diferentes investigaciones tanto a nivel nacional como internacional que están relacionadas con el problema de investigación.

2.1.1. A nivel nacional

Calderón (2012) a través de su investigación, se planteó como objetivo general el desarrollo de un método que permita la detección de metástasis de cáncer mamario a partir de datos de microarrays. La utilización de las máquinas de soporte vectorial como algoritmo de clasificación, el autor pretendió el mejoramiento del performance del algoritmo utilizando múltiples kernels (MKL) el cual involucró kernels lineales y no lineales. El autor evaluó el desempeño del modelo propuesto basado en MKL y SVM con distintos kernels y combinaciones para las clasificaciones binarias de microarrays para el descarte de cáncer mamario el cual es el objeto de estudio del investigador. Los resultados obtenidos demostraron un mayor desempeño con el modelo propuesto.

Alvarez (2014) en su investigación, tuvo como finalidad la elaboración de un sistema que detecte el cáncer de mama por medio del uso de redes neuronales, el cual pueda llevar el procesamiento de las imágenes para luego dar un diagnóstico el cual será validado mediante un conjunto de datos. La metodología que utilizó para esta investigación fue de tipo aplicativa, el diseño de estudio planteado por el autor paso por las fases de adquisición de las imágenes, preprocesamiento, entrenamiento de la red, codificación y diagnóstico. Los resultados que obtuvo el autor en su investigación es la

posibilidad de encontrar características en las imágenes el cual se clasificaron entre tejido sano y canceroso por medio de la utilización de una red perceptrón multicapa.

Eche (2017) efectuó un estudio de investigación cuyo objetivo general es el desarrollo de un algoritmo que detecte la presencia de líneas B en las imágenes de ultrasonido para prevenir la neumonía en menores de 5 años y estimar la sensibilidad, especificidad y exactitud de dicha detección. Para el análisis del umbral optimo fue seleccionado a través del uso de una curva roc; para la clasificación el autor utilizó las máquinas de soporte vectorial utilizando el kernel RBF. Los resultados que obtuvo cuando analizó una sola característica fueron del 77% de sensibilidad, 75% de especificidad y 75% de exactitud; al realizar la observación con dos características se logró apreciar una mejora con un 93% en la sensibilidad, 86% en la especificidad y 88% de exactitud; finalmente al analizar con más de cinco características tuvo una sensibilidad del 100%, especificidad con 98% y la exactitud del 98%.

Bardales (2013) realizó un estudio cuyo objetivo estuvo enfocado en el diseño de un algoritmo de reconocimiento de tumores de mama para mejorar la identificación de tumores cancerígenos. El tipo de investigación es de tipo aplicativa, el diseño que manejo fue la aplicación de una pre-prueba y post-prueba utilizando un solo grupo de control. El resultado con el cual concluye el autor es el desarrollo de un algoritmo de reconocimiento es beneficioso para la mejora de la identificación y diagnóstico del cáncer de mama. El autor recomienda para trabajos futuros la consideración el algoritmo SVM para el reconocimiento de tumores mamarios.

Paulino (2019) en su investigación se propuso como objetivo general el desarrollo de un sistema experto probabilístico basado en Redes Bayesianas para la predicción de riesgo de cáncer cervical, así mismo realiza un estudio sobre los diferentes modelos

existentes como máquinas de soporte vectorial, redes neuronales, arboles de decisiones entre otros métodos, para luego poder desarrollar un sistema confiable el cual contenga una precisión no menor al 80%. Para el desarrollo del proyecto realizó una base de datos con 322 registros y 15 atributos de la información de pacientes; la metodología que propuso consta de tres etapas, el primero es la selección del conjunto de datos, continuó con el tratamiento de datos y finalmente la construcción del modelo probabilístico. El resultado que obtuvo al desarrollar el proyecto resulto con una tasa de éxito del 96%.

2.1.2. A nivel internacional

Pedraza (2015) realizando su investigación tuvo por objetivo diseñar, implementar y validar un algoritmo para la identificación de microcalcificaciones malignas en mamografías empleando técnicas de machine learning. La metodología que implementó para la clasificación es un modelo de aprendizaje supervisado, el autor construyó una base de datos de entrenamiento de alrededor de 5000 observaciones donde 1000 son de tipo maligno y 4000 de tipo normal, luego se planteó el entrenamiento de los clasificadores empleando máquinas de soporte vectorial, regresión logística, redes neuronales, MetaCost y bagging; para realizar la validación del modelo de entrenamiento el autor utilizó validación cruzada 10-fold para ver que algoritmo de clasificación tiene mejor desempeño; además el autor realizó una aplicación para la interacción con el usuario donde se realizó una aplicación para la carga de imágenes mamográficas y la aplicación devolverá un análisis estadísticos de los datos ingresados. Los resultados que se obtuvieron luego de realizar el entrenamiento con los algoritmos planteados obtuvieron un error de clasificación mínimo de 0,92% $\pm$ 6,07% y un error máximo de 2,86% $\pm$ 6,7%.

Vijayarajeswari et al. (2019) en su proyecto de investigación el autor realizó la clasificación imágenes mamográficas utilizando características que fueron extraídas por

medio de la utilización de la transformada de Hough, a través de esta técnica resaltó el aislamiento de las características de una imagen. En el artículo de investigación el autor hace referencia sobre la extracción de características y sus estrategias para su clasificación, el autor seleccionó el algoritmo de máquinas de soporte vectorial para realizar la clasificación. La metodología utilizada constó de las etapas de recolección de las imágenes, extracción de las características y clasificación. Los resultados que obtuvo el autor luego de aplicar la transformada de Hough y el clasificador SVM logrando obtener una exactitud del 94% destacándose de entre otros clasificadores.

Kaur et al. (2019) en el estudio de investigación que realizaron, enfatizan la aplicación del aprendizaje profundo para mejorar los diagnósticos de los especialistas en salud. Los investigadores utilizaron el dataset MiniMIAS, para el procesamiento y extracción de las características y el algoritmo de k-means. Para el proceso de clasificación el autor hace la separación de los datos de entrada de un 70% para el entrenamiento y un 30% para el test, así mismo, el autor hace uso de los clasificadores de redes neuronales profundas y máquinas de soporte vectorial multiclase. Los resultados obtenidos en la investigación luego de realizar la clasificación entre las clases benigno, maligno y normal a través de los algoritmos de clasificación seleccionados, para la clase normal obtuvieron un 95%, la clase benigno resultó con un 94% y para la clase maligno obtuvieron un 98%; los resultados obtenidos por los autores demuestran que el algoritmo de SVM tiene mejores resultados comparados con otros clasificadores.

Navya et al. (2020) los autores propusieron la utilización de máquinas de soporte vectorial y una red neuronal como métodos de clasificación para el pronóstico de tumores de mama. Para la obtención de buenos resultados los autores aplican una validación cruzada de 10 y 5 veces. Como fuente de datos para el desarrollo de su investigación

utilizaron una base de datos que obtuvieron del repositorio de aprendizaje automático UCI. Los resultados luego de realizar la experimentación por medio de la herramienta weka, así mismo, los investigadores obtuvieron una validación cruzada de 5 veces y teniendo un mejor resultado con la utilización de máquinas de soporte vectorial frente a las redes neuronales.

Pharswan et al. (2020) en la investigación que realizaron, plantearon la realización de un sistema basado en machine learning para el descarte de cáncer de seno utilizando mamografías, seleccionaron la selección de la región de interés y para la extracción de las características utilizó una matriz de coocurrencia a nivel de grises, para la clasificación utilizó SVM y KNN. El proyecto de investigación utilizó la herramienta de MATLAB y el dataset utilizado fue MiniMias. Para evaluar el modelo para cada algoritmo de clasificación y comparar el rendimiento de cada uno de ellos, utilizaron una matriz de confusión utilizando las variables de medición como la exactitud, especificidad, precisión y puntaje f1. Los resultados obtenidos demostraron que los algoritmos de máquinas de soporte vectorial obtuvieron una exactitud del 94% superando al algoritmo k-nn.

2.2. Marco teórico

2.2.1. Cáncer de mama

Según Sumbaly (2021) define al cáncer como un problema a nivel mundial que viene afectando la vida de muchas personas (Sumbaly et al., 2021), además por su alta incidencia de mortalidad y el alto grado de costo social y económico que este conlleva (MINSA, 2020).

La Organización Mundial de la Salud (OMS) utiliza la definición del cáncer de seno como un amplio grupo de enfermedades, logrando afectar a cualquier parte del organismo (OMS, 2021), el inicio de esta enfermedad se da con la multiplicación de

células anormales (INC, 2015) formando una masa anormal de tejido llamado tumor (Rodríguez, 2009).

La aparición de esta enfermedad se puede dar en distintas partes del área del seno, la mayoría de los casos se pueden iniciar en los conductos que llevan la leche materna al pezón o iniciarse en las glándulas que producen leche materna (Escudero, 2018). Pedraza (2015) hace referencia que esta enfermedad no presentar algún tipo de síntoma en sus etapas iniciales, en las etapas que son avanzadas el síntoma a presentarse es la aparición de masas palpables, inflamación o cambio de color en el pezón.

Una enfermedad compleja como el cáncer de mama, tiene la posibilidad de ser producida por cierta combinación de factores de riesgo, estos pueden ser prevenibles (obesidad, alcoholismo) y no prevenibles(genero, edad, genética) (Sumbaly et al., 2021).

2.2.2. Estadísticas sobre el cáncer de mama

A nivel mundial el número de diagnósticos realizados es aproximadamente 19 millones de nuevos casos cada año, evidenciando que el nivel del control del cáncer no demuestra alguna mejora. Además, se estima que para el 2040 el número de casos de cáncer aumentará alrededor de 30 millones. Así mismo, la evidencia demuestra que la tercera parte de las muertes a nivel mundial por cáncer se produce en países de bajos y medianos ingresos (MINSA, 2020, 2021)

La OMS a través de su plataforma GLOBOCAN, indica que en 2020 el número de casos de cáncer reportados a nivel mundial fue alrededor de 20 millones, el cáncer de mama, es ubicada como la principal enfermedad neoplásica con un total de 2,261,419 casos representando el 11,7% seguido por los casos de cáncer al pulmón, colon, próstata

y estómago. Así mismo, los casos por fallecimiento a causa del cáncer de mama son alrededor del 7% (GLOBOCAN, 2021).

La Organización Panamericana de la Salud (OPS), indica que en América Latina el 45% de las muertes por cáncer se encuentran en esta región y en 2030 se estima que tendrá un incremento de aproximadamente de 2,1 millones de casos (OPS, s.f.). A pesar que el nivel de incidencia en América Latina es menor con respecto a otras regiones como Europa y Estados Unidos, el número de casos de mortalidad es mucho mayor (MINSA, 2021). Los casos de cáncer de mama en mujeres menores a 45 años en la región de Latinoamérica son alrededor del 20%, mientras que en América del Norte y la Unión Europea es del 12,3% (Romieu et al., 2019).

En el Perú la situación anual de incidencias de cáncer de mama según la Agencia Internacional de Investigación en Cáncer indica que en 2014 fue de 40 casos por cada 100 mil habitantes, mientras que la tasa de mortalidad fue de 9,2 casos por cada 100 mil habitantes (MINSA, 2018). En comparación a años anteriores en 2020 se ha visto el incremento de los casos de cáncer en las mujeres que están en las edades de 40 a 69 años, se estima que 90 de cada 100 mil padecieron esta enfermedad y la tasa de mortalidad fue de 22 de cada 100 mil casos.

En 2017 se registró que el 73% de los especialistas en oncología clínica están en la capital y en 2018 el registro de los especialistas en radioterapia que representan el 82,6% del total a nivel nacional están ubicados en Lima y Callao (MINSA, 2020). Los equipos registrados en 2017 a nivel nacional, se cuenta un total de 110 equipos a nivel nacional, sin embargo en los departamentos como Cajamarca, Cusco, Piura, Lima, Loreto, Ica, La Libertad y Puno la cantidad de equipos destinados son insuficientes para su población objetivo (Ramos & De La Cruz-Vargas, 2020).

2.2.3. Mamografía

En las primeras etapas el cáncer viene a ser más tratables y requiere de un menor costo para su tratamiento (Akram, 2019), el cáncer de mama puede ser clasificado según el tipo que este desarrolle, esta clasificación puede ser de tipo benigno o maligno (Sumbaly et al., 2021).

La prueba de diagnóstico que presenta un menor riesgo para el paciente y una elevada seguridad para el descarte de cáncer de mama es la mamografía, para este tipo de diagnóstico es necesario un equipo en buen estado y un especialista debidamente entrenado para minimizar los riesgos de una biopsia innecesaria (Diaz et al., 2012).

Cada seno se comprime entre placas de rayos X para aplanar los tejidos del seno (Akram, 2019), así mismo, Rodríguez (2009) indica que “se toma a un ángulo de 45° permitiendo analizar el perfil del seno (proyección vertical) y observar un mayor volumen del tejido mamario. Por otra parte, la proyección cráneo-caudal se toma desde la parte superior del seno” (p. 8).

La sensibilidad que se obtiene a través del tamizaje por mamografía tiene un aproximado de 63% en mamas muy densas y en los casos en que el paciente tiene las mamas muy grasas tienen un 87%, los estudios demuestran que las mamografías digitales tienen mejor contraste entre el tejido mamario y los tumores mamarios (MINSA, 2017).

En las mamografías las microcalcificaciones aparecen como pequeños puntos menores a 5mm y lesiones típicamente mayores a 2.5mm, de los cuales pueden tener forma circular u ovalada con bordes definidos y nítidos. Las lesiones menores a 2.5mm son catalogados como cáncer de seno no invasivo (Pedraza, 2015).

2.2.4. BI-RADS

Es utilizada para la descripción de una categoría al cual pertenece un tumor, el BI-RADS es considerada como un estándar en los informes de mamografías, ultrasonidos, resonancias. En la mamografía el BI-RADS se describe los hallazgos encontrados como las masas (describe la forma, margen, densidad y tamaño), calcificaciones (estas acumulaciones pueden dividirse en malignos o benignos), distorsión arquitectónica (una región anormal en la mama de forma aleatoria), lesión de la piel y asimetría mamaria (el área del tejido mamario del seno es diferente al otro seno) (Santos, 2016).

2.2.5. Machine Learning

Machine Learning (ML) o Aprendizaje Automático, es definido como un conjunto de métodos que permiten a las computadoras automatizar la creación de modelos que son basados en datos a través del descubrimiento de patrones (Bhavsar et al., 2017). Arthur Samuel (1959) define el termino de ML como un "campo de estudio que brinda a las computadoras la capacidad de aprender sin estar programadas explícitamente", centrándose en el desarrollo de programas informáticos que puedan realizar cambios cuando estos muestran nuevos datos (Kalali et al., 2019). Así mismo, Tom M. Mitchell (1997) realizó la definición de ML donde "un programa de computadora aprende de una experiencia E con respecto a alguna clase de tarea T y una medida de performance P, si su performance en la tarea T, medida por P, mejora con la experiencia E".

2.2.6. Modelos de Machine Learning

Los modelos principales que se desarrollan en las máquinas de aprendizaje son los tipos de aprendizajes supervisados, no supervisados y semi-supervisados; a continuación, se realiza una descripción de los modelos de aprendizaje.

2.2.6.1. Aprendizaje Supervisado

El método de aprendizaje supervisado, consta del entrenamiento de una función para que este calcule las variables de salidas en función de los datos de entrada (Bhavsar et al., 2017), de tal manera se define como su objetivo el de incitar un modelo el cual pueda aprender a través de un conjunto de datos donde cada uno de los datos poseen atributos y etiquetas (Bueno, 2018).

A partir de una base de datos utilizada como ejemplos de entrenamiento con una etiqueta de destino específica, de tal forma que $\{(x_1, y_1), \dots, (x_i, y_i)\}$ donde x_i corresponde al vector de salida del i-ésimo ejemplo y y_i es su etiqueta, así mismo, su objetivo es la creación de un modelo $g : X \rightarrow Y$, donde se pueda predecir la etiqueta Y a partir de futuros datos X (Bejnordi, 2017).

Para el modelo de aprendizaje supervisado se cuenta con dos tipos de problemas cuando son de tipo regresión y clasificación, a continuación, se describe estos problemas como:

- **Regresión:** Para los problemas de tipo regresión, el objetivo es desarrollar una relación entre los datos de salida y los datos de entrada, el cual pueda utilizar una función continua para ayudar a las máquinas de aprendizaje a entender cómo es el cambio en las salidas para las entradas dadas (Bhavsar et al., 2017).

- **Clasificación:** La clasificación consiste en establecer las clases a las que pertenece un nuevo dato de entrada, sobre la base de un conjunto de datos de entrenamiento cuya cualidad es conocida (Arnedo, 2016).

2.2.6.2. Aprendizaje no Supervisado

Los modelos de aprendizaje no supervisado usa un conjunto de datos de entrada $\{x_1, x_2, \dots, x_n\}$ el cual no contiene una salida determinada, esto es demostrado en la función g; tal que $g: X \to Y$, para este algoritmo la función g tiene que realizar el mapeo sin una salida Y, este tipo de algoritmos encuentran patrones a partir de los datos que se pueden tomar como ruido no estructurado (Geras, 2011). De tal manera podemos definir que el principal objetivo de este tipo de aprendizaje es descubrir patrones entre las muestras y revelar las clases ocultas detrás de las características (Ortiz, 2019).

2.2.6.3. Aprendizaje semi-supervisado

Este tipo de modelos generalmente cuenta con una cantidad de datos de entrenamiento que pueden estar etiquetados y otra gran cantidad de datos que no poseen etiquetas; esto indica que los datos sin etiquetas al utilizarse conjuntamente con una pequeña cantidad de datos etiquetados, logran mejorar sustancialmente el aprendizaje (Jayaram et al., 2015).

Para los problemas de aprendizaje semi-supervisado se pueden describir de la siguiente manera:

Para el conjunto de datos de entrenamiento X que tengan l datos etiquetados:

$$X_l = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$$

Y de los datos sobrantes, son aquellos que no poseen etiqueta alguna.

$$X_u = \{x_{l+1}, \dots, x_n\}$$

Los datos sin etiquetar llegan a ser más grande que los datos etiquetados.

$$n - l \gg l$$

2.2.7. Algoritmos de aprendizaje supervisado

2.2.7.1. K-Nearest Neighbors

El algoritmo k-nearest neighbors o vecinos más próximos tiene por objetivo encontrar un número de muestras de entrenamiento en una distancia cercana a un nuevo punto y a partir de esos puntos poder predecir un valor (Islam et al., 2020).

2.2.7.2. Árbol de decisiones

Este algoritmo se enfoca en la representación de los datos más comprensibles, realizando la división de los datos por sus atributos logrando una separación de los datos en sus distintas clases existentes hasta que alcance cierto criterio de parada (Luckert & Schaffer-Kehnert, 2015). Según la definición de Bell (2015) para el objetivo de estos algoritmos es “crear un modelo viable que predecirá el valor de una variable objetivo en función del conjunto de variables de entrada” (p. 45).

2.2.7.3. Redes neuronales

Este algoritmo está basado en la composición de la estructura biológica de las neuronas, las cuales están conectadas entre sí formando una gran red compleja (Shalev-Shwartz & Ben-David, 2013), este algoritmo está compuesto por elementos adaptativos con bases jerárquicas, las cuales procesan la información en respuesta a entradas externas (Matich, 2001).

2.2.7.4. Redes bayesianas

Este tipo de algoritmo es utilizado en entornos de aprendizaje supervisado y no supervisado. Según la definición de Enrirque Sucar (2011) las redes bayesianas "son una representación gráfica de dependencias para razonamiento probabilístico, en el cual los nodos representa variables aleatorias y los arcos representan relaciones de dependencia directa entre las variables" (p. 2).

2.2.8. Máquinas de Soporte Vectorial

El algoritmo de máquinas de soporte vectorial (por sus siglas en ingles SVM), este método fue propuesto por Vapnik y Cortes en los años 90 (Cortes & Vapnik, 1995), también es considerada como una gran herramienta de clasificación para modelos de aprendizaje supervisado (Afifi et al., 2019), utilizado para clasificación binaria y multi clasificación; considerado como un clasificador lineal por el uso de hiperplanos para la separación de un conjunto de datos de entrada estos pueden ser separables o cuasi separables (Carmona, 2014). Las SVM tienen por objetivo mapear los datos del vector de entrada en un espacio de características, para luego ser divididas por medio de la utilización de un hiperplano, así mismo, tener un margen de separación entre una clase y otra (Shihong et al., 2003).

2.2.8.1. Casos linealmente separables

Para los casos que son linealmente separables donde n es el número de instancias de entrenamientos de un conjunto de datos $X = \{(x_1, y_1), \dots, (x_i, y_i)\}$, para cada muestra (x_i, y_i) para todo $i = 1, \dots, n$, donde $x_i \in \mathbb{R}^d$ y una etiqueta $y_i \in \{+1, -1\}$ para todo $y_i \in \mathbb{R}$ (Carmona, 2014; González et al., 2017).

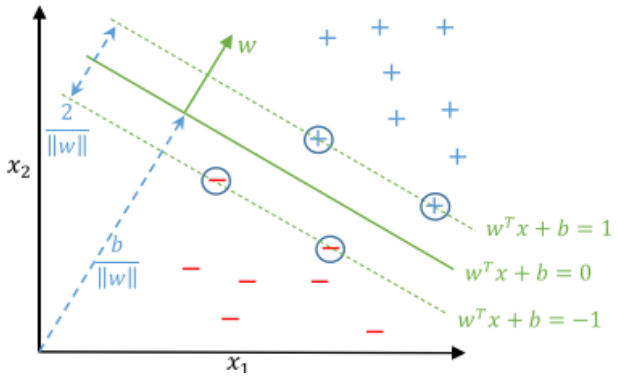
La siguiente ecuación describe una función lineal que pueda ser capaz de separar un conjunto de datos:

$$D(x) = (w_1x_1 + w_1x_1 + \cdots + w_dx_d) + b = < w, x > + b$$

Para los datos que pueden ser linealmente separables, se puede dividir por medio de una recta que divida el grafico x_1 vs x_2, siempre cuando la dimensión sea $d = 2$.

El hiperplano puede ser descrito como $w.x + b = 0$ donde $\boldsymbol{w}$ es la normal al hiperplano y $\frac{b}{||w||}$ es la distancia perpendicular desde el hiperplano al origen. Pueden existir diversos hiperplanos que puedan satisfacer las condiciones para la separación de las clases, Los vectores de soporte son los puntos de los dataset ubicados en $w^T x^i + b = 1$ y $w^T x^i + b = -1$, el hiperplano $\mathcal{H}$ definido por $w^T x + b = 0$ el cual logra separar el conjunto de datos en dos partes, en la Figura 1 se aprecia un conjunto de datos el cual está compuesto por "+" que representa a clases positivas y "-" a las clases negativas. Para encontrar el hiperplano $\mathcal{H}$ que realicé la mejor separación de las clases, pude estar compuesto de múltiples hiperplanos $\mathcal{H} = \{h_1, h_2, \dots, h_n\}$, el objetivo de este algoritmo es encontrar el mejor hiperplano de separación a través del margen definido por $\frac{2}{\|w\|}$.

Figura N° 1: SVM - casos linealmente separables



Fuente: (Flach, 2012; Fletcher 2009, como se citó en Aguilar & Vásquez, 2016)

Se dice que un conjunto de datos es linealmente separable si existe un vector w y un escalar b, así las desigualdades serán:

$$w \cdot x_i + b \geq +1 \text{ para } y_i = +1$$

$$w \cdot x_i + b \leq -1 \text{ para } y_i = -1$$

Las anteriores ecuaciones se pueden combinar como:

$$y_i(w \cdot x_i + b) \geq 1, i = 1, \dots, l$$

2.2.8.2. Casos linealmente no separables

Cuando los datos no pueden ser separados linealmente, es necesario realizar el procesamiento de los datos de su dimensión de origen a una dimensión de mayores características. Las SVM deben lograr maximizar el margen y minimizar el número de datos no separables (Havrylchyk & Poncet, 2007).

Agregando una variable de holgura $\xi \geq 0$ para cada vector de entrenamiento, se hace la modificación de las ecuaciones de los casos linealmente separables:

$$(w \cdot x_i) + b \geq 1 - \xi_i \text{ si } y_i = 1$$

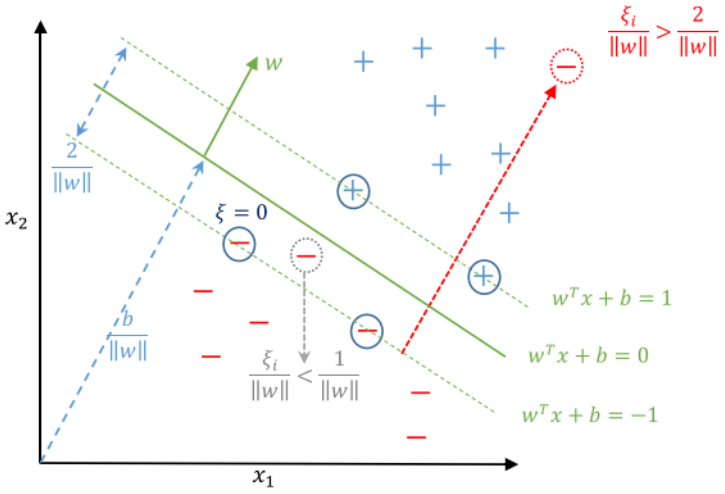
$$(w \cdot x_i) + b \leq -1 + \xi_i \text{ si } y_i = -1$$

$$\xi_i \geq 0, \forall\, i = 1, \dots, n$$

Las anteriores ecuaciones se pueden combinar como:

$$y_i[(w \cdot x_i) + b] \geq 1 - \xi_i, \forall\, i = 1, \dots, n$$

Figura N° 2: SVM - casos linealmente no separables



Fuente: (Flach, 2012; Fletcher 2009, como se citó en Aguilar & Vásquez, 2016)

Para los puntos de datos que se encuentran en el lado incorrecto del límite del margen tienen una penalización, el problema en forma primaria ahora es una minimización de la función:

$$\min\{\frac{1}{2}w^2 + C\sum_{i=1}^{L}\xi_i\}$$

$$y_i[(x_i \cdot w) + b] - 1 + \xi_i \geq 0\,, \forall\, i = 1, \ldots, n$$

$$\xi_i \geq 0\,, \forall\, i = 1, \ldots, n$$

Donde C controla la compensación de la penalización de la variable de holgura y el tamaño del margen. Aplicando el teorema de Lagrange, se necesita minimizar con respecto a w, b y ξ_i con respecto a a_i donde $a_i \geq 0,\ r_i \geq 0,\ \forall\, i = 1, \ldots, n.$

$$L(w, b, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}a_i\left[y_i\left((w \cdot x_i) + b\right) - 1 - \xi_i\right] - \sum_{i=1}^{n}r_i\xi_i$$

2.2.8.3. Clasificación multi-clase

La implementación de SVM para multiclases, se iniciará introduciendo variables de holgura ξ a las restricciones que se utilizan para las clasificaciones binarias obteniendo las siguientes ecuaciones:

$$w \cdot x_i + b \geq +1 - \xi_i \, , si \; y_i = +1$$

$$w \cdot x_i + b \leq -1 + \xi_i, si \; y_i = -1$$

$$\xi_i \geq 0, \forall \; i = 1, \dots, n$$

Combinando las anteriores ecuaciones:

$$y_i(w \cdot \Phi(x_i) + b) \geq 1 - \xi_i, \forall \; i = 1, \dots, n$$

$$\xi_i \geq 0, \forall \; i = 1, \dots, n$$

La función kernel es determinada por Φ, la variable de holgura ξ trabaja con un margen suave donde $0 < \xi \leq 1$.

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \xi_i$$

$$sujeto \; a \; y_i(w \cdot \Phi(x_i) + b) \geq 1 - \xi_i, \forall \; i = 1, \dots, m$$

$$\xi_i \geq 0, \forall \; i = 1, \dots, m$$

2.2.8.4. Kernels

Cuando se aplica los algoritmos de SVM en datos que son linealmente separables, se inicia con la creación de una matriz H:

$$H_{ij} = y_i y_j k(x_i, x_j) = x_i \cdot x_j = x_i^T x_j$$

Para $k(x_i, x_j)$ de la anterior ecuación, es un ejemplo de un tipo de kernel utilizado, para $x_i^T x_j$ que representa la utilización de un kernel lineal. Las funciones kernel son muy útiles cuando hay problemas de clasificación que no pueden ser linealmente separables en su espacio de entrada original, el cual tendrá que redimensionara el espacio vectorial a uno mayor (Fletcher, 2009). Existen diversos tipos de kernels que pueden ser utilizados por las SMV y pueden ser descritos como:

- **Kernel Lineal:** Para los problemas de clasificación donde los datos son linealmente separables en el espacio definido por:

$$k(x_i, x_j) = x_i \cdot x_j$$

- **Kernel Polinomial:** Uno de los kernels más utilizados para las relaciones no lineales:

$$k(x_i, x_j) = (1 + x_i \cdot x_j)^d$$

- **Kernel Gaussiano:**

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

- **Kernel Neural:**

$$k(x_i, x_j) = \tanh\left(a x_i \cdot x_j + b\right)$$

- **Anova Kernel**

$$k(x_i, x_j) = \tanh\left(a x_i \cdot x_j + b\right)$$

- **Fourier Series Kernel**

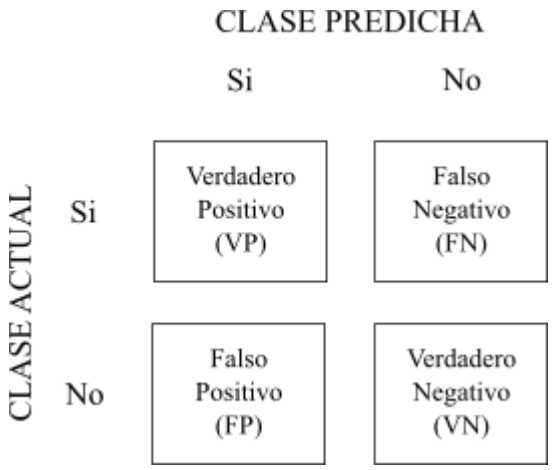
$$k(x_i, x_j) = \frac{sen(N + \frac{1}{2})(x_i - x_j)}{sen(\frac{1}{2}(x_i - x_j))}$$

2.2.9. Evaluación del desempeño del modelo de clasificación

2.2.9.1. Matriz de confusión

Para la validación de los modelos de entrenamiento se crea una matriz de confusión para las clasificaciones binarias y modelos de clasificación múltiple, esta matriz compara las etiquetas de la clase predicha de un punto de datos con su etiqueta de clase real (Kirk, 2015), cada fila representa la clase real, mientras que las columnas representan las clases predichas por el algoritmo como en la Figura 3.

Figura N° 3: Esquema de una matriz de confusión



Elaboración propia.

Según Bhattacharjee (2020, p. 273) los resultados de las predicciones pueden resultar en alguna de estas cuatro probabilidades:

- **Verdaderos positivos (VP):** Representa el número total de elementos de una clase positiva donde su etiqueta es de tipo verdades es igual que la etiqueta predicha.
- **Verdaderos Negativos (VN):** Representa el número total de elementos que fueron clasificados correctamente por el modelo como clase de tipo negativo.

- **Falsos Positivos (FP):** Representa el número total de elementos de una clase falsa, donde el modelo lo clasificó erroneamente como positivo.
- **Falsos Negativos (FN):** Representa el número total de clases positivas que fuerón incorrectamente clasificados como una clase negativa.

2.2.9.2. Métricas de rendimiento

La información que se obtiene a partir de la matriz de confusión, son indicadores para medir el desempeño del clasificador (Awad & Khanna, n.d.; Raschka, 2014), estas pueden ser descritas como:

Exactitud (Accuracy): Hace referencia al número de predicciones que fueron correctamente clasificados por el algoritmo:

$$Accuracy = \frac{VP + VN}{VP + FP + VN + FN}$$

Precisión: Es la referencia al total de casos positivos detectados o predichos:

$$Precision = \frac{VP}{VP + FP}$$

Sensibilidad (Recall): La cantidad de casos positivos que fueron correctamente predichos por el algoritmo clasificador.

$$Recall = \frac{VP}{VP + VN}$$

Especificidad: El número total de casos negativos que el algoritmo clasificó correctamente:

$$Especificidad = \frac{VN}{VN + FP}$$

F1 Score: Busca el equilibrio entre la precisión y la sensibilidad, y hay una distribución de clase desigual

$$F1 = 2 * \frac{Precisión * Recall}{Precisión + Recall}$$

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1. Enfoque y diseño de investigación

3.1.1. Enfoque de investigación

El tipo de investigación para este proyecto según sus características es marcado dentro del enfoque cuantitativo. Según Ñaupas (2018) indica que el enfoque cuantitativo tiene la característica de "utilizar métodos y técnicas cuantitativas y por ende tiene que ver con la medición, el uso de magnitudes, la observación y medición de las unidades de análisis" (p. 140).

3.1.2. Diseño de investigación

El diseño de investigación seleccionado para este proyecto de tesis es de tipo cuasi experimental y exploratorio.

La descripción de un diseño cuasi experimental según hace referencia Hernández Sampieri (2014). indica que estos sistemas "manipulan deliberadamente al menos una variable independiente para observar su efecto y relación con una o más variables dependientes, solo que difieren de los experimentos en el grado de confiabilidad que se pueda tener sobre la equivalencia inicial en los grupos, puesto que son grupos intactos" (p. 151).

El estudio de tipo exploratorio tiene como objetivo examinar un tema relativamente poco estudiado para buscar desarrollar estudios más completos sobre un contexto en particular, así también, establecer prioridades para las investigaciones futuras (Hernández, 2014).

3.2. Población

Para el desarrollo del proyecto se empleó los dataset Sociedad de Análisis de Imágenes Mamográficas (MIAS), Base de Datos Digital Mamografía de Tamizaje (DDSM), Repositorio Digital de Cáncer de mama (BCDR) y un conjunto de imágenes mamográficas proporcionas por el Hospital Regional "Manuel Núñez Butron".

Tabla N° 1: Descripción de los dataset utilizados

Nombre	Descripción
MiniMIAS	Base de datos creada por la Sociedad de Análisis de Imágenes mamográficas (MIAS). La base de datos MiniMIAS es la versión reducida a un tamaño 1024x1024 pixeles, disponible en https://www.kaggle.com/kmader/mias-mammography
DDSM	DDSM proviene de las siglas en ingles Digital Database for Screening mammography, desarrollada por el Hospital General de Massachusetts y por la Universidad del Sur de Florida por el departamento Ciencias de la Computación, disponible en https://www.kaggle.com/cheddad/miniddsm
BCDR	Repositorio Digital de Cáncer de Mama, desarrollada con el propósito de estudio del cáncer de mama y el desarrollo de nuevos métodos para el diagnóstico y detección por medio de las computadoras, disponible en https://bcdr.ceta-ciemat.es/patient/list#
HRMNB	Imágenes proporcionadas por el Hospital Regional "Manuel Núñez Butron" de la ciudad de Puno, del departamento de diagnóstico por imágenes.

Elaboración propia

3.3. Muestra

A continuación, se describe los dataset usados para el proyecto de investigación y el número de imágenes seleccionada por cada una de ellas:

Tabla N° 2: Descripción datasets

Nombre	Tipo de datos	Tipo de atributos	Número de datos
MiniMIAS	Multivariado	Categórico, Numérico	186
DDSM	Multivariado	Categórico, Numérico	120
BCDR	Multivariado	Numérico	120
HRMNB	Multivariado	Categórico	100

Elaboración propia

3.3.1. BCDR

Tabla N° 3: Composición del dataset BCDR

I.V	Node	Calci	Micro	Archi	Stroma	Den
1:177	0:452	0:432	0:606	0:779	0:726	1:165
2:221	1:356	1:376	1:202	1:029	1:082	2:222
3:187						3:342
4:223						4:79

Fuente: (Santos, 2016)

3.3.2. MiasMammography

Tabla N° 4: Composición y descripción del dataset MiniMIAS

Columna	Descripción	Tipo	Atributo	Descripción
1ro	Número de referencia de la base de datos MiniMIAS	Categórico	[Mdb001 – mdb322]	ID
2do	Carácter del tejido	Categórico	F	Graso

			G	Graso Glandular
			D	Denso Glandular
3ro	Clase de anomalía presente	Categórico	CALC	Calcificación
			CIRC	Masas definidas
			MISC	Otras masas mal definidas
			ARCH	Distorsión arquitectónica
			ASYM	Asimetría
			NORM	Normal
			SPIC	Masas espiculadas
4to	Gravedad de la anormalidad	Categórico	B	Benigno
			M	Maligno
5to, 6to	Coordenadas de imagen x, y del centro de anomalía	Numérico		
7mo	Radio aproximado (en píxeles) de un círculo que encierra la anormalidad	Numérico		

Elaboración propia

3.3.3. DDSM

Tabla N° 5: Composición y descripción del dataset DDSM

Tipo	**Descripción**
Sutileza	El valor de sutileza indica cuán difícil es encontrar una lesión, cuanto más grande es más fácil (1 es "Sutil y 5 es "obvio").
Forma de masa	**Redondo:** La anormalidad es de forma circular. **Ovalada:** La anormalidad es de tipo elíptica. **Lobulado:** La anormalidad presenta contornos con ondulaciones. **Irregular:** La anormalidad no presenta forma. **Distorsión:** Equivalente a una distorsión arquitectónica en BI-RADS.
Margen de masa	**Circunscrito:** Los márgenes están marcadamente delimitados con una transición abrupta entre la lesión y el tejido circundante. **Mala Definición:** Definición irregular de los márgenes.

	Oculto: La anomalía está oculto por tejido normal. **Masa espiculada:** Presenta líneas que irradian desde los márgenes. **Microbulado:** Los márgenes de las anormalidades contienen pequeñas ondulaciones.
Tipo de Clasificación	**Punteados:** Son circulares, de menos de 0,5mm. **Calcificaciones:** Pequeñas y que no se pueden caracterizar. **Pleomórfico:** Más grande que amorfo. **Centro luminoso:** Calcificaciones menores de 1mm hasta más de un centímetro. **Ramificación lineal fina:** Calcificaciones delgadas e irregulares.
Distribución de la calcificación	**Agrupado:** Utilizada cuando se encuentra múltiples calcificaciones ocupando un pequeño volumen de tejido. **Lineal:** Calcificaciones dispuestas en línea. **Regional:** Calcificaciones esparcidas en un gran volumen de tejido mamario. **Difuso:** Calcificaciones distribuidas aleatoriamente por toda la mama.

Elaboración propia

3.3.4. HRMNB

Las imágenes proporcionadas por el Hospital Regional “Manuel Núñez Butron” por el departamento de diagnóstico por imágenes, está compuesto por 10 casos de estudio, cada caso cuenta con mamografías del seno izquierdo y derecho, con un total de 20 imágenes, donde 14 imágenes son de tipo normal, 5 de tipo benigno y 1 de tipo maligno. Para resguardar la confidencialidad de los pacientes se tuvo que resguardar los datos personales, así como los datos del médico a cargo de la evaluación, se realiza una descripción de la composición del conjunto de datos obtenido.

Tabla N° 6: Composición y descripción de los datos de HRMNB

Tipo	Descripción	Evaluación
ID	Número de identificación para la mamografía.	[M1 – M20]
Lado	Describe si la mamografía pertenece al lado izquierdo o derecho del paciente.	Derecho: R Izquierdo: L
Evaluación	El resultado final realizado por el especialista del área, esto involucra la clasificación BI-RADS	**BI-RADS 0**: Estudios incompletos **BI-RADS 1**: Descubrimiento Negativo **BI-RADS 2**: Descubrimiento Benigno **BI-RADS3**: Descubrimiento probablemente benigno. **BI-RADS 4**: Descubrimiento sospechoso **BI-RADS 5**: El estudio confirma la presencia de cáncer.

Elaboración propia

A través de la librería Pillow de Python se realizó la manipulación de las imágenes cambiando la rotación, escala y la orientación, generando 5 imágenes por cada mamografía, obteniendo un total de 100 imágenes para su evaluación.

3.4. Material experimental

Los recursos utilizados para el desarrollo del proyecto de investigación son descritos tanto como software y hardware.

- Componentes de PC: Intel Core i7 10th, 32GB RAM y una tarjeta de video 1050ti.
- Lenguaje de programación seleccionado Python.
- Editor de texto Jupyter Notebook y Visual Studio Code.

- Para el control de versiones se hace uso de la herramienta GIT.

3.5. Definición de procesos

Basados en los antecedentes de proyectos analizados se propone desarrollar 4 módulos principales compuestos por el preprocesamiento, extracción de características, clasificación y evaluación del modelo.

a) Preprocesamiento

El dataset utilizado MiniMIAS está compuesto por imágenes de 1024x1024 pixeles con una matriz de 1,048,576 pixeles cada imagen, así también el dataset cuenta con la dimensión y el radio donde se allá la anormalidad. En este módulo de desarrollo se tiende a realizar el mejoramiento de las imágenes para mejorar la variación entre los objetos y el ruido de fondo innecesario que pueda existir la imagen.

Las imágenes de la base de datos utilizada para el entrenamiento están compuestas por datos que pueden ser irrelevantes para la investigación, así como también la existencia de ruido de fondo. Se aplica un filtro para para la eliminación del ruido existente, un filtro para el suavizado de la imagen no lineal. Se realiza un recorte de cada imagen contenidas en el dataset reduciendo el tamaño a 150x150 obteniendo una matriz de 22,500 pixeles por cada imagen.

b) Extracción de características

En este módulo se seleccionan algoritmos para la extracción de las características en las mamografías para realizar la clasificación de las anormalidades. En este caso para el reconocimiento de la región de interés se realiza de forma manual ya que se necesita el

apoyo de un especialista para determinar las regiones con anomalías, el dataset utilizada para el entrenamiento cuenta con las regiones de interés ya definidas.

c) Clasificación

Para el proyecto de investigación se implementará un clasificador de máquinas de soporte vectorial para realizar la clasificación entre las clases benigno, maligno y normal; se establece parámetros de configuración según el kernel seleccionado.

d) Evaluación del modelo

El objetivo de este módulo tiene como propósito realizar la evaluación del modelo, las métricas seleccionadas para este apartado son la exactitud, error, precisión, sensibilidad, especificidad y f1 score.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1. Resultados

4.1.1. Analizar los modelos de machine learning y el clasificador de máquinas de soporte vectorial

Para el desarrollo de este objetivo, se realizó la selección del modelo acorde al problema a tratar, para esto se elige el modelo de aprendizaje supervisado de tipo clasificación por las propiedades que este brinda, además se desarrolla el análisis de los principales algoritmos de clasificación que compone este modelo, así mismo, se define el performance de los algoritmos en temas de clasificación de imágenes en las áreas de salud.

El modelo de aprendizaje supervisado de tipo clasificación se caracteriza por el conjunto de datos de entrada $\{(x_1, y_1), \dots, (x_i, y_i)\}$ cuando x_i representa al vector del i-esimo elemento y y_i representa su etiqueta. Los problemas de aprendizaje que son asociados a respuestas cualitativas y tiene como interés que los resultados de las observaciones puedan ser clasificadas en una de las categorías o clases evaluadas a través de las variables de respuesta (Santos, 2018). Para este tipo de modelos existen distintos algoritmos que son enfocados a la clasificación (binaria o multiclase). Los algoritmos que fueron analizados para el proyecto de investigación son k-NN, ANN, NB, DT y SVM; para lo cual se realiza una descripción del modelo y como está formulado matemáticamente.

a) **K-vecinos más próximos (K-NN):** Este algoritmo es utilizado para el reconocimiento de patrones, utilizando los puntos más cercanos al punto

evaluado, estos son determinados mediante métricas para encontrar las distancias, una de las métricas más utilizadas es la distancia euclidiana. El cálculo de la distancia se realiza a través de la ecuación entre los puntos (x, x') determinada por:

$$\rho(x, x') = \sqrt{\sum_{i=1}^{n} (x_i - x'_i)^2}$$

Donde:

ρ : función de distancia euclidiana.

x_i: i – ésimo ejemplo de entrenamiento.

x'_i: i – ésimo ejemplo de predicción.

n: cantidad de atributos

b) Redes neuronales artificiales (ANN): El uso de este algoritmo enfocados a temas de clasificación hace uso de la creación de un perceptrón simple o multi-capa teniendo una forma más sofisticada para la resolución de problemas (Islam et al., 2020), se define matemáticamente como:

$$Salida = b_i + \sum_{j=1}^{n_x} w_{ij} x_i$$

Donde:

w_{ij}: ponderación de la capa de entrada a la de la salida

b_i: valor de polarización

x_i: valor de entrada

Para el valor de salida se realiza la aplicación de una función de activación estos pueden ser consideradas como:

- **Sigmoidea:**

$$Activación(x) = \frac{1}{1 + e^{-x}}$$

- **Tangente hiperbólica:**

$$Activación(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

- **Unidad Lineal rectificada:**

$$Activación(x) = \begin{cases} 0, & para\ x \leq 0 \\ x, & para\ x > 0 \end{cases}$$

- **Unidad Lineal rectificada con fugas:**

$$Activación(x) = \begin{cases} 0.01x, & para\ x < 0 \\ x, & para\ x \geq 0 \end{cases}$$

- **Función Softmax:**

$$Activación(x) = \frac{e^{x_i}}{\sum_{j=1}^{J} e^{x_j}}, Donde\ i = 1,2, \dots, j$$

c) Redes bayesianas (NB): Según Luckert & Schaffer-Kehnert (2015) inca que la existencia de este algoritmo radica en la creación de nodos y las conexiones que existe entre estos, simbolizando sus dependencias existentes entre estos nodos; el teorema de Bayes está definido por:

$$P(h|D) = \frac{P(D|h) * P(h)}{P(D)}$$

Donde:

$P(h)$: es la probabilidad a priori de la hipótesis h.

$P(D)$: es la probabilidad al observar el conjunto de entrenamiento D.

$P(D|h)$: es la probabilidad de observar el conjunto de entrenamiento D en un universo para la verificación de la hipótesis h.

$P(h|D)$: es la probabilidad a posteriori de h cuando se ha observado el con junto de entrenamiento D.

d) Árboles de decisiones (DT): Este algoritmo realiza un mapeo de los atributos y los valores de las características en los datos de entrada o muestra, formando un árbol por distintos tipos de nodos, se define como objetivo principal de este algoritmo predecir el valor de una variable a través de reglas de decisión, esto es definido como:

$$\Delta I(t) = I(t) - \frac{N_{tS}}{N_t} I(t_s) - \frac{N_{tN}}{N_t}(t_N)$$

Donde:

t: representa a un nodo.

N_t: el número total de muestras en el nodo padre t.

N_{tS}: el número total de muestras enviados al nodo *SI*.

N_{tN}: el número total de muestras enviadas al nodo *NO*.

e) **Máquinas de soporte vectorial (SVM):** El propósito de este algoritmo es la utilización de hiperplanos óptimos para la separación de los datos de entrada que pueda maximizar la separación de las clases de los datos de entrenamiento. Se define la siguiente ecuación del hiperplano para casos que son linealmente no separables $y_i(w \cdot \phi(x_i) + b) \geq 1 - \xi_i$, donde la función kernel es determinada por ϕ y la variable de holgura definida por ξ.

Este algoritmo hace uso de kernel para que la dimensión de los datos de entrada pueda ser elevada a una dimensión mayor a fin de que se pueda realizar la separación de los datos a través del uso de un hiperplano, el kernel de función de base radial (RBF) es considerada como uno de los mejores kernels para temas de clasificación esta función es definida como:

$$K(x_i, x_j) = \exp\left(-\frac{\left|\left|x_i - x_j\right|\right|^2}{2\sigma^2}\right)$$

Donde:

σ: la varianza al hiper parámetro.

$\left|\left|x_i - x_j\right|\right|$: es la distancia euclidiana entre dos puntos x_i y x_j.

Se realizó una comparativa de los algoritmos analizados anteriormente aplicados en trabajos de investigación enfocados en el modelo de aprendizaje supervisado para clasificación, de esta manera se evidencia en la Tabla 7 el performance obtenido por los algoritmos de clasificación. Los resultados obtenidos a través de esta comparativa se obtuvieron que en la mayoría de trabajos el performance de SVM obtuvo resultados sobresalientes frente a los otros algoritmos, así también se demuestra la popularidad de este algoritmo en trabajos de investigación enfocados a la clasificación.

Tabla N° 7: Comparativa y análisis de proyectos de investigación

N°	Autor	Algoritmos	Dataset	Exactitud (Accuracy)
1	Sun et al. (2013)	- SVM - DT - K-NN - RNA - Árboles aleatorios	Elaboración propia	- SVM = 93% - DT = 75% - K-NN = 73% - RNA = 89% - Árboles aleatorios = 90%
2	Pedraza (2015)	- Regresión logística - SVM - RNA	MiniMIAS	- Regresión logística = 97,45% - SVM = 98,88% - RNA = 97,45%
3	Ruchika et al. (2020)	- SVM - KNN	MIAS	- SVM = 94% - KNN = 92%
4	Navya et al. (2020)	- SVN - RNA	UCI machine learning repository	- SVM = 67% - RNA = 64%
5	Osisanwo et al. (2017)	- SVM - Random Forest - NB - Árbol de decisiones - RNA	Diabetes dataset	- SVM = 77,34% - Random Forest = 74,74% - NB = 76,30% - Árbol de decisiones = 73,83% - RNA = 75,13%
6	Kazerouni et al. (2020)	- KNN - SVM - RNA - Regresión logística	UCI machine learning repository	- KNN = 91% - SVM = 95% - RNA = 93% - Regresión logística = 95%
7	Díaz (2021)	- SVM - KNN - NB	GSE10886	- SVM = 97% - KNN = 88% - NB= 90%
8	Castro (2015)	- SVM - RNA - Random Forest - NB	Elaboración propia	- SVM = 95,2% - RNA = 94,4 % - Random Forest = 91,2 % - NB = 86,5%
9	Priya et al. (2020)	- PSO-SVM - SVM - KNN - ANN	MIAS	- PSO-SVM = 94.61 % - SVM = 94% - KNN = 65% - ANN = 71%
10	Islam et al. (2020)	- SVM - K-NN - RF - ANN	Wisconsin Breast Cancer dataset	- SVM = 97.14% - K-NN = 97.14% - RF = 95.71% - ANN = 98.57%

Elaboración propia

Para realizar la evaluación de la comparativa en el performance de los algoritmos de aprendizaje según Paulino (2019) se define los criterios y características a evaluar para cada algoritmo, estos criterios son descritos como:

- **Naturaleza de los datos:** Los datos de entrada para los métodos pueden ser discretos, continuos u ambos.
- **Cantidad de datos de entrenamiento**: La cantidad de datos necesarios para que el algoritmo pueda desempeñarse correctamente.
- **Interpretabilidad:** Representa el nivel de facilidad para interpretar los resultados.
- **Velocidad**: El nivel de rapidez para procesar los datos de entrada.
- **Precisión:** Se mide que tan exacto es el resultado según la comparación de los resultados esperados.
- **Manejo de ruido:** Que tan bien manejan los datos que necesitan preprocesamiento para poder ser analizados.
- **Área de dominio:** La limitación de las técnicas respecto a la complejidad de los diferentes dominios.
- **Nivel de complejidad**: El esfuerzo humano requerido para desarrollar la técnica seleccionada.

Tabla N° 8: Muestra de criterios para la evaluación de algoritmos.

Cod.	Criterio	Valor	Puntaje
C1	**Naturaleza de los datos**	Continuo Discreto	1
		Ambos	2
C2	**Cantidad de datos de entrenamiento**	Bajo	2
		Alto	1
C3	**Interpretabilidad**	No Interpretable	0

		Complejo	1
		Fácil	2
C4	**Velocidad**	Bajo	1
		Alto	2
C5	**Precisión**	Bajo (<= 70%)	1
		Medio ([70-90%])	2
		Alto (90%>)	3
C6	**Manejo de ruido**	Bajo	1
		Medio	2
		Alto	3
C7	**Área de dominio**	Bajo	3
		Medio	2
		Alto	1
C8	**Nivel de complejidad**	Bajo	3
		Medio	2
		Alto	1

Fuente: (Paulino, 2019)

Luego de realizar el análisis de los distintos algoritmos y los antecedentes, se designa una puntuación a criterio propio según los criterios descritos en la Tabla 8.

Tabla N° 9: Resultado de comparativas de performance

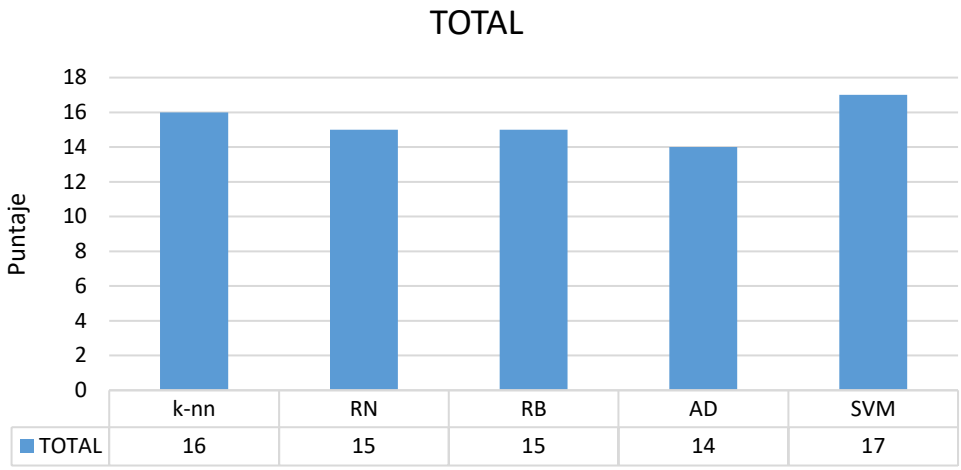
Algoritmo	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**Total**
K-NN	2	1	2	1	3	2	2	3	16
ANN	1	1	1	2	3	3	2	2	15
NB	2	2	1	1	2	3	2	2	15
DT	2	2	2	1	2	1	2	2	14
SVM	2	1	2	1	3	3	3	2	17

Fuente: Elaboración propia.

Los resultados observados a partir de la Tabla 9, luego de realizar la evaluación y puntuación a cada algoritmo según los criterios propuestos, como resultados se obtiene que el algoritmo SVM obtuvo la mayor puntuación con 17 puntos destacándose en los criterios C5, C6 y C7, así mismo, el algoritmo K-NN fue el segundo algoritmo con mejores resultados con una puntuación de 16 puntos y destacándose en los criterios C5 y

C8, los algoritmos ANN y NB ambos algoritmos obtuvieron 15 puntos, y el algoritmos AD obtuvo 14 puntos.

Figura N° 4: Gráfico de barras, puntuación final por algoritmo

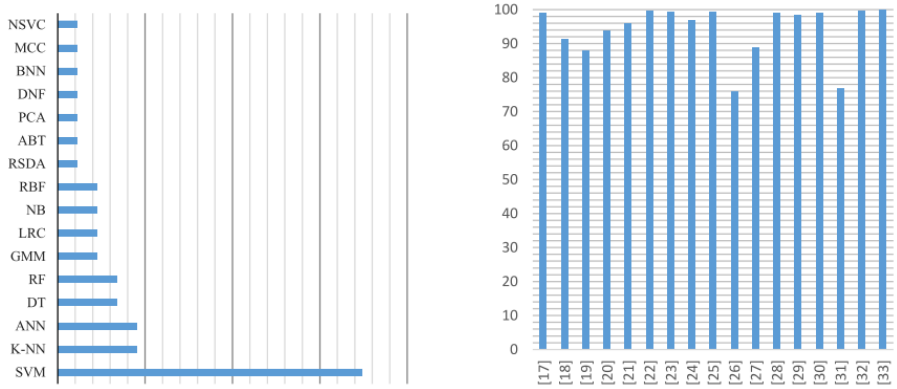


Elaboración propia

De manera similar Tahmooresi et al. (2018) realizaron en su investigación una evaluación y comparación de proyectos de investigación enfocados a la detección y clasificación del cáncer de mama, en su investigación contemplaron los algoritmos de ANN, SVM, K-NN, DT, AdaBoost, NB y arboles aleatorios. Los autores concluyeron que el algoritmo más utilizado es el algoritmo SVM además de tener los mejores resultados aplicándolo solo o en combinación con otros algoritmos, en la Figura 5 se observa los gráficos de evaluación obtenidos por los autores.

Figura N° 5: Resultados obtenidos a través del análisis comparativo

Fuente: (Tahmooresi et al., 2018)



Nota. La imagen del lado izquierdo representa el grado de popularidad por cada algoritmo de clasificación, la imagen del lado derecho representa en nivel de exactitud obtenida en los proyectos de investigación analizados.

El trabajo desarrollado por Bernilla (2021) en su pro en su proyecto de tesis, donde realizó un análisis comparativo de los principales algoritmos para la detección de subtipos de cáncer como se aprecia en la Tabla 10, el autor tomo como referencias de evaluación las métricas de precisión, recall, f1 y error. Los resultados obtenidos por el autor demostraron que el algoritmo SVM obtuvo mejores resultados en comparación de los demás algoritmos de clasificación.

Tabla N° 10: Análisis comparativo de algoritmos

Algoritmos	Precisión	Recall	F-Measure	Error
NB	0,95	0,94	0,94	0,10
ZeroR	0,55	0,74	0,63	0,38

SVM	0,98	0,98	0,98	0,02
K-NN	0,81	0,74	0,64	0,20
MLP	0,67	0,72	0,68	0,40

Fuente: (Díaz Bernilla, 2021)

4.1.2. Implementar un modelo de machine learning utilizando el clasificador de máquinas de soporte vectorial

Para realizar este objetivo específico se inició con el entrenamiento del modelo de machine learning a través de los procesos de adquisición de datos, preprocesamiento, extracción de características y clasificación; así mismo se realizó el desarrollo de una interfaz web para la interacción con el usuario y poder realizar un fácil manejo de las imágenes mamográficas para realización de las pruebas.

a) Adquisición de la base de datos

Para el entrenamiento del modelo se adquirió el dataset MiniMIAS, así también la adquisición de los dataset DDSM, BCDR e imágenes del HRMNB para las pruebas del modelo entrenado. Para obtener las imágenes mamográficas del HRMNB se contó con el apoyo del departamento de diagnóstico por imágenes, se analizaron 10 caso donde cada caso está compuesto por una mamografía del lado izquierdo y derecho. Para incrementar la cantidad de imágenes del HRMNB se utilizó la librería PILLOW de Python para la manipulación e incremento de las imágenes.

b) Pre procesamiento de imágenes

Para el entrenamiento el dataset MiniMIAS cuenta con imágenes de 1024x1024 pixeles, obteniendo una matriz de 1,048,576 pixeles, se tiene un fondo negro alrededor de la mama, asimismo cada anormalidad puede estar ubicada en cualquier parte de la imagen sin tener una ubicación fija, para el reconocimiento de las anormalidades se necesitaría el apoyo de un especialista que defina el área donde se ubica la anormalidad, el dataset MiniMIAS ya cuenta con las áreas definidas que presenten alguna anormalidad, se define un total de 186 imágenes de las cuales estuvo compuesto de las clases benigno, maligno y normal según está definido en la Tabla 11 y este conjunto de imágenes se separó en datos para entrenamiento, test y pruebas del modelo.

Tabla N° 11: Número de imágenes por cada clase

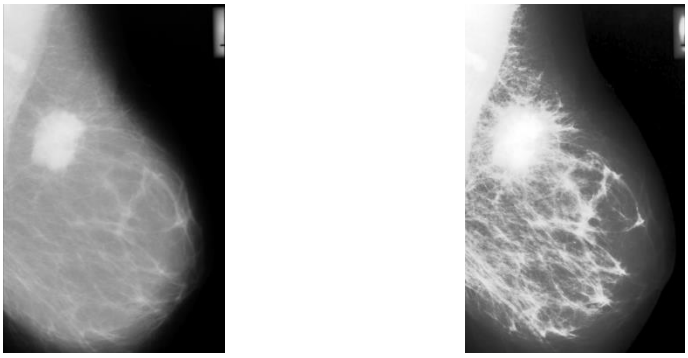
Tipo	Cantidad
Benigno	64
Maligno	52
Normal	70
TOTAL	**186**

Elaboración propia

Para mejorar la relación entre el área de estudio y la existencia del ruido de fondo, se utilizó el filtro gaussiano utilizado para el procesamiento de las imágenes logrando un suavizado de la imagen para que mejore el contraste, además del uso de la ecualización del histograma para uniformizar los distintos niveles de intensidad.

Figura N° 6: Mejora de la imagen a través de los filtros

Fuente: Elaboración propia.



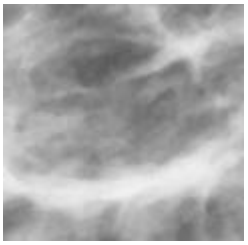
(a) (b)

Nota. La representación de las imágenes está compuesta por una imagen original (a) sin ningún tipo de tratamiento, mientras que la imagen (b) representa a una mamografía aplicando mejoras

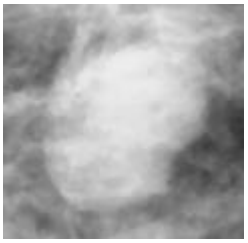
c) Extracción de características

Como se mencionó en el apartado del pre procesamiento el dataset MiniMIAS cuenta con la información de las áreas o regiones de interés (ROI) de cada imagen, para un fácil manejo se realiza el recorte de las imágenes de 150x150 pixeles tal como se aprecia en la Figura 7 además de obtener una matriz de 22,500 pixeles por cada imagen. Para esta sección se aplicó un algoritmo que permita separar la anomalía con respecto al fondo a través del uso de la transformada de hough por ser tolerante a los ruidos encontrados y pueda delimitar el área anómala.

Figura N° 7: Muestra de las regiones de interés



(a)



(b)

Elaboración propia.

Nota. La imagen (a) representa a una región de interés normal, mientras que la imagen (b) representa una anormalidad.

d) Clasificación

Para este proceso se hace uso de la selección del algoritmo SVM para realizar una clasificación múltiple, el entrenamiento del modelo se realizó con un 70% para el entrenamiento y un 30% para las pruebas de validación, se toma estos porcentajes a partir del análisis de la literatura, de tal manera poder evitar los casos de sobre ajuste.

El algoritmo SVM se basa en el uso de planos de decisión el cual define los límites de las clases a partir de los datos de entrada $X = \{(x_1, y_1), \dots, (x_n, y_n)\}$ donde n es el vector de entrada en un espacio de $\mathbb{R}^d$ y y_n denota el índice de clase.

Para la implementación del algoritmo, se hizo uso de las librerías de Python como scikit-learn y libsvm, el cual implementa los procedimientos para el algoritmo SVM, estas herramientas dieron la posibilidad de realizar una clasificación multi clase, para este caso se realizó la definición de parámetros C y γ re queridos para el kernel RBF como se aprecia en la Tabla 12.

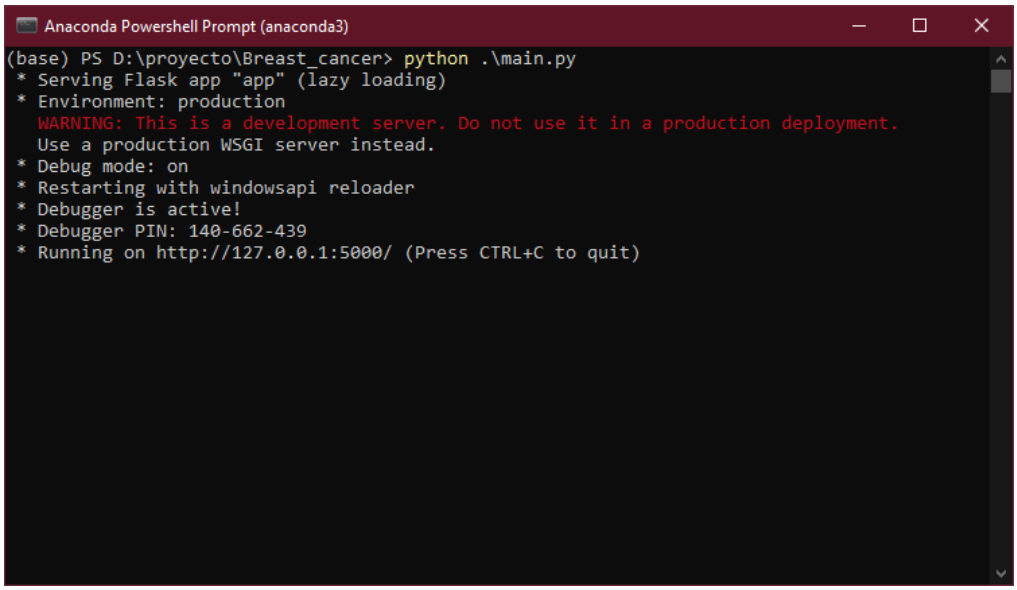
Tabla N° 12: Parámetros para el kernel RBF

Costo C	Función Kernel	Parámetro γ	Tipo de Problema
100	$k(x^{(i)}, x^{(j)}) = \exp(-\gamma \|x^{(i)} - x^{(j)}\|^2)$	2^{-9}	Multi-clasificación

Elaboración propia

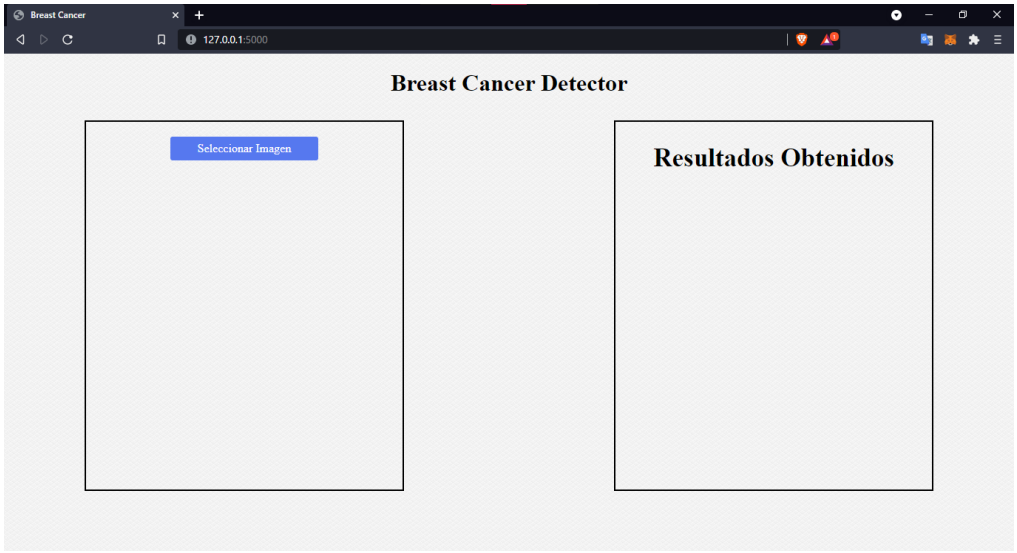
Asimismo, se desarrolló un interfaz gráfico desarrollado en Flask un Framework de Python especializada en el desarrollo de aplicaciones rápidas, esta aplicación se desarrolló para el fácil manejo de las imágenes y visualización de los resultados que se muestre a través del modelo entrenado, en las Figuras 8, 9 y 10 se aprecia el interfaz web desarrollado.

Figura N° 8: Inicio del servidor de la interfaz web



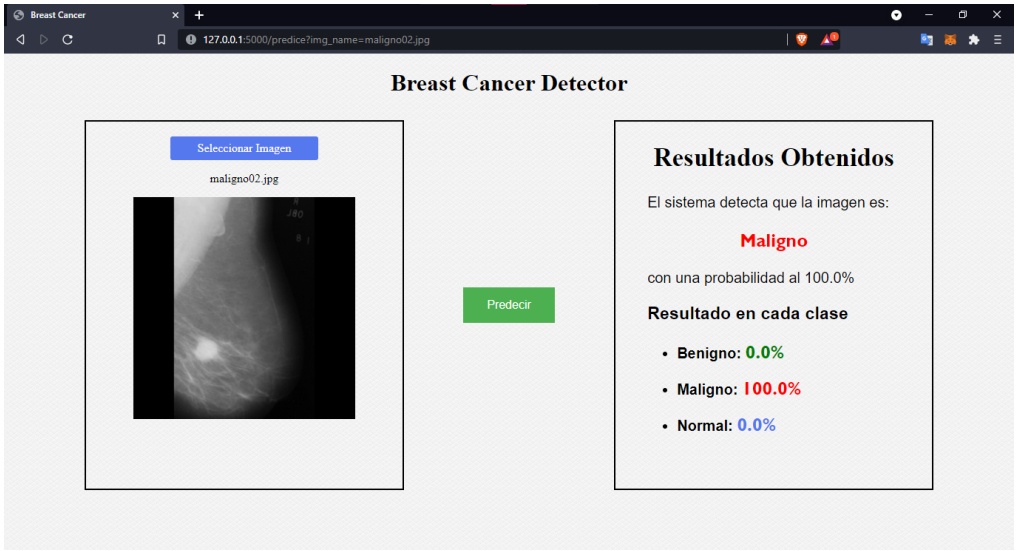
Elaboración propia.

Figura N° 9: Interfaz web



Elaboración propia.

Figura N° 10: Resultado de la evaluación de la imagen



Elaboración propia.

El resultado obtenido a través del desarrollo de esta aplicación luego de realizar la implementación del modelo de aprendizaje, evidencia una mejor manera de realizar la selección y visualización de las imágenes mamográficas, así también observar el resultado a través del clasificador SVM pudiendo clasificar entre las clases benignos, malignos y normales.

4.1.3. Identificar el nivel de precisión del modelo de machine learning con el clasificador SVM para la detección y clasificación del cáncer de mama

En este apartado se mostrará los resultados del objetivo específico obtenidos del modelo entrenado, para realizar su validación se utilizó una matriz de confusión y el uso de una validación cruzada k-10 fold. Las métricas utilizadas para la matriz de confusión se definen en la Tabla 13.

Tabla N° 13: Métricas para la evaluación de los modelos de aprendizaje

Métricas para la evaluación	
Exactitud (accuracy)	$\frac{Verdaderos\ Positivos}{Total}$
Precisión	$\frac{VP}{FP + VP}$
Sensibilidad (recall)	$\frac{VP}{VP + FN}$
Especificidad	$\frac{VN}{VN + FP}$
F1 score	$F1 = 2 * \frac{Precisión * Recall}{Precisión + Recall}$

Elaboración propia

Donde:

VP: Verdaderos Positivos

VN: Verdaderos Negativos

FP: Falso Positivo

FN: Falso Negativo

- Los resultados obtenidos en la matriz de confusión descrita en la tabla 14, nos indica que el total de imágenes mamográficas clasificados como ‘Benigno’ fueron clasificados 31 correctamente por el algoritmo clasificador, así mismo

las clases definidas como ‘Maligno’ tuvieron 26 elementos clasificados como correctos y las mamografías catalogadas como ‘Normal’ fueron clasificados 33 correctamente por el algoritmo clasificador.

Tabla N° 14: Resultados de la matriz de confusión

Matriz de confusión		Predicciones		
		Benigno	Maligno	Normal
Reales	Benigno	**31**	1	3
	Maligno	4	**26**	1
	Normal	2	0	**33**

Elaboración propia

- En la Tabla 15 se muestra el resumen de la cantidad obtenida de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos de cada clase obtenida de la matriz de confusión de la Tabla 14.

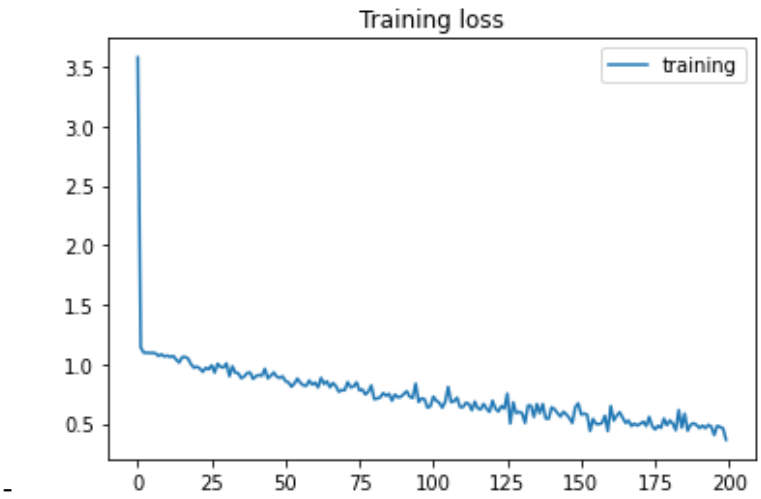
Tabla N° 15: Resumen de la evaluación de la matriz de confusión por clases

Benigno		Maligno		Normal	
Verdadero Positivo (VP)	31	Verdadero Positivo (VP)	26	Verdadero Positivo (VP)	33
Falso Positivo (FP)	4	Falso Positivo (FP)	5	Falso Positivo (FP)	2
Verdadero Negativo (VN)	60	Verdadero Negativo (VN)	69	Verdadero Negativo (VN)	62
Falso Negativo (FN)	6	Falso Negativo (FN)	1	Falso Negativo (FN)	4

Elaboración propia

- En la figura 11 se puede observar que la tasa de error se va disminuyendo progresivamente a medida que se realiza el entrenamiento, estabilizándose alrededor de $[1, 0.5]$.

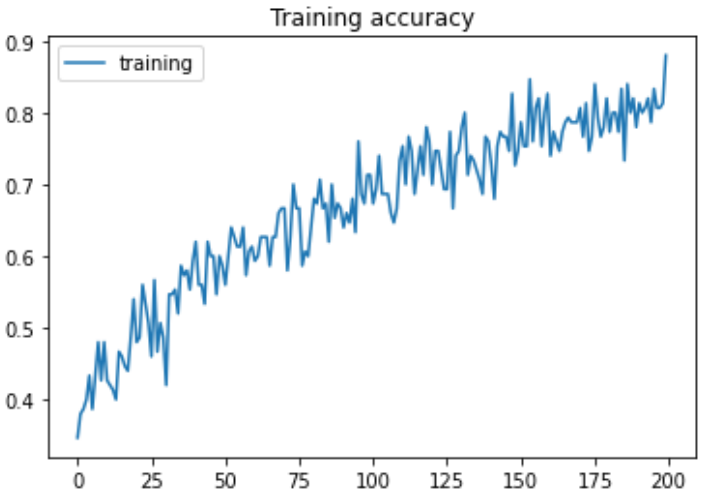
Figura N° 11: Tasa de error del modelo entrenado



Elaboración propia.

- En la figura 12 se observa la tasa de precisión, a medida que el entrenamiento del modelo se va realizando la exactitud va incrementándose hasta alcanzar picos de alrededor de 0.9, resultando que el nivel de precisión sea aceptable y se demuestra que el grado de confiabilidad es alto para el modelo que se está proponiendo.

Figura N° 12: Tasa de exactitud (accuracy) del modelo entrenado



Elaboración propia.

- En la Tabla 16 se muestran los resultados sobre la evaluación de los desempeños del modelo entrenado para las clases benigno, maligno y normal, para la evaluación se toma las métricas de precisión, sensibilidad y f1 score.

Tabla N° 16: Resultados de precisión, sensibilidad y f1 score

Clases	Precisión	Sensibilidad	F1
Benigno	0.886	0.838	0.861
Maligno	0.839	0.963	0.897
Normal	0.943	0.892	0.917

Elaboración propia

- Los resultados indican que la precisión con mayor puntuación es de 0.943 en la clase ‘Normal’, seguido de la clase ‘Benigno’ que obtuvo 0.886 y con una menor puntuación en la clase ‘Maligno’ que obtuvo 0.839. Para los resultados de la sensibilidad se obtuvieron que la clase ‘Maligno’ esta con una mayor puntuación de 0.963, seguido de la clase ‘Normal’ que obtiene 0.892 y la clase

'Benigno' obtuvo una menor puntuación de 0.838. Para f1 los resultados para las clases 'Normal', 'Benigno' y 'Maligno' obtuvieron un resultado de 0.917, 0.897 y 0.861 respectivamente.

- El método k-fold, consta de la división de los datos en forma aleatoria de k grupos del mismo tamaño, donde $k - 1$ grupos son empleados para lograr el entrenamiento del modelo y uno de los grupos empleándose como validación. En la Tabla 17 se observa el resultado de este método de validación, donde el conjunto de entrenamiento es de 10, los resultados obtenidos para el conjunto de entrenamiento número 7 se obtiene una exactitud del entrenamiento del 93.99% y una exactitud en el test del 90%.

Tabla N° 17: Resultados de evaluación de la validación cruzada

Train Set	Acc Train	Acc Test	Tiempo
1	88%	87%	1h 20min
2	89.67%	89%	1h 22min
3	89.25%	86%	1h 17min
4	91.5%	90%	1h 24min
5	87%	85%	1h 21min
6	92%	91%	1h 22min
7	93.99%	90%	1h 20min
8	88.75%	88%	1h 15min
9	90.5%	89%	1h 05min
10	90%	86%	1h 29min

Elaboración propia

4.1.4 Desarrollar un modelo de machine learning utilizando un clasificador de máquinas de soporte vectorial para detectar y clasificar el cáncer de seno a partir de imágenes mamográficas

Para el desarrollo de este objetivo se realizó la prueba del modelo entrenado usando otros dataset para comprobar si es posible la clasificación, para ellos se contó con el uso de los dataset DDSM, BCDR y HRMN, para cada datset se realiza una selección de imágenes que estén en las clases normal, benigno y maligno.

Tabla N° 18: Resultado del modelo aplicado a otros datasets

Dataset	Clases			Total	Clasificados		% clasificados
	N	B	M	Img.	Correctos	Incorrectos	correctamente
DDSM	40	40	40	120	107	13	89%
BCDR	40	40	40	120	101	19	84%
HRMNB	70	25	5	100	83	17	83%

Elaboración propia.

Los resultados mostrados en la Tabla 18, demuestran que el modelo en general es capaz de realizar clasificaciones utilizando otros dataset con imágenes mamográficas proporcionados por organizaciones enfocadas a la investigación científica y el uso de un dataset obtenida de un entorno local del centro de salud de Puno. Para los dataset DDSM y BCDR la cantidad de imágenes seleccionadas aleatoriamente para cada clase suman un total de 120 mamografías y un total de 100 imágenes obtenidas del HRMNB. Para los resultados obtenidos indica que el modelo fue capaz de clasificar correctamente el dataset DDSM en un 89% donde 107 fueron clasificadas correctamente y 13 incorrectas; para el dataset BCDR se obtuvo un 84% de clasificaciones correctamente realizadas donde 101 imágenes son clasificadas correctamente y 19 fueron clasificadas incorrectas; para el

conjunto de imágenes mamográficas del HRMNB se aprecia un 83% de clasificaciones realizadas correctamente donde el número de figuras correctamente clasificadas fueron 83 y 17 son clasificadas incorrectamente.

V. CONCLUSIONES

Primero: Para el análisis de los modelos de aprendizaje se abordó los algoritmos K-NN, ANN, NB, DT y SVM enfocados a temas de clasificación realizando una descripción y enfoques matemáticos, así mismo, se plantearon 8 criterios de avaluación y la evaluación del performance de los algoritmos en proyectos de investigación enfocados a la clasificación de imágenes, dando como resultado que el algoritmo SVM es el más utilizado para estos tipos de problemas y obteniendo un puntaje de 17 a partir de los resultados de los criterios de evaluación. Luego de evaluar los resultados del análisis de los algoritmos y realizar la comparación de estos resultados con otras investigaciones, se concluye que el algoritmo SVM es una de las mejores alternativas para clasificación de imágenes médicas.

Segundo: Para la implementación se contó con datasets proporcionadas por plataformas digitales, así como la adquisición de imágenes de un entorno real ubicadas en la ciudad de Puno, al ser imágenes locales se tuvo la necesidad de realizar un tratamiento de datos manuales, eliminando datos personales de los pacientes y médicos por la confidencialidad medico paciente, así mismo realizar el tratamiento de estas imágenes para las pruebas con el modelo desarrollado. La aplicación del filtro gaussiano y la ecualización de histograma mejoraron significativamente la calidad de las imágenes asegurando mejores resultados en las demás etapas del proceso. La clasificación del algoritmo SVM uso las variables $C = 100$ y $\gamma = 2^{-9}$ utilizadas por la librería de Python para el uso del kernel RBF. Se implementó una interfaz web para el fácil manejo de las imágenes y el desarrollo del modelo, esta interfaz mejora el manejo de las pruebas con imágenes.

Tercero: El modelo desarrollado fue evaluado a través de métricas como la exactitud, error, precisión, sensibilidad y f1 score, tuvo una exactitud del 90% con un margen de

error del 10%. La evaluación de las métricas de precisión, sensibilidad y f1 escore se realizó por cada clase evaluada, donde los resultados obtenidos de la clase benigno tuvieron un 88.6%, 83.8% y 86.1%, para la clase maligno se obtuvo un 83.9%, 96.3% y 89.7%, la clase normal obtuvo un 94.3%, 89.2% y 91.7% respectivamente a las métricas evaluadas por cada clase. A través de la evaluación cruzada se vio un aumento en un 93.99% en la exactitud del entrenamiento. Estas métricas demuestran que el modelo no tiende a problemas de sobre ajuste o sub ajuste, así mismo se demuestra que a través del análisis de trabajos enfocados en la detección del cáncer de mama, el modelo propuesto logra competir con otros trabajos de investigación que fueron analizados en este proyecto de tesis.

VI. RECOMENDACIONES

Primero: Para futuras investigaciones, se recomienda realizar el análisis comparativo entre los modelos de machine learning y deep learning enfocados a temas de clasificación de imágenes en el área de salud.

Segundo: Para el mejoramiento de los datos de entrada al modelo, se recomienda el uso de algoritmos de aprendizaje supervisados que puedan ser enfocados en la etapa de extracción de características y usar el algoritmo de SVM como clasificador para comprobar el mejoramiento de la exactitud del modelo de clasificación.

Tercero: Se recomienda en futuras investigaciones realizar la comparación del modelo utilizando el algoritmo SVM para clasificación múltiple a través del enfoque One-against-All (OAA) y One-against-One (OAA), esto proporcionara una mejor perspectiva sobre el desempeño del modelo a través de estos enfoques.

VII. REFERENCIAS

ACS. (2019). Breast Cancer Facts & Figures. *Occupational Cancers*, 417–438. https://doi.org/10.1007/978-3-030-30766-0_24

Afifi, S., GholamHosseini, H., & Sinha, R. (2019). A system on chip for melanoma detection using FPGA-based SVM classifier. *Microprocessors and Microsystems*, *65*, 57–68. https://doi.org/10.1016/j.micpro.2018.12.005

Aguilar, L., & Vásquez, Y. (2016). *Principal component analysis (PCA) para mejorar la performance de aprendizaje de los algoritmos Support Vector Machine (SVM) y Red Neuronal Multicapa (MLNN)"*. http://www.gonzalezcabeza.com/documentos/CRECIMIENTO_MICROBIANO.pdf

Akram, Z. (2019). *Detection and Classification of Mammographic Abnormalities*. https://core.ac.uk/download/pdf/196350286.pdf

Alvarez Mayta, J. A. (2014). *Sistema de detección de cáncer de mama en mujeres, mediante el uso de redes neuronales*. 93.

Anvesh, M. (2014). *Machine Learning Approaches for Breast Cancer Diagnosis and their Comparison*. *May*.

Arnedo, C. J. V. (2016). *Sistema predictivo progresivo de clasificación probabilística como guía para el aprendizaje*. http://hdl.handle.net/10045/54256

Awad, M., & Khanna, R. (n.d.). *Efficient Learning Machines. Theories, Concepts, and Applications for Engineers and System Designers*.

Balas, V. E., Kumar, V., Raghvendra, S., & Editors, K. (2020). *Intelligent Systems*

Reference Library 180 Internet of Things and Big Data Applications Recent Advances and Challenges. http://www.springer.com/series/8578

Bardales Bruno, E. (2013). *Diseño de un algoritmo de reconocimiento en la identificación de tumores de mama*. *February*, 102–104.

Bejnordi, B. E. (2017). *Histopathological diagnosis of breast cancer using machine learning*. 190.

Bell, J. (2015). Machine Learning: Hands-On for Developers and Technical Professionals. In *Paper Knowledge . Toward a Media History of Documents*.

Bhattacharjee, J. (2020). Practical Machine Learning with Rust. In *Practical Machine Learning with Rust*. https://doi.org/10.1007/978-1-4842-5121-8

Bhavsar, P., Safro, I., Bouaynaya, N., Polikar, R., & Dera, D. (2017). Machine Learning in Transportation Data Analytics. In *Data Analytics for Intelligent Transportation Systems*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-809715-1.00012-2

Bueno, I. (2018). *Exploring the intersections between Information Visualization and Machine Learning* [Universidade de São Paulo]. http://www.teses.usp.br/teses/disponiveis/55/55134/tde-02012019-110149/

Calderón Niquin, M. A. (2012). *Detección de metástasis de cáncer mamario usando Máquinas de Soporte Vectorial a partir de datos de microarray*.

Carmona, E. (2014). *Tutorial sobre Máquinas de Vectores Soporte (SVM)*. Universidad Nacional de Educación a Distancia.

Castro, G. (2015). *Aplicación de algoritmos inteligentes para reconocimiento automático de enfermedades foliares de cultivo de palta*. *052*, 1–18.

Cazap, E., Buzaid, A., Garbino, C., de la Garza, J., Orlandi, F., Schwartsmann, G., Vallejos, C., Guercovich, A., & Breitbart, G. (2010). Breast cancer in Latin America: Experts perceptions compared with medical care standards. *Breast*, *19*(1), 50–54. https://doi.org/10.1016/j.breast.2009.10.011

Chandra, S., Srujan, S. K., Krishna, K. S. D. R., & Editors, M. N. F. (2020). *Learning and Analytics in Intelligent Systems 3 Advances in Decision Sciences , Image Processing , Security and Computer Vision* (Vol. 3).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *IEEE Expert-Intelligent Systems and Their Applications*, 273–297. https://doi.org/10.1109/64.163674

Díaz Bernilla, N. M. (2021). *Análisis comparativo de clasificadores para la detección de subtipos de cáncer*.

Diaz, J., Ruibal, A., & Tejerina, A. (2012). *Cáncer de mama Aspectos de interés actual*.

Eche Zapata, G. M. (2017). *Exploración de técnicas automáticas de detección de líneas-b en imágenes de ultrasonido para diagnóstico de neumonía en pacientes pediátricos*.

Escudero, E. (2018). Diseño de un árbol de decisión bayesiana para el tratamiento del carcinoma ductal in situ de mama. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699.

Fletcher, T. (2009). Support Vector Machines Explained. *Online]. Http://Sutikno. Blog. Undip. Ac. Id/Files/2011/11/SVM-Explained. Pdf.[Accessed 06 06 2013]*, 1–19. http://sutikno.blog.undip.ac.id/files/2011/11/SVM-Explained.pdf

Geras, K. J. (2011). Prediction Markets for Machine Learning. *Artificial Intelligence*, *1*, 76.

GLOBOCAN. (2021). *Incidencia de cancer en el mundo*. Obtenido de https://gco.iarc.fr/today/online-analysis-pie?v=2020&mode=cancer&mode_population=continents&population=900&populations=900&key=total&sex=0&cancer=39&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=7&group

González, R., Barrientos, A., Toapanta, M., & Del Cerro, J. (2017). Aplicación de las Máquinas de Soporte Vectorial (SVM) al diagnóstico clínico de la Enfermedad de Párkinson y el Temblor Esencial. *RIAI - Revista Iberoamericana de Automatica e Informatica Industrial*, *14*(4), 394–405. https://doi.org/10.1016/j.riai.2017.07.005

Guades Santos, T. A. (2016). *Weighted Multiple Kernel Learning for Breast Cancer Diagnosis applied to Mammograms*. Universidade do Porto.

Hamidinekoo, A., Denton, E., Rampun, A., Honnor, K., & Zwiggelaar, R. (2018). Deep learning in mammography and breast histology, an overview and future trends. *Medical Image Analysis*, *47*, 45–67. https://doi.org/10.1016/j.media.2018.03.006

Havrylchyk, O., & Poncet, S. (2007). Machine Learning Paradigms: Applications in Recommender Systems. In *The World Economy* (Vol. 30, Issue 11). https://books.google.com/books?id=Mb7pCQAAQBAJ&pgis=1

Hepsağ, P. U., Özel, S. A., & Yazici, A. (2017). Using deep learning for mammography classification. *2nd International Conference on Computer Science and Engineering, UBMK 2017*, 418–423. https://doi.org/10.1109/UBMK.2017.8093429

Hernández, S. (2014). *Metodología de la investigación*.

IARC. (30 de Agosto de 2021). *GLOBOCAN*. Obtenido de Estimated age-standardized incidence and mortality rates (World) in 2020, worldwide, females, ages 45-69: https://gco.iarc.fr/today/online-analysis-multi-bars?v=2020&mode=cancer&mode_population=countries&population=900&populations=900&key=asr&sex=2&cancer=39&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=9&ages_group%5B%5D=13&nb_items=10&

INC. (9 de Febrero de 2015). *Instituto Nacional del Cancer*. Obtenido de https://www.cancer.gov/espanol/cancer/naturaleza/que-es

INEI. (2018). Enfermedades No Transmisibles y Transmisibles. *Perú Enfermedades No Transmisibles y Trasmisibles, 2018*, *53*(9), 1–192.

Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. *SN Computer Science*, *1*(5), 1–14. https://doi.org/10.1007/s42979-020-00305-w

Jayaram, V., Avseth, P. A., Azbel, K., Coléou, T., Devegowda, D., De Groot, P., Gao, D., Marfurt, K., Matos, M., Mukerji, T., Poupon, M., Roy, A., Russell, B., Wallet, B., & Kumar, V. (2015). Introduction to special section: Pattern recognition and machine learning. In *Interpretation* (Vol. 3, Issue 4). https://doi.org/10.1190/INT2015-0918-SPSEINTRO.1

Kalali, A., Richerson, S., Ouzunova, E., Westphal, R., & Miller, B. (2019). Digital Biomarkers in Clinical Drug Development. In *Handbook of Behavioral Neuroscience* (1st ed., Vol. 29). Elsevier B.V. https://doi.org/10.1016/B978-0-12-

803161-2.00016-3

Kaur, P., Singh, G., & Kaur, P. (2019). Intellectual detection and validation of automated mammogram breast cancer images by multi-class SVM using deep learning classification. *Informatics in Medicine Unlocked*, *16*(January), 100151. https://doi.org/10.1016/j.imu.2019.01.001

Kazerouni, F., Bayani, A., Asadi, F., Saeidi, L., Parvizi, N., & Mansoori, Z. (2020). Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression: A comparison of four data mining approaches. *BMC Bioinformatics*, *21*(1), 1–13. https://doi.org/10.1186/s12859-020-03719-8

Kirk, M. (2015). *Thoughtful Machine Learning*. O'Reilly Media.

Krithiga, R., & Geetha, P. (2021). Breast Cancer Detection, Segmentation and Classification on Histopathology Images Analysis: A Systematic Review. *Archives of Computational Methods in Engineering*, *28*(4), 2607–2619. https://doi.org/10.1007/s11831-020-09470-w

Luckert, M., & Schaffer-Kehnert, M. (2015). *Using Machine Learning Methods for Evaluating the Quality of Technical Documents*. 102. http://www.diva-portal.org/smash/get/diva2:920202/FULLTEXT01.pdf

Matich, D. (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones. *Historia*, 55. ftp://decsai.ugr.es/pub/usuarios/castro/Material-Redes-Neuronales/Libros/matich-redesneuronales.pdf

MINSA. (2017). *Plan Nacional Para La Prevención Y Control De Cáncer De Mama En El Peru 2017-2021*. http://bvs.minsa.gob.pe/local/MINSA/4234.pdf

MINSA. (2018). Boletín Epidemiológico del Perú. *Boletin Epidemiològico Del Perù*,

27(31), 698–699. https://www.dge.gob.pe/portal/docs/vigilancia/boletines/2018/31.pdf

MINSA. (2020). Análisis de la situación del cáncer en el Perú, 2018. In *Centro Nacional de Epidemiología, Prevención y Control de Enfermedades* (1ra. Edici).

MINSA. (2021). Programa Presupuestal 0024 Prevencion Y Control Del Cancer. *Ministerio de Salud*, 1–303.

Mitchell, T. (1997). Machine Learning. *McGraw Hill*, 870–877.

Ñaupas, H., Valdivia, M., Palacios, J., & Romero, H. (2018). Metodología de la investigación cuantitativa-cualitativa y redacción de la tesis. In *Journal of Chemical Information and Modeling* (5a.Edición, Vol. 53, Issue 9). https://doi.org/10.1017/CBO9781107415324.004

OMS. (26 de Marzo de 2021). *Cáncer de mama*. Obtenido de https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer

OMS. (26 de Marzo de 2021). *Organización Mundial de la Salud*. Obtenido de Cáncer de mama: https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer

OMS. (s.f.). *Organización Mundial de la Salud*. Recuperado el 09 de Julio de 2021, de Cáncer: https://www.who.int/es/health-topics/cancer#tab=tab_1

OPS. (2016). *Garantía de calidad de los servicios de mamografía: Normas básicas para América Latina y el Caribe*.

OPS. (s.f.). *Organización Panamericana de la Salud - Cáncer*. Obtenido de https://www.paho.org/es/temas/cancer#:~:text=El%20c%C3%A1ncer%20es%20una%20de,1%20millones%20en%20el%202030.

OPS. (s.f.). *Organización Panamericana de la Salud*. Recuperado el 09 de Julio de 2021, de Cáncer: https://www.paho.org/es/temas/cancer

Orozco, R., Suarez aday, E., & Perez diaz, M. (2020). *Diseño de Sistema Automatizado para Detección de Anomalías en Imágenes Digitales de Mama*. *January*.

Ortiz, R. (2019). *Radiomics for diagnosis and assessing brain diseases: an approach based on texture analysis on magnetic resonance imaging* [Universitat Politècnica de València]. https://dialnet.unirioja.es/servlet/tesis?codigo=250441&info=resumen&idioma=SPA%0Ahttps://dialnet.unirioja.es/servlet/tesis?codigo=250441

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., & Akinjobi, J. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, *48*(3), 128–138. https://doi.org/10.14445/22312803/ijctt-v48p126

Paulino Flores, L. A. (2019). *Sistema experto probabilístico basado en redes bayesianas para la predicción de riesgo de cáncer cervical*.

Pedraza, C. (2015). *Implementación De Técnicas De Machine Learning Para La Identificación Asistida De Lesiones Tumorales en Imágenes Médicas*. Universidad de los Andes.

Pinillo, V. baffigo torre de. (2016). Detección temparana del cáncer de mama en EsSalud. *2016*, 32. http://www.essalud.gob.pe/ietsi/pdfs/guias/DIREC_DETECCION_TEMP_CANCER_MAMA.pdf

Priya, T. S., & Ramaprabha, T. (2020). An Effective Feature Extraction Based Particle

Swarm Optimization with Support Vector Machine for Biomedical Mammogram Image Diagnosis. *Proceedings of the 5th International Conference on Inventive Computation Technologies, ICICT 2020*, 348–352. https://doi.org/10.1109/ICICT48043.2020.9112486

Ramos, W., & De La Cruz-Vargas, J. A. (2020). Presentación del documento técnico "Análisis de la situación del cáncer en el Perú, 2018." *Revista de La Facultad de Medicina Humana*, *20*(1), 10–11. https://doi.org/10.25176/rfmh.v20i1.2704

Raschka, S. (2014). Python Machine Learning. Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics. In *Paper Knowledge . Toward a Media History of Documents*.

Rodríguez, G. (2009). *Detección de Microcalcificaciones utilizando Discriminantes Lineales de Fisher*.

Romieu, I., Biessy, C., Torres-Mejía, G., Ángeles-Llerenas, A., Sánchez, G. I., Borrero, M., Ossa, C. A., Porras, C., Rodríguez, A. C., Ocampo, R., Garmendia, M. L., Bustamante, E., Olivier, M., Porter, P., & Rinaldi, S. (2019). Project profile: A multicenter study on breast cancer in young women in Latin America (PRECAMA study). *Salud Publica de Mexico*, *61*(5), 601–608. https://doi.org/10.21149/10466

Samuel, A. L. (1959). Eight-move opening utilizing generalization learning. (See Appendix B, Game G-43.1 Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal*, 210–229.

Santos, H. G. dos. (2018). *Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina*. 207. http://www.teses.usp.br/teses/disponiveis/6/6141/tde-09102018-132826/

Shalev-Shwartz, S., & Ben-David, S. (2013). Understanding machine learning: From theory to algorithms. In *Understanding Machine Learning: From Theory to Algorithms* (Vol. 9781107057). https://doi.org/10.1017/CBO9781107298019

Shihong, Y., Ping, L., & Peiyi, H. (2003). *SVM Classification: Its contents and challenges*. *8*, 332–342.

Sucar, L. E. (2011). Introduction to bayesian networks and influence diagrams. *Decision Theory Models for Applications in Artificial Intelligence: Concepts and Solutions*, 9–32. https://doi.org/10.4018/978-1-60960-165-2.ch002

Sumbaly, R., Vishnusri, N., & Jeyalatha, S. (2021). Diagnosis of Breast Cancer using Decision Tree Data Mining Technique. *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, *98*(10), 1530–1535. https://doi.org/10.1109/ICACCS51430.2021.9442043

Sun, T., Wang, J., Li, X., Lv, P., Liu, F., Luo, Y., Gao, Q., Zhu, H., & Guo, X. (2013). Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set. *Computer Methods and Programs in Biomedicine*, *111*(2), 519–524. https://doi.org/10.1016/j.cmpb.2013.04.016

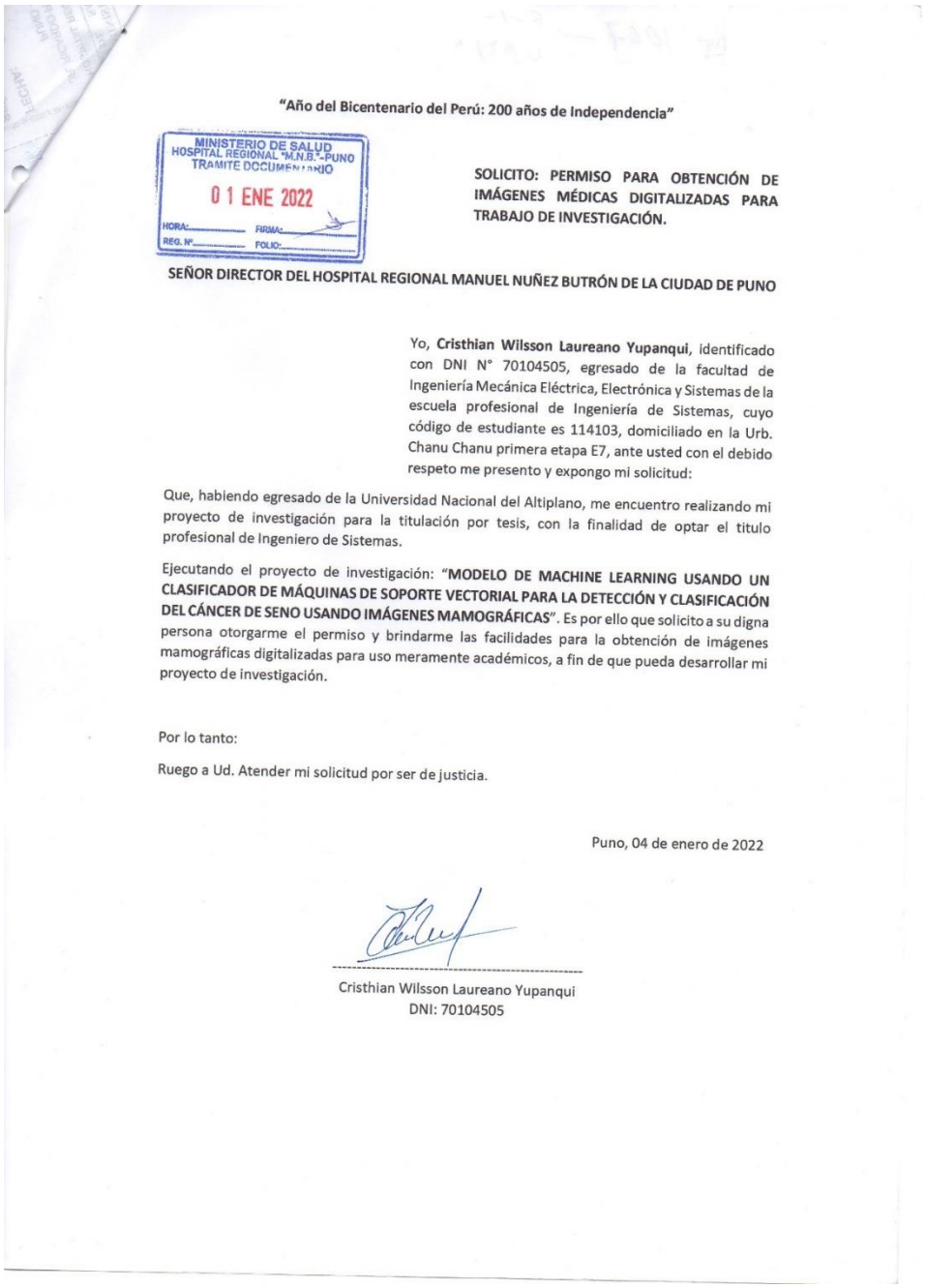
Tahmooresi, M., Afshar, A., Bashari, B., Nowshath, K. B., & Bamiah, M. A. (2018). Early detection of breast cancer using machine learning techniques. *Journal of Telecommunication, Electronic and Computer Engineering*, *10*(3–2), 21–27.

Tahmooresi, M., Afshar, A., Bashari Rad, B., Nowshath, K. B., & Bamiah, M. A. (2018). Early detection of breast cancer using machine learning techniques. *Journal of Telecommunication, Electronic and Computer Engineering*, *10*(3–2), 21–27.

Vijayarajeswari, R., Parthasarathy, P., Vivekanandan, S., & Basha, A. A. (2019). Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform. *Measurement: Journal of the International Measurement Confederation*, *146*, 800–805. https://doi.org/10.1016/j.measurement.2019.05.083

ANEXOS

ANEXO 1: Documento de solicitud de imágenes mamográficas dirigida al Hospital Regional "Manuel Núñez Butron" y comprobante de pago



"Año del Bicentenario del Perú: 200 años de Independencia"

MINISTERIO DE SALUD
HOSPITAL REGIONAL "M.N.B."-PUNO
TRAMITE DOCUMENTARIO
01 ENE 2022
HORA:_____ FIRMA:_____
REG. N°:_____ FOLIO:_____

SOLICITO: PERMISO PARA OBTENCIÓN DE IMÁGENES MÉDICAS DIGITALIZADAS PARA TRABAJO DE INVESTIGACIÓN.

SEÑOR DIRECTOR DEL HOSPITAL REGIONAL MANUEL NUÑEZ BUTRÓN DE LA CIUDAD DE PUNO

Yo, **Cristhian Wilsson Laureano Yupanqui**, identificado con DNI N° 70104505, egresado de la facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas de la escuela profesional de Ingeniería de Sistemas, cuyo código de estudiante es 114103, domiciliado en la Urb. Chanu Chanu primera etapa E7, ante usted con el debido respeto me presento y expongo mi solicitud:

Que, habiendo egresado de la Universidad Nacional del Altiplano, me encuentro realizando mi proyecto de investigación para la titulación por tesis, con la finalidad de optar el titulo profesional de Ingeniero de Sistemas.

Ejecutando el proyecto de investigación: "**MODELO DE MACHINE LEARNING USANDO UN CLASIFICADOR DE MÁQUINAS DE SOPORTE VECTORIAL PARA LA DETECCIÓN Y CLASIFICACIÓN DEL CÁNCER DE SENO USANDO IMÁGENES MAMOGRÁFICAS**". Es por ello que solicito a su digna persona otorgarme el permiso y brindarme las facilidades para la obtención de imágenes mamográficas digitalizadas para uso meramente académicos, a fin de que pueda desarrollar mi proyecto de investigación.

Por lo tanto:

Ruego a Ud. Atender mi solicitud por ser de justicia.

Puno, 04 de enero de 2022

Cristhian Wilsson Laureano Yupanqui
DNI: 70104505

ANEXO 2: Carta de presentación obtenida del director del Hospital Regional "Manuel Núñez Butron" Puno, para el departamento de diagnóstico por imágenes.

PERÚ Ministerio de Salud

REGION DE SALUD PUNO
"HOSPITAL REGIONAL "MANUEL NÚÑEZ BUTRON"
UNIDAD DE APOYO A LA INVESTIGACION Y DOCENCIA
Jr. Ricardo Palma N° 120 – Telefax: 351021 – Telef.: 369696 – 367777

"Año del Fortalecimiento de la Soberanía Nacional"

07 Ene 2022

OFICIO N° 000090 -2022 - UAID - HR "MNB"-PUNO.

Señora.:
NARDY MONTES DE OCA VELASCO
JEFE DEL DPTO. DIAGNOSTICO POR IMAGENES DEL H.R. "MNB" - PUNO

Presente.-

ASUNTO: Presentación de Tesista.

Es grato dirigirme a usted para saludarla y presentar al señor ***CHISTHIAN WILSSON LAUREANO YUPANQUI,*** *egresado de la* ***Universidad Nacional del Altiplano – Escuela Profesional de Ingeniería de Sistemas,*** *quien realizara el Proyecto de Investigación Titulado:* ***"MODELO DE MACHINE LEARNING USANDO UN CLASIFICADOR DE MAQUINAS DE SOPORTE VECTORIAL PARA LA DETECCIÓN Y CLASIFICACIÓN DEL CÁNCER DE SENO USANDO IMÁGENES MAMOGRAFICAS"*** *con Autorización de su Jefatura. Se solicita brindar las facilidades del caso.*

Es propicia la oportunidad para expresarle mis consideraciones más distinguidas.

Atentamente,

Dr. Juan Marcia VELASCO CARDENAS
CMP. 25266 RNE. 14541
DIRECTOR
HOSPITAL REGIONAL "MNB" - PUNO

JMVC/JMBA
CC.Arch. Direcc.
CC.Arch. UADI

HOSPITAL REGIONAL "M.N.B."- PUNO
DPTO. DIAGNOSTICO POR IMAGENES
SECRETARIA - RECEPCION
18 ENE 2022
HORA:_____ FIRMA:_____
REG. N°_____ FOLIO:_____