UNIB.E

UNIVERSIDAD IBEROAMERICANA DEL ECUADOR

FACULTAD DE LA INGENIERÍA DE SOFTWARE

CARRERA: INGENIERÍA DE SOFTWARE

Diseño de un Modelo basado en Técnicas de Machine Learning para la Clasificación de Imágenes Médicas del Cáncer Pulmonar: Contribuciones al Diagnóstico Médico

Trabajo de Integración Curricular para la obtención del Título de Ingeniería de Software

Autora:

Odalis Fernanda Rea Chamorro

Tutor:

Msc. Tonysé de la Rosa

Quito, Ecuador

Septiembre, 2024

DECLARACIÓN DE AUTORÍA Y AUTORIZACIÓN PARA LA DIFUSIÓN DEL TRABAJO DE INTEGRACIÓN CURRICULAR

1. Yo, Odalis Fernanda Rea Chamorro**,** declaro en forma libre y voluntaria, que los criterios emitidos en el presente Trabajo de Integración Curricular, titulado: “Diseño de un Modelo basado en Técnicas de Machine Learning para la Clasificación de Imágenes Médicas del Cáncer Pulmonar: Contribuciones al Diagnóstico Médico”, previo a la obtención del título profesional de Ingeniería de Software, así como también los contenidos, ideas, análisis, conclusiones y propuestas son exclusiva responsabilidad de mi persona, como autor/a.

2. Declaro, igualmente, tener pleno conocimiento de la obligación que tiene la Universidad Iberoamericana del Ecuador, de conformidad con el artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT, en formato digital una copia del referido Trabajo de Integración Curricular para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública, respetando los derechos de autor.

3. Autorizo, finalmente, a la Universidad Iberoamericana del Ecuador a difundir a través del sitio web de la Biblioteca de la UNIB.E (Repositorio Digital Institucional), el referido Trabajo de Integración Curricular, respetando las políticas de propiedad intelectual de la Universidad Iberoamericana del Ecuador.

Quito, DM., a los 05 días del mes de septiembre de 2024.

Odalis Fernanda Rea Chamorro

1752442267

AUTORIZACIÓN DE PRESENTACIÓN FINAL DEL TRABAJO DE INTEGRACIÓN CURRICULAR POR PARTE DEL TUTOR

Msc. Sandino Jaramillo
Director de la Carrera Ingeniería de Software
Presente. -

Yo, Msc. Tonysé de la Rosa, Tutor del Trabajo de Integración Curricular realizado por el estudiante Odalis Fernanda Rea Chamorro de la carrera de Ingeniería de Software informo haber revisado el presente documento titulado *"Diseño de un Modelo basado en Técnicas de Machine Learning para la Clasificación de Imágenes Médicas del Cáncer Pulmonar: Contribuciones al Diagnóstico Médico"*, el mismo que se encuentra elaborado conforme a lo establecido en el Reglamento de Titulación y el Manual de Estilo de la Universidad Iberoamericana del Ecuador, UNIB.E de Quito, por lo tanto, autorizo la entrega del Trabajo de Integración Curricular a la Unidad de Titulación para la presentación final ante el tribunal evaluador.

Atentamente,
Msc. Tonysé de la Rosa
Tutor

UNIB.E

UNIVERSIDAD IBEROAMERICANA DEL ECUADOR

ACTA DE APROBACIÓN DEL TRABAJO DE INTEGRACIÓN CURRICULAR

Facultad: Comunicación y Tecnologías

Carrera: Ingeniería de Software

Modalidad: Semipresencial

Nivel: 3er nivel de Grado

En el Distrito Metropolitano de Quito a los dieciocho días del mes de septiembre del 2024 (18-09-2024) a las once horas con cero minutos (11:00), ante el Tribunal de Presentación Oral, se presentó la señorita: **REA CHAMORRO ODALIS FERNANDA**, titular de la cédula de ciudadanía No. **1752442267** a rendir la evaluación oral del Trabajo de Integración Curricular: "**Diseño de un Modelo basado en Técnicas de Machine Learning para la Clasificación de Imágenes Médicas del Cáncer Pulmonar: Contribuciones al Diagnóstico Médico**", previo a la obtención del Título de Ingeniera de Software. Luego de la exposición, la referida estudiante obtiene las calificaciones que a continuación se detallan:

	Calificación
Lectura del Trabajo de Integración Curricular	**8.4/10**
Evaluación Oral del Trabajo de Integración Curricular	7 /10
Calificación Final del Trabajo de Integración Curricular	7.7/10

Para constancia de lo actuado, los miembros del Tribunal de Presentación Oral del Trabajo de Integración Curricular, firman el presente documento en unidad de acto, a los dieciocho días del mes de septiembre del 2024 (18-09-2024).

PhD. Luisa Taborda
DIRECTOR ACADEMICO

Mgst. Sandino Jaramillo
DIRECTOR DE CARRERA

Mgst. Tanysé de la Rosa
TUTOR

UNIB.E
SECRETARÍA ACADÉMICA
GRADO

Mgst. Byron Moreno
LECTOR

DEDICATORIA

Dedico al que es mi fuerza cuando estoy débil, al que cuando estoy sola es mi apoyo, al que cuando estoy perdida es mi sendero por ello este trabajo es para Dios. De igual forma, dedico este trabajo de titulación a mis padres que siempre estuvieron conmigo fomentándome la humildad, responsabilidad y disciplina, admirando su fortaleza. A mi hermano quien es un pilar en mi vida.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Odalis Fernanda Rea Chamorro. ***"Diseño de un Modelo basado en Técnicas de Machine Learning para la Clasificación de Imágenes Médicas del Cáncer Pulmonar: Contribuciones al Diagnóstico Médico".*** Carrera Ingeniería de Software. Universidad Iberoamericana del Ecuador. Quito Ecuador. 2024. 2-63 pp.

RESUMEN

La investigación se enfocó en diseñar un modelo de Machine Learning (ML) capaz de clasificar imágenes médicas del cáncer pulmonar, con la finalidad de apoyar el diagnóstico médico. Se ha utilizado una metodología deductiva, revisión documental, un enfoque cuantitativo, con un diseño no experimental, un nivel descriptivo y se basa en un paradigma positivista. El instrumento en el que se analiza es a través de la estadística descriptiva. El dataset está compuesto por tres categorías: cáncer de pulmón benigno y maligno, pulmones normales, el total de imágenes es de 1097. Se analizó el primer objetivo específico mediante una matriz de análisis documental, para escoger el modelo a través de la validación de los 5 expertos. Los resultados mostraron un tiempo de entrenamiento de 12.88 segundos, la división de los datos fue el 20% de prueba equivale a 220 imágenes y 80% entrenamiento corresponde a 877 imágenes. De las 220 imágenes de prueba se detectó el cáncer en 136 imágenes y sin cáncer en 84. En conclusión, el modelo Máquina de Vectores de Soporte (SVM) ha demostrado ser muy eficaz en la clasificación de imágenes médicas del cáncer pulmonar, superando problemas relacionados con la calidad y cantidad de datos. Se recomienda llevar a cabo una validación exhaustiva con datos adicionales y elaborar un informe detallado en 1 a 3 meses para asegurar la robustez del modelo antes de su implementación.

Palabras Clave: Machine Learning, entrenamiento, prueba, SVM y metodología.

ABSTRACT

The research focused on designing a Machine Learning (ML) model capable of classifying medical images of lung cancer, with the purpose of supporting medical diagnosis. A deductive methodology, documentary review, a quantitative approach has been used, with a non-experimental design, a descriptive level and is based on a positivist paradigm. The instrument in which it is analyzed is through descriptive statistics. The dataset is composed of three categories: benign and malignant lung cancer, normal lungs, the total images are 1097. The first specific objective was analyzed using a documentary analysis matrix, to choose the model through the validation of the 5 experts. The results showed a training time of 12.88 seconds, the division of the data was 20% testing equals 220 images and 80% training corresponds to 877 images. Of the 220 test images, cancer was detected in 136 images and no cancer in 84. In conclusion, the Support Vector Machine (SVM) model has proven to be very effective in classifying medical images of lung cancer, overcoming related problems. with the quality and quantity of data. It is recommended to perform a thorough validation with additional data and prepare a detailed report in 1 to 3 months to ensure the robustness of the model before implementation.

Keywords: Machine Learning, training, testing, SVM and methodology.

INTRODUCCIÓN

Este proyecto de investigación se compone de cinco capítulos. El primer capítulo, "El Problema," examina las principales variables del estudio y el desafío que enfrentan los profesionales de la salud para detectar enfermedades pulmonares en etapas tempranas debido a la falta de tecnología avanzada en IA y ML, lo que conduce a diagnósticos tardíos. El objetivo es diseñar un modelo de Machine Learning (ML) capaz de clasificar imágenes médicas del cáncer pulmonar, con la finalidad de apoyar el diagnóstico médico. Los objetivos incluyen definir el modelo ML, analizar el dataset para asegurar la calidad de los datos, desarrollar una abstracción matemática, evaluar el desempeño y desarrollar el modelo.

Se formulan las preguntas de investigación y se justifica la elección del conjunto de datos, que incluye casos benignos, malignos y en otros casos normales, argumentando que no es necesario validar el modelo con profesionales de la salud para su integración clínica. El alcance del estudio abarca la implementación de herramientas y frameworks como Python, Google Colab, TensorFlow, Keras y OpenCV para el desarrollo del modelo. El segundo capítulo, “Marco Teórico,” presenta los antecedentes y fundamentos teóricos necesarios, incluyendo la definición y clasificación de las variables de estudio. Se destacan métricas como exactitud, precisión, sensibilidad y F1-Score para evaluar el rendimiento del modelo, junto con técnicas como validación cruzada, preprocesamiento, aumentación de datos y optimización con SGD para mejorar y validar su eficacia.

En el tercer capítulo detalla la metodología de investigación con un enfoque deductivo y cuantitativo, revisión documental, utilizando un diseño no experimental y un paradigma positivista, con análisis mediante estadística descriptiva y la metodología Kanban. La investigación descriptiva evalúa modelos de ML en el dataset "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD", incluyendo operacionalización de variables y validación por expertos. En el cuarto capítulo de la tesis, se expone detalladamente el análisis e interpretación de los resultados del modelo de ML. En el quinto capítulo, se presentan las conclusiones y recomendaciones derivadas del presente trabajo de investigación. En los anexos están los 5 juicios de expertos, el proceso del modelo a través de la metodología Kanban y el modelo se encuentra en el repositorio de GitHub.

CAPÍTULO I

EL PROBLEMA

El problema de investigación se refiere según Arias (2012) “Independientemente de su naturaleza, un problema es todo aquello que amerita ser resuelto. Si no hay necesidad de encontrar una solución, entonces no existe tal problema” (pág. 37). El problema de la investigación es que a los profesionales de la salud se les dificulta detectar el cáncer de pulmón, debido a la ausencia de alta tecnología de IA. El 70% de los diagnósticos son demasiado tardíos por falta de infraestructura médica, falta de adopción de herramientas de ML y otros avances tecnológicos, menor precisión temprana y precisión en el diagnóstico (Cortes, 2019).

Planteamiento del Problema

La IA es un campo de la ciencia y la ingeniería enfocado en desarrollar máquinas capaces de realizar tareas que normalmente requieren inteligencia humana, mejorando la capacidad de las máquinas para aprender y adaptarse de manera autónoma (Russell & Norvig, 2016). Por su parte Rodríguez, Flores y Vitón (2022) refieren que “ML es el estudio de herramientas y métodos para identificar patrones en los datos. Estos patrones pueden usarse luego para aumentar la comprensión del mundo actual o hacer predicciones sobre el futuro” (pág. 2). Con lo cual, se busca diseñar un modelo de ML para clasificar imágenes médicas de cáncer pulmonar en tipos malignos, benignos y también imágenes normales para el entrenamiento del modelo, con el propósito de mejorar la capacidad diagnóstica en el ámbito médico. Este proyecto surge como respuesta a la necesidad de elevar la precisión en la detección de esta enfermedad, aprovechando los avances en la tecnología de ML y al modelo Support Vector Machine (SVM), se espera que los resultados obtenidos conduzcan a una mejora en la precisión de la clasificación de imágenes médicas de cáncer pulmonar, para disminuir en los errores de diagnóstico y optimizando el proceso de detección.

A nivel mundial, en España se desarrolló el proyecto Anorak, el objetivo fue crear un modelo de IA para analizar imágenes a nivel de píxeles y ayudar a los patólogos en la realización de pronósticos tumorales más precisos, así como en la predicción de la reproducibilidad y el riesgo del adenocarcinoma de pulmón a nivel mundial. Este sistema ha sido aplicado a más de 5500 portaobjetos de diagnóstico correspondientes a 1372 casos de adenocarcinoma de pulmón de cuatro cohortes independientes de pacientes. Anorak surgió de la necesidad de mejorar el diagnóstico del cáncer de pulmón, aprovechar los avances en la tecnología de IA y superar las limitaciones de los métodos tradicionales. Sus resultados incluyen mejorar el pronóstico, reducir los errores de diagnóstico y optimizar los recursos médicos (Roche, 2024).

A nivel continental, en Estados Unidos se desarrolló Sybil, un modelo de aprendizaje profundo o Deep Learning (DL) diseñado para analizar exploraciones y predecir el riesgo de padecer enfermedades pulmonares. Las motivaciones para su creación incluyen la alta prevalencia de estas enfermedades, las limitaciones de los métodos de detección actuales y los avances en tecnologías de IA y DL. Las consecuencias de Sybil son significativas: mejora la detección temprana, reduce los errores diagnósticos y optimiza los recursos sanitarios. El equipo validó Sybil con tres conjuntos de datos independientes, incluido uno del National Lung Screening Trial (NLST) con exploraciones de más de 6.000 participantes, de los cuales el 92% eran estadounidenses blancos (Ecancer, 2023).

En Ecuador, no existe una herramienta tecnológica que permita identificar y detectar, de forma temprana y eficiente, las cuatro enfermedades: pulmonares, diabetes, cardiovasculares, cerebrovasculares en personas independientemente de sus edades. Se ha realizado un análisis exhaustivo de diferentes modelos de ML con el objetivo de predecir el riesgo de contraer las cuatro enfermedades. Para este análisis, se dividió los datos en un 25% para prueba y un 75% para entrenamiento. Además, mostró un desempeño sólido en la predicción del cáncer de pulmón. Las razones detrás de este desarrollo incluyen la alta incidencia de enfermedades crónicas, los avances tecnológicos y la disponibilidad de datos, así como la necesidad de soluciones locales y personalizadas. Este avance tiene varias consecuencias positivas, como la mejora en la detección y el pronóstico de enfermedades, el fortalecimiento de las capacidades técnicas locales. (Valdés, Intriago & Felipe, 2022).

El problema de investigación es que el 70% de los diagnósticos son demasiado tardíos. La detección temprana del cáncer de pulmón es difícil debido a la falta de recursos y tecnologías avanzadas, así como al desconocimiento y resistencia al uso del ML, resultando en diagnósticos tardíos y menos precisos. Esto eleva los costos de tratamiento y tiene un impacto emocional significativo en pacientes y familias. Además, la falta de infraestructura y tecnología avanzada aumenta la inequidad en la atención médica. La integración del ML en el proceso podría contribuir a mejorar el reconocimiento oportuno y, por ende, se pude mejorar las tasas de supervivencia y la calidad de vida (Cortes, 2019).

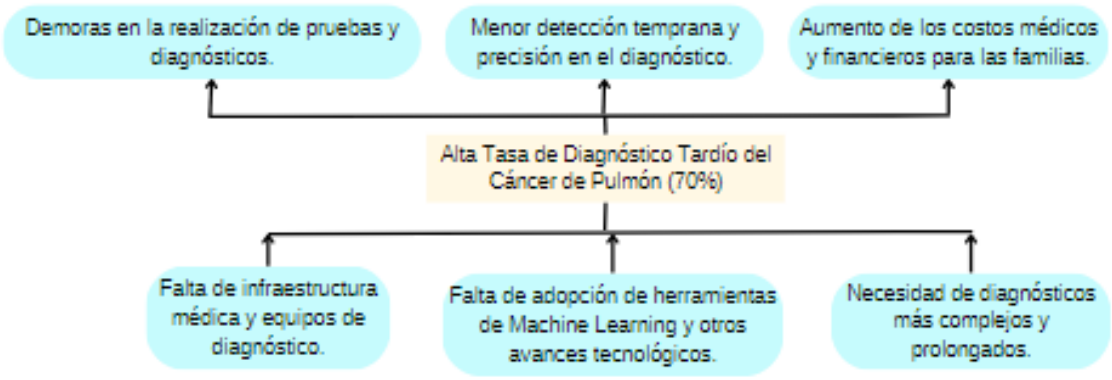

Figura 1. *Árbol del problema de la investigación*

Se propone una solución al problema identificado mediante el desarrollo de un modelo de algoritmos de ML diseñado para categorizar imágenes médicas relacionadas con el cáncer de pulmón e imágenes de pulmones sanos, mejorando la precisión y rapidez del diagnóstico, en el año 2024. Este modelo busca mejorar el proceso al permitir una identificación más rápida y precisa de los casos. Con esta solución, se busca abordar diversos desafíos significativos asociados con la clasificación de esta enfermedad, utilizando el dataset "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD". Se propone desarrollar un modelo para mejorar la precisión y rapidez del diagnóstico. Ante esto, surge las preguntas: ¿Cuál modelo de ML es el más adecuado para la clasificación de imágenes radiológicas del cáncer pulmonar, categorizándolas como benignas, malignas, entre otros casos normales o sanas, utilizando el dataset "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD"? ¿Cómo se calcula la

abstracción matemática para el desarrollo del modelo ML y la evaluación del desempeño?

Objetivos de la Investigación

Objetivo general

Diseñar un modelo de Machine Learning (ML) capaz de clasificar imágenes médicas del cáncer pulmonar, con la finalidad de apoyar el diagnóstico médico.

Objetivos específicos

- Definir el modelo de Machine Learning para la clasificación de imágenes radiológicas del cáncer pulmonar.
- Analizar el dataset de imágenes clínicas del cáncer de pulmón, para el aseguramiento de la calidad de los datos previo al entrenamiento y a la validación del modelo.
- Desarrollar una abstracción matemática para el modelo Machine Learning y la evaluación del desempeño.
- Desarrollar el modelo de Machine Learning para la clasificación de imágenes médicas del cáncer pulmonar.

Justificación e Impacto de la Investigación

El desarrollo de la propuesta nos permitirá explorar la viabilidad de incorporar un extenso conjunto de imágenes médicas de neoplasias malignas pulmonares de pacientes de todas las edades, basándonos en la investigación de ML. Para entender los algoritmos se debe tomar en cuenta que “Los algoritmos de aprendizaje automático pueden analizar grandes cantidades de datos médicos, incluyendo imágenes de tomografía, para identificar patrones y características relacionadas con el cáncer de pulmón” (Salvat, 2023, pág. 9). Lo que demuestra que la implementación de este modelo podría incrementar la precisión y disminuir el tiempo requerido para clasificar la condición médica. El conjunto de datos incluye tres directorios para casos benignos, malignos y normales, por lo que se considera que no es necesario validar

el modelo con profesionales de la salud para asegurar su integración en los flujos de trabajo clínicos.

El propósito del modelo de ML es que clasifica imágenes pulmonares con cáncer en benignas, malignas y en otros casos normales o sanas con alta interpretabilidad, mejorando la detección temprana y la precisión del diagnóstico. Esta iniciativa facilita el acceso a tecnologías avanzadas de diagnóstico y optimiza los recursos médicos. La integración del aprendizaje automático contribuye a una identificación más efectiva de la enfermedad, mejorando las tasas de supervivencia y la calidad de vida de los pacientes. Además, alivia la carga emocional y financiera para los pacientes y sus familias, proporcionando un apoyo significativo al diagnóstico médico.

Este proyecto de investigación tiene un doble propósito: por un lado, fortalecer mi perfil profesional al explorar y demostrar las posibilidades y resultados de aplicar los modelos de ML y las arquitecturas de redes neuronales más avanzadas; por otro lado, proporcionar a los estudiantes de la Universidad Iberoamericana una herramienta de estudio que sirva como base para el desarrollo de proyectos más complejos. El modelo SVM aprende de forma autónoma a interpretar imágenes pulmonares para facilitar el diagnóstico clínico. Este modelo se eligió porque tiene mejor tiempo, alta interpretabilidad, permite calcular las métricas de evaluación y entrenamiento como la precisión, sensibilidad, especificidad, puntaje F1, AUC-ROC y es adecuado al dataset “El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD”.

Alcance de la investigación

La investigación se centra en definir y desarrollar un modelo de ML para la clasificación de imágenes radiológicas de carcinoma pulmonar. Se basará en la interpretación de radiografías pulmonares en tres categorías: maligno, benigno y normal, utilizando el dataset "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD". Este conjunto de datos es de alcance internacional, la mayoría de ellos provienen de lugares de la región central de Irak y contiene un total de 1.097 imágenes, que representan casos con diversidad de género, edad, nivel educativo, área de residencia y estado de vida. Las imágenes corresponden a tomografías computarizadas de pacientes diagnosticados con neoplasias pulmonares malignas en

diversas etapas, así como de individuos sanos (Solano, 2021). Se va a desarrollar una abstracción matemática del modelo de ML y evaluará el desempeño, utilizando Python, Google Colab, TensorFlow, Keras y OpenCV como herramientas y frameworks para el desarrollo del modelo.

CAPÍTULO II

MARCO TEÓRICO

Para comprender que es el marco metodológico se debe tener en cuenta que según Azuero (2019) “La formulación del marco metodológico en una investigación, es permitir, descubrir los supuestos del estudio para reconstruir datos, a partir de conceptos teóricos habitualmente operacionalizados. Significa detallar cada aspecto seleccionado para desarrollar dentro del proyecto de investigación (…)” (pág. 1). Esto sugiere que, en el proceso de construcción del marco teórico, luego de caracterizar el problema y establecer los objetivos, se lleva a cabo una cuidadosa selección de investigaciones previas que constituyen las bases teóricas. Estas investigaciones anteriores proporcionan antecedentes teórico-empíricos, como la interpretación de imágenes pulmonares, a través de la revisión de libros, fuentes y textos que permiten enfocar el estudio.

Antecedentes de la investigación

El estudio en el ámbito internacional, Salvat (2023), en España, desarrollaron la tesis titulada “Aplicaciones del machine learning en el diagnóstico del cáncer de pulmón,” cuyo objetivo fue desarrollar de una red neuronal convolucional (CNN) que a partir de imágenes histopatológicas de tumores pulmonares puede clasificar tumores benignos, carcinomas o adenocarcinomas. Se ha escogido la metodología Agile para diseñar varias versiones de un mismo modelo y probarlos. Se pudo observar que la CNN desarrollada produce predicciones prácticamente perfectas, fallando únicamente al clasificar ciertos adenocarcinomas como carcinomas. Se ha usado el ordenador personal, pero todas las funciones son ejecutadas en Google Colab, de manera que pueda contar con un rendimiento mayor al que podría ofrecer el ordenador por sí solo. En conclusión, el objetivo principal de la red neuronal se cumple satisfactoriamente, ya que todo paciente es clasificado correctamente dependiendo de si su tumor es benigno o cancerígeno.

Este estudio destaca que la metodología ágil es altamente flexible y adaptable al proceso de desarrollo de modelos de Machine Learning (ML) para la clasificación de imágenes. Esto se debe a que permite ajustes continuos basados en la retroalimentación y los cambios en los requisitos del proyecto. Por lo tanto, posibilita la creación de varias versiones del mismo modelo y su posterior prueba. Esto implica que con esta metodología se pueden trabajar en iteraciones cortas y entregas incrementales del modelo de ML. Como resultado, todas las imágenes médicas de los pacientes pueden ser clasificadas adecuadamente según si el tumor es benigno o maligno.

Otro estudio ámbito internacional, Rivas (2023), en Perú, desarrollaron la tesis titulada “Clasificación de cáncer de pulmón en imágenes de tomografías mediante procesamiento de imágenes y aprendizaje automático,” la cual tiene como objetivo de clasificar el cáncer de pulmón en imágenes de tomografías mediante procesamiento de imágenes digitales y aprendizaje automático. La metodología es cuantitativa, este trabajo de investigación se realizó mediante la automatización del software Parsifal costando de 4 partes revisión, planificación, conducción y reportes. En los resultados de análisis se obtuvo métricas porcentuales de 97% o más. Sin embargo, las pruebas revelaron que las métricas de exactitud y precisión alcanzaron porcentajes de 95% y 91%, respectivamente. En conclusión, el entorno de trabajo y lenguaje de programación seleccionados son Google Colab y Python respectivamente por la factibilidad y eficiencia con la que se puede desarrollar todo el proceso de implementación, pruebas y despliegue del sistema.

En este estudio aporta que los algoritmos de ML analizaron grandes cantidades de datos médicos, incluyendo imágenes de tomografía, para identificar patrones y características del cáncer de pulmón, mediante procesamiento de imágenes digitales. Al entrenar la red neuronal, se puede usar el modelo entrenado con el conjunto de datos de Testing, para mejorar el desempeño y la escalabilidad del modelo ante nuevas entradas de datos. La utilización de la metodología cuantitativa afronta los problemas de carácter tecnológico, el producto es tecnológico, se puede utilizar en las investigaciones con conocimiento teórico científico, a su vez se emplearon conocimientos teóricos probados al problema planteado.

Además, el estudio en el ámbito internacional, Leivi (2019), en Argentina, desarrollaron la tesis titulada “Análisis de la implementación de Machine Learning en el diagnóstico por imágenes,” con el objetivo de determinar beneficios y barreras que expliquen el estado actual de Machine Learning (ML) aplicado al Diagnóstico por Imágenes (DPI). La metodología es de tipo exploratoria, para el análisis el uso de ML para el DPI, que, si bien, viene siendo estudiado desde una perspectiva médica-científica para casos puntuales. Los resultados de las investigaciones han sido satisfactorios, en una gran medida, no se ha llegado a una etapa avanzada en el ciclo de adopción de los productos tal que garantice éxitos y permanencias de las empresas. En conclusión, se considera importante tener en cuenta que, dado el auge reciente de la tecnología de ML en DPI, la industria no está exenta de transitar caminos sinuosos o fallidos.

Este estudio destaca el uso de ML en el diagnóstico por imágenes ha generado resultados satisfactorios, pero todavía existen barreras que dificultan la adopción que garantice éxitos en la aceptación por parte de los profesionales médicos y las empresas. La industria no está exenta de transitar caminos sinuosos o fallidos. Además, es fundamental abordar las limitaciones y desafíos señalados en este estudio para avanzar hacia una aplicación más amplia y efectiva del ML en el diagnóstico por imágenes, lo que podría conducir a mejoras significativas en la atención médica y en los resultados para los pacientes.

El estudio en el ámbito nacional, Lopez & Terranova (2023), en Guayaquil, desarrollaron la tesis titulada “Técnicas de Machine Learning basadas en Aprendizaje Supervisado para la predicción de enfermedades respiratorias y/o pulmonares ocasionadas y derivadas por el Covid19,” con el objetivo de realizar un análisis de imágenes radiográficas mediante la implementación con Python de diferentes modelos de redes neuronales artificiales para identificar distintos tipos de enfermedades respiratorias y/o pulmonares producidas por el covid-19. Se empleó el método cuasi experimental, se utilizan métodos similares a los de un experimento. Se siguió un proceso para seleccionar el dataset para entrenar y validar los modelos de redes neuronales convolucionales, los cuales fueron desarrollados usando Python, Tensor Flow. La evaluación de los modelos demostró valores superiores al 85% en

todas las métricas utilizadas. La investigación concluye que los resultados de ambos modelos fueron muy satisfactorios, ya que presentaron un buen rendimiento y se obtuvo métricas de clasificación que son consideradas satisfactorias para las predicciones y detecciones de los tipos de enfermedades pulmonares descritos anteriormente.

Este estudio ilustra que se llevó a cabo la evaluación de varios enfoques de ML para predecir el Covid-19, obteniendo resultados altamente precisos, lo cual fue satisfactorio. Sin embargo, es fundamental destacar que se utilizó lenguaje de alto nivel de programación interpretado de Python y el framework de TensorFlow, se obtuvo métricas de clasificación que son consideradas satisfactorias para las predicciones y detecciones de los tipos de enfermedades pulmonares, los resultados fueron satisfactorios. El proceso que se realizó fue desde la selección del dataset, luego el entrenamiento y la validación de los modelos.

Otro estudio en el ámbito nacional, Valdés, Intriago & Felipe (2022), en Manabí, desarrollaron la tesis titulada “Predicción de las principales enfermedades que afectan la salud en Ecuador a partir de factores de riesgo,” con el objetivo de hacer una predicción de las principales enfermedades que afectan la salud en Ecuador, a partir de factores de riesgo, que sirva de insumo médico para el personal de salud, con el uso de modelos de ML, para así, dar soporte a la toma de decisiones médicas. Se implementaron las metodologías Crisp-DM (Cross – Industry Standard Process for Data Mining) y el proceso KDD (Knowledge Discovery Databases). Los resultados son que el modelo Random Forest posee mejor rendimiento para las enfermedades cerebrovasculares; el modelo de Regresión Logística tiene un mayor desempeño en la diabetes mellitus. El índice de Shapley fue usado para explicar los factores de riesgo que más influyeron en la aparición de las mismas. La investigación concluye que el índice de Shapley fue utilizado para validar estos resultados.

Este estudio resalta que la aplicación de un modelo de ML facilita el apoyo en la toma de decisiones médicas, ofreciendo predicciones precisas y rápidas sobre enfermedades. Esto capacita a los profesionales de la salud para anticipar posibles afecciones. Los métodos de ML analizan extensas bases de datos médicas para detectar patrones y relaciones. Además, el procesamiento de la información en estos

modelos es rápido. Estas herramientas de ML son recursos valiosos para la toma de decisiones clínicas y contribuyen a mejorar la calidad de la atención médica, optimizando los recursos y proporcionando un mejor cuidado a los pacientes.

Además, el estudio en el ámbito nacional, Tuarez & Vera (2022), en la Maná, desarrollaron la tesis titulada “Desarrollo de software biomédico mediante modelos Deep Learning para la detección de tumores pulmonares en la aplicación de procesamiento de imágenes espectrales para el departamento médico,” con el objetivo de desarrollar un software biomédico basado en modelos DL para la detección de tumores pulmonares en la aplicación de procesamiento de imágenes espectrales. Se aplica la metodología documentada, permite en el desarrollo de la investigación considerar bases teóricas de investigaciones donde se está analizando la problemática del proyecto, aplicado a la detección de tumores pulmonares mediante radiografías DICOM (Estándar de transmisión de imágenes). Se aplicó 9 casos lo cual hubo una probabilidad de 8 aciertos que representan una tasa de error del 5%, por lo tanto, existe un 88% de acierto que es tumor pulmonar tomando también el fallo predictivo de un 8.55%. En conclusión, permitió obtener resultados precisos en el tratamiento de imágenes en base al proceso de preparación e implementación dentro del software biomédico que está orientado a la detección de tumores pulmonares.

Este estudio resalta la eficacia en el desarrollo de software biomédico para la detección de tumores pulmonares mediante el procesamiento de imágenes espectrales. Los resultados obtenidos mostraron una precisión notable en la identificación de dichos tumores. Estos hallazgos sugieren que la integración de esta tecnología en el ámbito médico podría tener un impacto significativo en el diagnóstico precoz y preciso de enfermedades, lo que podría mejorar tanto la calidad de la atención médica como los resultados para los pacientes. Además, el éxito en el desarrollo de este software biomédico resalta un potencial para abordar desafíos médicos complejos y generar nuevas oportunidades en la investigación y el tratamiento de enfermedades.

Bases Teóricas

Machine Learning (ML)

El objetivo del ML

A continuación, explican Vargas et al. (2022) que “El aprendizaje automático tiene como objetivo identificar patrones a partir de los datos con el fin de hacer predicciones, detecciones o clasificaciones” (p. 2). Se refiere a que el ML se centra en identificar patrones en los datos, lo que implica descubrir relaciones, estructuras o regularidades que pueden ser utilizadas para hacer predicciones, clasificaciones o tomar decisiones. Este enfoque permite que los algoritmos se entrenen para reconocer y aprovechar estos patrones, facilitando así diversas aplicaciones prácticas como la detección y clasificación de enfermedades, mejorando la eficiencia y precisión de los procesos de análisis de datos.

Proceso del ML

Plantean Vargas et al. (2022) que “Llegar a una solución analítica es un proceso que parte desde comprender el problema, estructurar y recopilar los datos para posteriormente controlar, monitorear y utilizar los resultados de la solución con el fin de calibrar el modelo” (p. 2). En otras palabras, proponen un enfoque metódico y cíclico donde la recopilación y análisis de datos son esenciales, no solo para obtener una solución, sino también para ajustar y perfeccionar el modelo continuamente basado en los resultados obtenidos. Para llegar a una solución analítica, el cual comienza con la comprensión del problema.

Algoritmos de ML

Algoritmos del ML e imágenes de tomografía

En esta sección, de acuerdo con Salvat (2023) “Los algoritmos de aprendizaje automático pueden analizar grandes cantidades de datos médicos, incluyendo imágenes de tomografía, para identificar patrones y características relacionadas con el cáncer de pulmón” (p. 9). Se refiere que los algoritmos de ML tienen la capacidad

de procesar grandes cantidades de información clínica, incluyendo registros de pacientes, resultados de pruebas médicas e imágenes de tomografía, de manera eficiente y efectiva. Esto permite identificar patrones y características relacionados con el cáncer de pulmón, facilitando su detección temprana. La implementación del ML en el análisis de datos médicos mejora significativamente la precisión y rapidez del diagnóstico, lo que puede contribuir a mejores resultados en el tratamiento y manejo de la enfermedad.

El rendimiento del algoritmo de ML

Es comprensible que Altuna (2023) advierta que “Gran cantidad de algoritmos de aprendizaje automático tienen un rendimiento mejor cuando las variables se encuentran en una escala parecida y cercana a una distribución normal” (p. 57). En otras palabras, la estandarización implica ajustar los valores de tal manera que la desviación estándar de la distribución sea igual a 1, lo que equivale a normalizar el rango de los valores de las características. Para llevar a cabo esta normalización, existen varios métodos disponibles. Uno de los métodos utilizados es la función StandardScaler de Scikit-Learn, la cual resta la media de todos los valores de una característica.

El algoritmo K-nearest neighbours o K-vecinos más cercanos

De acuerdo con Salvat (2023) “Primero, se crea un objeto KneighborsClassifier sin especificar ningún hiperparámetro, y seguidamente se calcula la 10-fold cross validation, con la que obtendremos un valor de 95.79%, con una varianza de 0.0016 (…)” (p. 47). El algoritmo K-nearest neighbours o K- vecinos más cercanos se emplea para predecir la propensión de un paciente a padecer cáncer de pulmón o no. Esta perspectiva se respalda en la afirmación de Abelaira, Ruano & Fernández (2020) que menciona: “Para que un algoritmo reciba crédito y aceptación, debe explicar cómo ha llegado a los resultados generalmente, utilizando un mapa de prominencia que destaca las áreas específicas de la imagen que contribuyeron a la salida final del algoritmo” (p. 3). Los resultados hacen referencia al problema de las "capas ocultas"

presentes en ciertas áreas de las imágenes del cáncer, de acuerdo con la salida del algoritmo.

Estructura básica de un modelo de ML

Plantean Vargas et al. (2022) que “La estructura de un sistema de ML puede simplificarse en cuatro componentes fundamentales” (p. 3). Denota que estos incluyen la información de entrada, que comprende los datos suministrados al sistema para resolver la tarea particular; la tarea a realizar o resolver, como la clasificación de datos o imágenes según enfermedades o fenotipos; el resultado producido por la ejecución del sistema, como la categorización de radiografías de tórax en diferentes patologías; y por último, el proceso de entrenamiento, optimización y ajuste, que implica el entrenamiento del sistema y la modificación de los resultados obtenidos en relación con los resultados esperados. Este procedimiento garantiza la precisión del análisis y su adecuación a las exigencias específicas del problema.

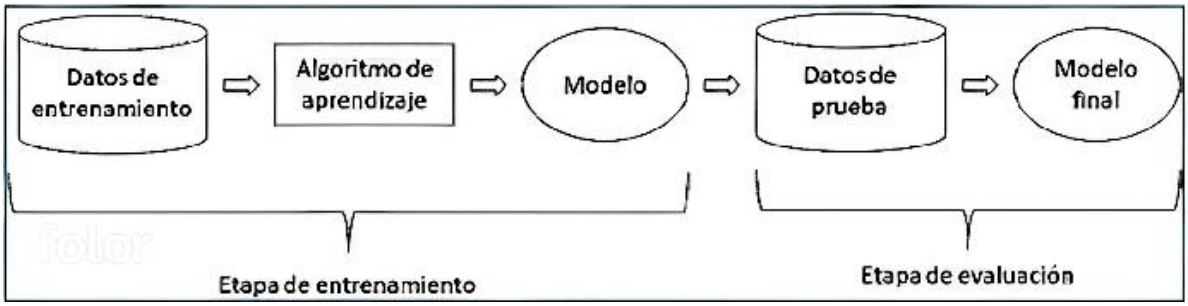

Figura 2. *Proceso de entrenamiento y evaluación de un modelo de ML*

Nota: Tomado de Vargas et al., 2022

Las herramientas más importantes del ML

Según Sarmiento (2020), “Dentro de las herramientas más importantes del ML se encuentran las redes neuronales artificiales, conocidas en inglés como artificial neural networks (ANNs), y el aprendizaje profundo, conocido en inglés como deep learning”

(p. 2). Se destaca que las redes neuronales artificiales (ANNs) y el aprendizaje profundo (deep learning) son herramientas esenciales en el aprendizaje automático. Las ANNs, inspiradas en las redes neuronales biológicas del cerebro humano, tienen la capacidad de aprender y generalizar, lo que les permite reconocer patrones, predecir comportamientos y tomar decisiones de manera efectiva.

Métodos o Modelos de ML

Los cuatro métodos de aprendizaje distintos

Por su parte, Fuentes, Cabrera y Rodríguez (2022) “Se utilizaron cuatro métodos de aprendizaje distintos: máquinas de vectores de soporte, vecinos más cercanos, redes neuronales y árboles de decisión. Se utilizó la base de datos HAM10000, la cual contiene imágenes médicas manualmente etiquetadas, lo que asegura que el proceso de entrenamiento sea efectivo” (p. 1). En este procedimiento se aplicaron los cuatro métodos de aprendizaje previamente mencionados, que incluyen clasificadores de conjuntos basados en kNN y SVM, además de un modelo basado en redes neuronales convolucionales (CNN). También se implementaron algoritmos de búsqueda y de optimización mediante enjambre de partículas.

Análisis del comportamiento de las predicciones en los modelos

Presentan Fuentes et al. (2022) “En este procedimiento se realiza también una implementación con Python para poder analizar el comportamiento de las predicciones en cada uno de los modelos, como K-Nearest-Neighbor (kNN), Support vector machine (SVM), Decisión Tree(Tree), Neural Network” (p. 3). El uso de Python es especialmente apropiado para este proceso, ya que permite a los modelos de aprendizaje automático analizar y evaluar el comportamiento del sistema en el procesamiento e interpretación de imágenes, mejorando así su precisión y capacidad para realizar predicciones acertadas.

Modelo SVM

Conforme a Fos (2016) “El mejor rendimiento se ha conseguido con un clasificador SVM” (p. 92). Lo que demuestra que el modelo SVM tiene una gran capacidad de generalización y, por lo tanto, es una opción destacada en numerosos contextos de clasificación para que se obtengan resultados precisos y fiables en diversas aplicaciones. En conclusión, el uso de SVM es altamente recomendable para alcanzar un rendimiento óptimo en tareas de clasificación. En recomendación, el uso de SVM es altamente aconsejable para alcanzar un rendimiento óptimo en tareas de clasificación.

Modelo de vecinos más cercanos (KNN)

De acuerdo a Prieto (2022) “Los métodos de vecinos más cercanos (KNN) suelen tener mejores resultados” (p. 32). Lo que demuestra que el modelo KNN ofrece un rendimiento competitivo y confiable en una variedad de contextos de ML, por lo tanto, es una opción a considerar en numerosas aplicaciones de clasificación y regresión para obtener predicciones precisas y efectivas en conjuntos de datos diversos y complejos. En conclusión, se recomienda considerar el modelo KNN como una herramienta valiosa en el arsenal de técnicas de aprendizaje automático, es recomendable explorar su aplicación en diferentes problemas de análisis de datos.

Random Forest

Según Salvat (2023) “Una vez comprobada la capacidad predictiva del modelo, se procede a entrenar el Random Forest Classifier con sus hiperparámetros predeterminados. Una vez entrenado el modelo, se realiza un estudio de los resultados brindados por este (…)” (p. 53). Se refiere al proceso de evaluación y análisis posterior del modelo Random Forest Classifier. Por lo tanto, este enfoque permite entender la efectividad del modelo en la predicción de datos nuevos y no vistos para mejorar la comprensión de su desempeño y tomar decisiones informadas sobre su aplicación en escenarios prácticos. En conclusión, este proceso de entrenamiento y evaluación es crucial para garantizar la fiabilidad y utilidad del modelo en diferentes contextos de aplicación.

Aprendizaje Supervisado

Regresión lineal

Proponen Vargas et al. (2022) que “Es un método estadístico para predecir una variable cuantitativa Y sobre la base de una única variable predictiva de regresión X, debido a que la relación entre ambas es lineal” (p. 3). Se utiliza una técnica de análisis para estimar el valor de una variable numérica en función de otra variable, asumiendo que existe una relación directa y proporcional entre ellas. Igualmente afirman Vargas et al. (2022) que “Matemáticamente, se escribe $Y \approx \beta_0 + \beta_1 X$, siendo “≈” leído como “se modela aproximadamente como…”. Los coeficientes β_0 y β_1 son desconocidos en la práctica, así que es necesario utilizar información o datos para estimarlos” (p. 3). Hace referencia a que la relación entre las variables se representa mediante una ecuación lineal, donde Y se aproxima a $\beta_0 + \beta_1 X$. Dado que los valores de los coeficientes β_0 y β_1 no se conocen de antemano, es necesario emplear datos para calcular estas estimaciones.

Regresión logística

Sostienen Vargas et al. (2022) que “Es una forma de modelar la probabilidad de que una variable cualitativa Y pertenezca a una categoría binaria. Se utiliza la función logística, expresada como $p(X) =$, e $\beta_0 + \beta_1 X$ 1 + $\beta_0 + \beta_1 X$ (…) (p. 4). Esto significa que esta técnica permite calcular la probabilidad de que una variable categórica tome uno de dos posibles valores, utilizando una fórmula que transforma los valores de X en una probabilidad que siempre se encuentra entre 0 y 1, representada visualmente por una curva sigmoide. Una curva sigmoide es una gráfica que tiene forma de "S" y es característica de la función logística. Esta curva se utiliza para modelar situaciones donde el crecimiento comienza lentamente, se acelera y luego se estabiliza.

El aprendizaje supervisado se requieren instancias etiquetadas

Según Fuentes et al. (2022) “El aprendizaje supervisado se requieren instancias etiquetadas como entrenamiento, que son usadas por el sistema automático para aprender, mientras que en el aprendizaje no supervisado no se requiere de estas instancias” (p. 2). En consecuencia, en el aprendizaje no supervisado no es necesario

tener en cuenta los datos de salida cuando se buscan detectar patrones en los datos de entrada. En contraste, en el aprendizaje supervisado se observan pares de datos de entrada y salida, ampliando así el concepto de resolución de problemas mediante algoritmos de regresión y clasificación.

Aprendizaje estadístico supervisado

Por su parte Bernavent, Colomer, Quecedo, Gol-Monserrat y Llano (2024) explican que “Aprendizaje estadístico supervisado: este tipo de algoritmo se utiliza cuando se conocen datos previos y se sabe el tipo de resultado que se quiere obtener. Se pueden realizar predicciones en base a los datos previamente introducidos en el sistema, mediante un proceso de entrenamiento con la información extraída de la comparación de los resultados obtenidos por el algoritmo con los esperables en el histórico. Así, se optimiza una función de ganancia/función de pérdida sobre una variable dependiente observada (exógena)” (p. 29). Con este enfoque, el Aprendizaje estadístico supervisado facilita la realización de predicciones basadas en datos previamente introducidos en el sistema. Mediante la evaluación de diversos algoritmos, se pueden determinar los resultados de salida (output) que mejor se relacionan con las entradas (input).

Dataset

Definición del Dataset

Según la investigación realizada, se define un Dataset como un conjunto de datos organizado en un sistema de almacenamiento que proporciona los principales criterios de búsqueda o estructura de la información a ser trabajada. Básicamente, representa el contenido de una tabla dentro de una base de datos, con diversas columnas que contienen registros almacenados en cada fila. Estas filas pueden considerarse como las categorías de los datos, mientras que las columnas representan las posibles variables que las conforman. La combinación de columnas y filas constituye lo que se conoce como un Dataset. Este conjunto de datos puede ser utilizado para diversos propósitos, dependiendo de la metodología, orientación o tratamiento que se desee aplicar a la información. Su objetivo es facilitar la vida de

las personas, automatizar tareas o simplemente analizar la información de manera más eficiente (Solis, 2023).

Descripción del Dataset

En esta investigación conceptual, se define un dataset como un conjunto organizado de datos utilizado para análisis o para alimentar modelos de aprendizaje automático. Estos conjuntos pueden contener una variedad de tipos de datos, que van desde números y texto hasta imágenes y sonidos. Su tamaño y representatividad están diseñados para permitir la extracción de patrones, realizar inferencias o entrenar algoritmos. A pesar de su poder, los datasets también plantean desafíos, como la calidad de los datos, la privacidad y la gestión de grandes volúmenes de información. Asegurar la integridad y relevancia de los datos es fundamental para obtener resultados precisos y confiables (KeepCoding, 2023).

Neurona artificial

Explica Sarmiento (2020) que “Una neurona artificial se compone de una o múltiples entradas, un peso, un bias, un sumador y una función de transferencia (o de activación)” (p. 4). En términos generales, una neurona artificial está compuesta por múltiples entradas, un peso, un sesgo, un sumador y una función de transferencia o activación. Sin embargo, incluso con múltiples entradas, no es suficiente para iniciar un proyecto o procedimiento. Esto significa que las ANNs utilizan varias neuronas dispuestas en paralelo, formando lo que se conoce como una capa. Una ANN puede tener una sola capa o varias capas.

Las redes neuronales artificiales

Multilayer perceptron (MLP)

Según Sarmiento (2020) “Es una ANN compuesta por múltiples capas de neuronas ubicadas entre la entrada y la salida, conocidas como capas ocultas. La adición estas capas permite incrementar el potencial de la red para solucionar problemas de clasificación y regresión con conjuntos de datos complejos” (p. 6). Se hace referencia

al concepto de Perceptrón Multicapa (MLP), el cual posee capas ocultas y es utilizado con conjuntos de datos complejos que son no lineales y discontinuos. Para su entrenamiento, se utiliza ampliamente el algoritmo de retropropagación (backpropagation, BP).

Las redes neuronales recurrentes

Conforme a Sarmiento (2020) “Las redes neuronales recurrentes poseen una arquitectura donde las conexiones entre neuronas forman un gráfico dirigido a lo largo de una secuencia temporal, lo que les permite ser útiles para analizar datos en series de tiempo, exhibir un comportamiento dinámico, tener una memoria interna y mantener información sobre lo que sucedió en pasos de tiempos anteriores” (p. 6). Se explica que las redes neuronales recurrentes (RNNs) son principalmente utilizadas en aplicaciones de análisis de texto y de habla en series temporales con un comportamiento dinámico.

Las redes neuronales convolucionales

Según Sarmiento (2020) “Las redes neuronales convolucionales son inspiradas en la corteza visual, la cual consiste en mapas de campos receptivos locales que responden a los estímulos únicamente en una región del campo visual, y disminuyen a medida que la corteza se mueve hacia adelante; los campos receptivos se superponen de modo que cubren todo el campo visual” (p. 7). Se propone que las Convolutional Neural Networks (CNN) están inspiradas en la estructura de la corteza visual, donde los mapas de campos receptivos locales adoptan la arquitectura de los Perceptrones Multicapa Profundos (DMLP). Sin embargo, la principal diferencia es que cada neurona se conecta solo a un área local de neuronas en la capa siguiente.

El uso de distintas redes neuronales convolucionales preentrenadas

De acuerdo con Guerrero (2022) “Para la construcción del modelo se hizo uso de distintas redes neuronales convolucionales preentrenadas, obteniendo el mejor resultado con la red DenseNet121. Finalmente se obtuvieron valores de precisión de

94.4% en el modelo final” (…) (p. 6). Lo que demuestra que los resultados obtenidos indican que el uso de la red neuronal convolucional preentrenada DenseNet121 ofrece un rendimiento superior en comparación con otras redes, por lo tanto, es altamente efectiva en tareas de clasificación para que se logren altos niveles de precisión en el modelo final.

Métricas que se utilizan para medir los modelos

Exactitud

Según Guerrero (2022) “Es la medida de rendimiento más intuitiva y representa una relación entre la clasificación correcta y el total de observaciones” (p. 50). Lo que demuestra que la exactitud de las métricas de rendimiento de ML es fundamental para medir la efectividad de un modelo, ya que muestra la relación entre las predicciones correctas y el total de predicciones realizadas. Así, la exactitud ofrece una comprensión clara de cuán eficaz es el modelo en cuanto a la cantidad de predicciones correctas en relación con el número total de observaciones. Por lo tanto, mantener un alto nivel de exactitud es vital para el éxito de los modelos de aprendizaje automático en aplicaciones prácticas.

Precisión

De acuerdo a Guerrero (2022) “Es la relación entre el número de clasificaciones positivas realizadas correctamente y el total de clasificaciones positivas realizadas. Una alta precisión se relaciona con la baja tasa de falsos positivos” (p. 50). Lo que demuestra que la precisión es una métrica esencial en el rendimiento de ML para evaluar la exactitud de las predicciones positivas. Por lo tanto, una alta precisión indica que el modelo tiene una baja tasa de falsos positivos y es eficaz en la identificación correcta de las clasificaciones positivas. Esto es crucial para aplicaciones donde los falsos positivos pueden tener consecuencias significativas, asegurando así la confiabilidad y efectividad del modelo en situaciones reales. Por lo tanto, mantener una alta precisión es vital para garantizar que el modelo haga predicciones positivas acertadas y minimice errores críticos.

Sensibilidad

De acorde a Guerrero (2022) “Es la proporción de clasificaciones positivas realizadas correctamente para todos los datos en la clase real” (p. 51). Lo que demuestra que la sensibilidad de las métricas de rendimiento de ML es clave para evaluar la capacidad del modelo para identificar correctamente los verdaderos positivos. Por lo tanto, una alta sensibilidad indica que el modelo es efectivo en detectar la mayoría de los casos positivos reales, lo que es crucial para que se minimicen los falsos negativos y se asegure una mayor fiabilidad en la detección de condiciones o eventos de interés en aplicaciones prácticas del ML.

Especificidad

Según Guerrero (2022) “Es una medida que indica el número de clasificaciones negativas que fueron clasificadas correctamente como negativas” (p. 51). Lo que demuestra que la especificidad de las métricas de rendimiento de ML es importante para evaluar la capacidad del modelo para identificar correctamente los verdaderos negativos. Por lo tanto, una alta especificidad indica que el modelo es efectivo en descartar correctamente la mayoría de los casos negativos reales, lo que es fundamental para minimizar los falsos positivos y garantizar una mayor precisión en la detección de los casos negativos.

F1 - score

A juzgar por Rivas (2023) “F1 - score representa una forma de medición entre la presión y sensibilidad de manera similar, se opta los tres primeros añadiendo a su estudio la exactitud que representa la cantidad de aciertos correctos” (p. 101). Se refiere a una métrica combinada que considera tanto la precisión (Precision) como la sensibilidad (Recall) de un modelo de clasificación. El F1-score es una medida que busca el equilibrio entre ambas métricas, proporcionando una evaluación más robusta del rendimiento del modelo en términos de su capacidad para identificar correctamente tanto los casos positivos como los negativos. Es particularmente útil en situaciones donde hay un desbalance significativo entre las clases de datos.

La curva ROC (AUC-ROC)

De acuerdo a Rivas (2023) “La AUC es la medida para evaluar el desempeño de los límites” (p. 101). Se refiere a que la AUC (Area Under the Curve) es una métrica utilizada para evaluar la capacidad de un modelo de clasificación para distinguir entre clases. En particular, se aplica a la curva ROC (Receiver Operating Characteristic), que representa la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) a diferentes umbrales de decisión. La AUC mide el área bajo esta curva ROC y proporciona una medida única del rendimiento del modelo, donde un valor más alto de AUC indica un mejor rendimiento en la clasificación de los datos.

Validación Cruzada

Dependiendo de Rivas (2023) “La segunda/última etapa utiliza estos clasificadores y utiliza la validación cruzada estrategia para asegurar la comparación de tarifas entre clasificadores” (p. 27). Se refiere a que en esta etapa del estudio o del proceso, se utilizan los clasificadores entrenados previamente para evaluar su desempeño de manera más robusta y generalizable. La validación cruzada es una técnica que se emplea para estimar cómo se comportará un modelo en un conjunto de datos independiente, dividiendo los datos en subconjuntos de entrenamiento y prueba repetidamente. Esto permite evaluar la variabilidad del rendimiento del modelo y asegurar que las comparaciones entre diferentes clasificadores sean justas y no estén sesgadas por una sola partición de los datos.

Preprocesamiento de los datos

Conforme a Guerrero (2022) “El preprocesamiento previo al entrenamiento de las redes se llevó a cabo utilizando la librería de Keras v.2.8, se realizó una normalización de los datos redimensionando todas las imágenes a una dimensión de 128X128 píxeles y un reescalado de 1/255” (p. 38). Se refiere a que, antes de entrenar las redes neuronales, se aplicaron técnicas de preprocesamiento utilizando Keras versión 2.8. Estas técnicas incluyeron la normalización de los datos, redimensionando todas las imágenes a un tamaño uniforme de 128x128 píxeles y reescalando los valores de

los píxeles para que se encuentren en el rango de 0 a 1 (dividiendo cada valor de píxel por 255).

Técnicas de aumentación de datos

Según Guerrero (2022) “Al tratarse de un conjunto de datos con un número limitado de elementos, se aplicaron distintas técnicas de aumentación de datos como el estiramiento, rotación, translación, corte y otras deformaciones de forma aleatoria” (p. 38). Se refiere a que el uso de diversas técnicas de aumentación de datos como el estiramiento, rotación, translación, corte y otras deformaciones aleatorias se implementó debido a que el conjunto de datos tenía un número limitado de elementos. Estas técnicas permitieron incrementar la variedad y cantidad de datos disponibles para el entrenamiento del modelo, mejorando así su capacidad de generalización y rendimiento predictivo.

Función de activación de Simplicidad

Basado en Toquero (2021) “Debemos usar funciones con bajo coste computacional, pues suele presentarse un uso masivo de ellas” (p. 28). Se refiere a que es importante utilizar funciones de activación que tengan un bajo costo computacional, ya que generalmente se emplean de manera intensiva en modelos de ML. La eficiencia computacional de estas funciones es crucial debido a su frecuente uso en la arquitectura de las redes neuronales, lo cual puede afectar significativamente el rendimiento y la velocidad adecuada en el procesamiento del modelo. Esto puede influir significativamente en la eficacia del modelo y en su capacidad para manejar grandes volúmenes de datos de manera eficiente.

Técnica Bagging

Los diferentes modelos creados con la técnica Bagging pueden considerarse como algoritmos que buscan respuestas (o hipótesis) en un data set (o espacio h). Como cada algoritmo tiene un set de datos diferentes, cada uno creará unas hipótesis diferentes sobre la realidad. Esto significa que La técnica de Bagging crea múltiples

modelos con diferentes subconjuntos de datos, cuyas hipótesis variadas al combinarse mejoran la precisión y robustez, reduciendo el sobreajuste. Esto aumenta la estabilidad y capacidad de generalización del modelo, reduciendo la variabilidad y el riesgo de sobreajuste (Parra, 2019).

Función de coste o pérdida

De acuerdo a Toquero (2021) “Representa la suma del error: la diferencia entre el valor predicho y el real. Se emplea en problemas supervisados, es decir, con la variable respuesta conocida” (p. 34). Se refiere a que, en problemas de aprendizaje supervisado, se calcula el error como la diferencia entre los valores que el modelo predice y los valores reales conocidos. Este error se suma a lo largo de todas las predicciones para evaluar el rendimiento del modelo. La suma del error es una medida fundamental para ajustar y mejorar los modelos, ya que proporciona una indicación clara de cuán cerca están las predicciones del modelo respecto a los valores reales.

Optimizador de descenso del gradiente estocástico o SGD

Dependiendo de Toquero (2021) “Descenso del gradiente estocástico o SGD: optimizador con descenso de gradiente y momento. Puede incluirse la aceleración de Nesterov” (p. 34). Se refiere a que el descenso del gradiente estocástico (SGD) es una técnica de optimización que utiliza el gradiente del error para ajustar los parámetros del modelo. Esta técnica incluye el uso del momento, que ayuda a acelerar el proceso de convergencia y reducir las oscilaciones en el camino hacia el mínimo del error. Además, la aceleración de Nesterov puede ser incorporada para mejorar aún más la velocidad y estabilidad del proceso de optimización.

Python

Según González (s/f) “Python es un lenguaje que todo el mundo debería conocer. Su sintaxis simple, clara y sencilla; el tipado dinámico, el gestor de memoria, la gran cantidad de librerías disponibles y la potencia del lenguaje, entre otros” (…) (pág. 8). Se refiere a que Python es un lenguaje accesible y poderoso, ideal tanto para

principiantes como para desarrolladores experimentados, facilitando la programación gracias a sus cualidades técnicas y su amplio ecosistema de herramientas. Además, la gran cantidad de librerías disponibles amplía su aplicabilidad en diversas áreas como la ciencia de datos, el desarrollo web y la inteligencia artificial.

Google Colab

Conforme a López (2023) “Google Colaboratory, comúnmente conocido como Google Colab, es una herramienta de investigación desarrollada por Google que permite a los usuarios escribir y ejecutar código en un entorno virtual basado en la nube”. Se refiere a que, al ser un entorno virtual, permite a los usuarios aprovechar recursos computacionales como GPU y TPU para realizar tareas de procesamiento intensivo, lo que es especialmente útil en proyectos de ML y análisis de datos. Además, Colab integra fácilmente con otras herramientas de Google y soporta la colaboración en tiempo real, lo que lo convierte en una herramienta valiosa tanto para la investigación como para la enseñanza.

TensorFlow y Keras

De acuerdo a Géron (s/f). “Utiliza TensorFlow y Keras para crear y entrenar redes neuronales para visión por ordenador, procesamiento del lenguaje natural, modelos generativos y aprendizaje profundo por refuerzo” (pág. 19). Se refiere a que estas herramientas permiten a los desarrolladores crear y entrenar redes neuronales de manera eficiente, gracias a sus funciones de alto nivel y su capacidad para manejar grandes volúmenes de datos. TensorFlow proporciona una estructura robusta para construir modelos complejos, mientras que Keras simplifica el proceso con una interfaz más accesible y amigable para los usuarios.

OpenCV

Según Igual & Medrano (2023) “Las siglas OpenCV provienen de los términos anglosajones (Open Source Computer Vision Library). Por lo tanto, OpenCV es una librería de tratamiento de imágenes, destinada principalmente a aplicaciones de visión

por computador en tiempo real” (pág. 7). Se refiere a que esta librería es ampliamente utilizada en el desarrollo de proyectos que requieren análisis visual, como la detección de objetos, el reconocimiento de rostros, y la automatización de tareas visuales. OpenCV proporciona herramientas y algoritmos que permiten a los desarrolladores trabajar eficientemente con imágenes y videos, facilitando el desarrollo de soluciones innovadoras en el campo de la visión artificial.

CAPÍTULO III

MARCO METODOLÓGICO

En este capítulo se presenta la metodología que constituye el plan o conjunto de fases en forma ordenada, que permite mostrar con claridad lo que se hizo, y el porqué, junto con las razones de la elección o realización de cada una de ellas. La metodología deductiva permite especificar e identificar los atributos del tema para la tipificación del cáncer de pulmón en el modelo ML para obtener conclusiones y recomendaciones en base a los resultados, se logra que la información sea relevante para el proyecto de investigación y el método analítico sintético permite que los investigadores lleguen a la autenticidad de los hechos, se recopila elementos de observación en la problemática de la investigación planteada, el modelo de IA está basada en modelos ML para dar soporte diagnóstico médico del mismo cáncer, con un análisis a profundidad para obtener buenos resultados (Tuarez & Vera, 2022).

Naturaleza de la Investigación

El presente capítulo se desarrolló bajo el paradigma positivista. Según lo plantea Sánchez (2013) citado por Julca (2020) “El paradigma es positivista, persigue la verificación rigurosa de proposiciones generales a través de la observación empírica, el experimento en muestras de amplio alcance, desde una aproximación cuantitativa, con el fin de verificar, perfeccionar leyes referidas a lo educativo.” (p. 42). Su finalidad fue realizar de una categorización de exámenes de rayos X del cáncer de pulmón. A fin de que las bases teóricas e indicadores permitan dar respuestas a las preguntas de investigación formuladas: ¿Identificar el modelo de ML que permita obtener una caracterización de radiografías médicas del cáncer de pulmón? ¿Definir el modelo de ML para la realización de la abstracción matemática de todo lo que tenga el modelo? Se preparó un algoritmo de DL para la caracterización de rayos X del mismo cáncer.

En este sentido, Hernández y Mendoza (2018) refieren que la investigación cuantitativa representa un conjunto de procesos organizados de manera secuencial

para comprobar ciertas suposiciones, en el cual cada fase precede a la siguiente y no se puede dejar de realizar ningún paso, el orden es riguroso, aunque se permite redefinir alguna etapa, si es necesario. La finalidad principal de los estudios cuantitativos es que sea determinado el modelo de ML que permitió clasificar imágenes radiológicas del cáncer de pulmón en las personas de todas las edades del dataset “El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD,” con el DL. Radica en el descubrimiento del conocimiento y la minería de datos, el reconocimiento de patrones por métodos de la neuro-computación para imágenes de Rayos X. La mayor parte de las técnicas de ML son basados en inteligencia artificial y en el análisis de regresión lineal.

Los diseños no experimentales, según Hernández y Mendoza (2018), son aquellos estudios que se realizan sin la manipulación de la variable independiente, debido a que ya ha sucedido y en los que solo se observan los fenómenos en su ambiente natural. Estos se clasifican por su dimensión temporal o el número de momentos o puntos en el tiempo en los cuales se recolectan datos, de la siguiente manera: Diseño Transeccional o Transversal, la recolección de datos se realiza en un solo momento, en un tiempo único (Hernández y Mendoza, 2018). Para la presente investigación se ha manipulado la patología computacional, en la que se ha utilizado el diseño no experimental, porque no se va a realizar experimentos.

El Diseño Transeccional o Transversal es la recolección de datos se realiza en un solo momento, en un tiempo único (Hernández y Mendoza, 2018). En este proyecto de investigación se va a analizar las diferentes técnicas de ML para clasificación de imágenes médicas pulmonares; examinar el dataset de “El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD,” que contengan información relevante y válida para diseñar el modelo de ML para la clasificación de imágenes de cáncer de pulmón; evaluar el o los modelos de clasificación de ML de imágenes de cáncer pulmonar; medir la precisión de los modelos de categorización de ML usando las imágenes de ese cáncer específico, realizar de la abstracción matemática de todo lo que tenga el modelo.

La investigación documental se basa en “la búsqueda, recuperación, análisis, critica e interpretación de datos secundarios, es decir obtenidos y registrados en fuentes documentales impresas, audiovisuales o electrónicas” (Arias, 2016, p. 27). Ha

desempeñado un papel fundamental de los requisitos necesarios para la obtención del repositorio de imágenes médicas denominado dataset de “El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD,” de esta manera se analizaron las redes neuronales aplicables al algoritmo de ML que integran tecnologías avanzadas de procesamiento de imágenes médicas, con el fin de que se ha logrado proporcionar un enfoque eficiente y confiable en la identificación temprana de esta enfermedad.

De esta manera, el nivel de investigación se llevó a cabo de manera descriptiva. Recolectando la información sobre la variable o variables a investigar para posteriormente describirla (s) (Hernández y Mendoza, 2018). Se centró en proporcionar una descripción detallada y sistemática de las características del cáncer del pulmón y los casos: maligno, benigno y normal. Proporcionando una representación precisa y objetiva, utilizando herramientas de IA. Este tipo de investigación se fundamenta en la caracterización del modelo ML con el fin de conocer su estructura o comportamiento con el dataset de “El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD,” con edades distintas. Los resultados de estos estudios se ubican en un nivel intermedio en cuanto a la profundidad de los conocimientos. Este nivel permite que el modelo ML clasifique los exámenes de rayos X de los casos del mismo cáncer.

Población

La población se refiere al conjunto de cosas, objetos, sujetos que guardan una característica en común, la muestra implica un subconjunto representativo de la población (Arias, 2012). En el presente proyecto de investigación se va a trabajar con el dataset “El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD,” del Hospital de Enseñanza de Irak-Oncología/Centro Nacional de Enfermedades del Cáncer (IQ-OTH/NCCD), contiene radiografías de pacientes diagnosticados con cáncer de pulmón en diferentes etapas, así como sujetos sanos. Este conjunto de datos es internacional, la mayoría de ellos provienen de lugares de la región central de Irak y comprende un total de 1.097 imágenes. Estas varían en género, edad, nivel educativo, área de residencia y estado de vida (Solano, 2021). La herramienta con la que se va a trabajar es mediante un algoritmo de ML que permita categorizar

imágenes radiológicas del mismo cáncer mediante los casos como: maligno, benigno, normal; utilizando el dataset mencionado.

Técnicas e instrumentos de recolección de datos

Una técnica son las estrategias empleadas para recabar la información requerida y así construir el conocimiento de lo que se investiga, la técnica cuantitativa son la recopilación documental, la recopilación de datos a través de cuestionarios que asumen el nombre de encuestas o entrevistas y el análisis estadístico de los datos (Gonzáles, 2020). Es necesario definir las técnicas de recolección de datos para seleccionar o construir los instrumentos que nos permitan obtenerlos de la realidad. El instrumento es un mecanismo que usa el investigador para recolectar y registrar la información: formularios, pruebas, test, escalas de opinión y listas de chequeo (Martín, Manjarrés & Martín, 2019).

Técnica de recolección de datos

La revisión documental es un componente crucial en la metodología de investigación “permite ubicar los pasos y acciones; este instrumento incluye protocolos de búsqueda, así como revisión de fuentes de información” (Bernate & Fonseca, 2023, pág. 6). Por lo tanto, la técnica que se llevó a cabo en el trabajo de titulación es de revisión documental. Está técnica permite analizar sobre los diferentes procesos de ML para clasificación de imágenes Rayos X del pulmón y examinación del dataset “El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD,” que contengan información relevante y válida para que se pueda colaborar en el diagnóstico del cáncer de pulmón.

La encuesta es como una entrevista basada en un cuestionario. No obstante, esta definición es cuestionable cuando se considera que la encuesta es autoadministrada. En tal caso, no existe un diálogo entre entrevistador y encuestado; más bien, el encuestado interactúa solo consigo mismo, guiado únicamente por el cuestionario que actúa como el instrumento metodológico. Por lo tanto, la clasificación de la encuesta como una entrevista es discutible en este contexto (Lanuez y Fernández, 2014). Está

técnica permite definir el modelo de ML para la clasificación de imágenes radiológicas del cáncer pulmonar.

Operacionalización de la variable

La operacionalización de variables es un proceso crucial en la investigación, ya que permite convertir conceptos abstractos en datos cuantificables y observables. Las variables como: antigüedad, organización, eficiencia, magnitud, productividad, entre otras; pueden ser atributos, cualidades o características observables de personas, objetos o instituciones que expresan magnitudes que varían discreta o continuamente (Ñaupas, 2014). A continuación, se presenta la operacionalización de la variable "Definir el modelo de ML para la clasificación de imágenes radiológicas del cáncer pulmonar", se presentan en la Tabla 1.

Tabla 1. *Matriz de Operacionalización de las variables*

Variable	Definición	Dimensión	Indicador	Items o Pregunta	Fuente (opcional)
Definir el modelo de ML para la clasificación de imágenes radiológicas del cáncer pulmonar	Modelo Random Forest	• Precisión • Sensibilidad • Especificidad • Puntaje F1 • AUC-ROC • Tiempo de Entrenamiento • Interpretabilidad	• Precisión 0.92 • Sensibilidad 0.89 • Especificidad 0.94 • Puntaje F1 0.91 • AUC-ROC 0.96 • Tiempo de Entrenamiento 2 horas • Interpretabilidad Moderada	1	Revisión documental

Modelo Red Neuronal CNN	• Precisión • Sensibilidad • Especificidad • Puntaje F1 • AUC-ROC • Tiempo de Entrenamiento • Interpretabilidad	• Precisión 0.94 • Sensibilidad 0.92 • Especificidad 0.95 • Puntaje F1 0.93 • AUC-ROC 0.97 • Tiempo de Entrenamiento 4 horas • Interpretabilidad Baja	2, 3
Modelo SVM	• Precisión • Sensibilidad • Especificidad • Puntaje F1 • AUC-ROC • Tiempo de Entrenamiento • Interpretabilidad	• Precisión 0.88 • Sensibilidad 0.85 • Especificidad 0.91 • Puntaje F1 0.87 • AUC-ROC 0.94 • Tiempo de Entrenamiento 1.5 horas • Interpretabilidad Alta	1, 2, 4, 5

Modelo K-NN	• Precisión • Sensibilidad • Especificidad • Puntaje F1 • AUC-ROC • Tiempo de Entrenamiento • Interpretabilidad	• Precisión 0.86 • Sensibilidad 0.83 • Especificidad 0.89 • Puntaje F1 0.85 • AUC-ROC 0.92 • Tiempo de Entrenamiento 1 hora • Interpretabilidad Moderada	5, 6

Instrumento de recolección de datos

El cuestionario consiste en un conjunto de preguntas, normalmente de varios tipos, preparado sistemática y cuidadosamente, sobre los hechos y aspectos que interesan en una investigación o evaluación y que puede ser aplicado en formas variadas, entre las que destacan su administración a grupos o su envío por correo La finalidad del cuestionario es obtener, de manera sistemática y ordenada, información acerca de la población con la que se trabaja, sobre las variables objeto de la investigación o evaluación (García, 2003). Por lo tanto, se ha realizado un cuestionario para definir el modelo de ML para la clasificación de imágenes radiológicas del cáncer pulmonar.

Validez

La validez es la pertinencia de un instrumento de medición, para medir lo que se quiere medir, se refiere a la exactitud con que el instrumento mide lo que se propone medir. También validez es conocida como validez hipótesis de trabajo y se determina en base al juicio de expertos (Paitán, Mejía, Ramírez & Paucar, 2014). El juicio de expertos es una estrategia frecuente para contrastar la validez de contenido donde se estima el grado de concordancia entre los expertos, en la que los investigadores experimentados en el tema emiten sus valoraciones sobre los indicadores o ítems de la herramienta (Zamora, Serrano, & Martínez, 2020). Por lo tanto, en este proyecto de investigación se basa en expertos, se va a realizar en base a cinco expertos, cuatro en el área de software y uno de en área de metodología. A fin de determinar si cumple con la finalidad establecida de acuerdo a los objetivos propuestos. A continuación, la lista de los expertos.

Tabla 2. *Lista de Expertos*

Expertos	Título profesional
Msc. Sandino Jaramillo	Director de Carrera de Ingeniería de Software
MSc. Carlos Antonio Ayala Tipán	Académico/Investigador
Msc. Miguel Ángel Fernández Marín	Profesor
Msc. Miguel Flores	Doctor en Estadística e Investigación Operativa
PhD. Roberto Andrade	Investigador/ Docente

Técnicas de análisis de los datos

Según Arias (2012) indica que “se describen las distintas operaciones a las que serán sometidos los datos que se obtengan: clasificación, registro, tabulación y codificación si fuere el caso” (p. 111). En otras palabras, las técnicas de análisis de los datos “se refiere a los cruces, análisis de varianza, regresiones, análisis de conglomerados, regresiones logísticas y cuanta cosa haya en la batería de herramientas estadísticas al uso (y en particular las que aparecen en el SPSS)” (Cerón & Cerâon, 2006). En este proyecto el instrumento se analiza a través de la estadística descriptiva. La estadística descriptiva se enfoca en la clasificación de imágenes de cáncer pulmonar utilizando técnicas de aprendizaje automático (ML). El análisis de las radiografías de tórax se lleva a cabo considerando varios parámetros, tales como el procesamiento de imágenes; el entrenamiento, la evaluación y la visualización de resultados del modelo SVM, cada sección está delimitada y se incluyen gráficos relevantes para la evaluación del modelo, utilizando el dataset "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD".

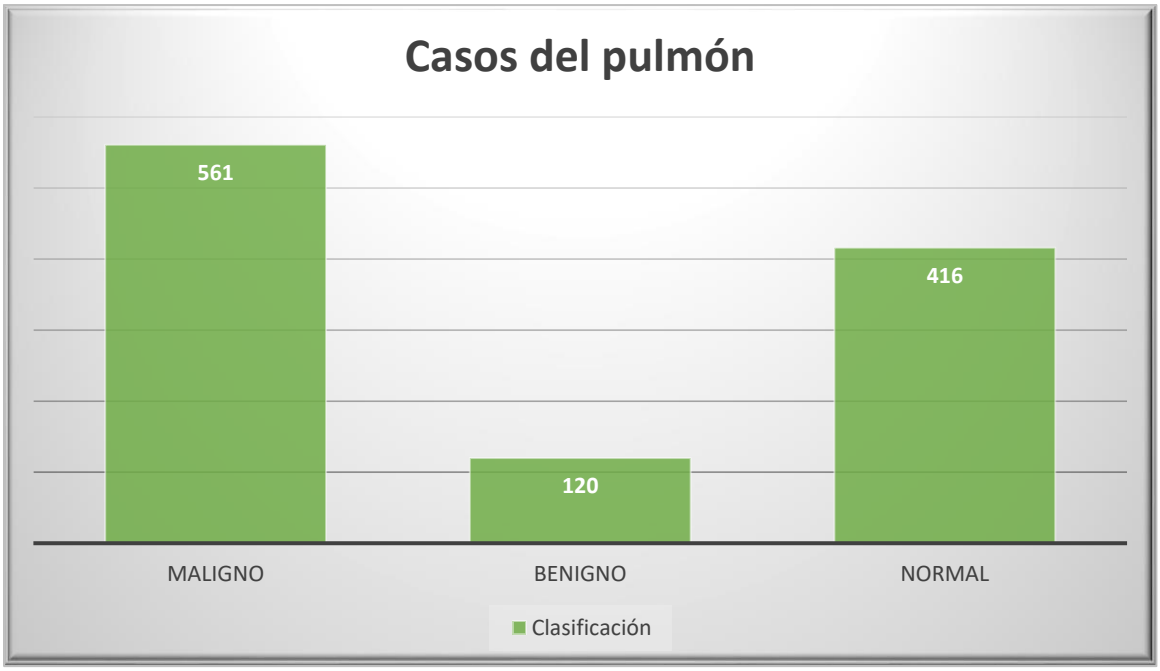

Figura 3. *Casos del pulmón*

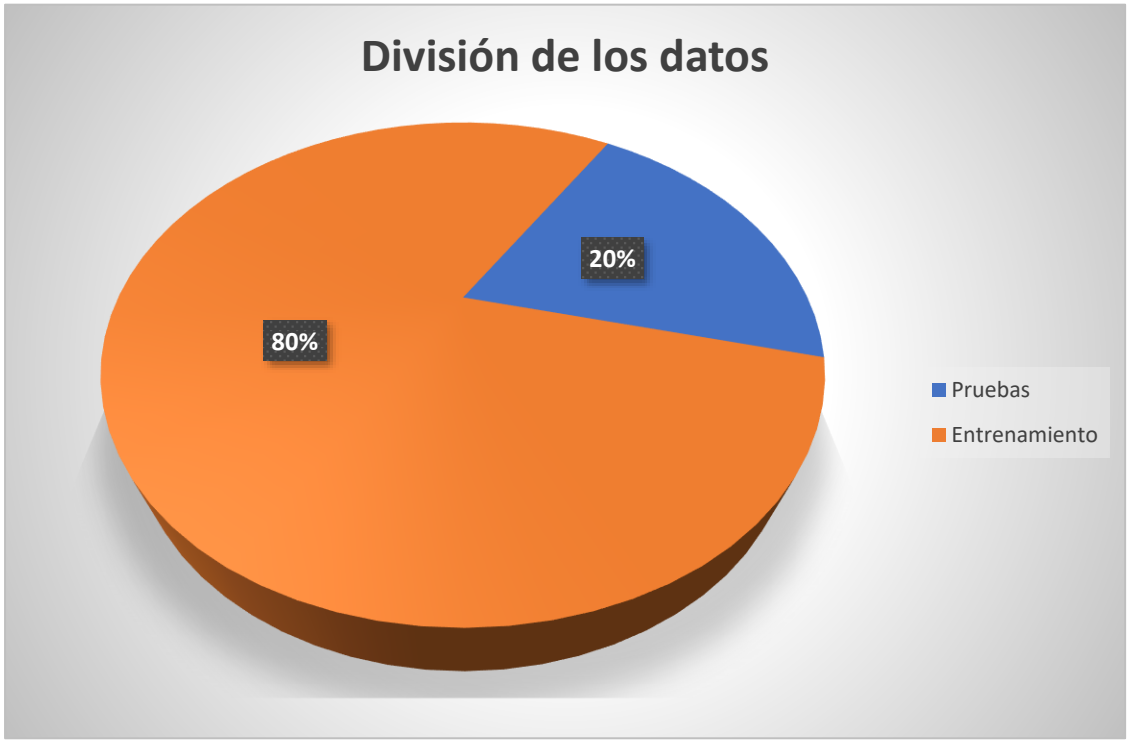

Figura 4. *División de los datos*

Metodología del producto

La metodología Kanban es una herramienta eficaz para la gestión de proyectos y la optimización del flujo de trabajo, tanto en equipos como de forma individual. Con su enfoque visual y su capacidad para promover la mejora continua, permite a los equipos gestionar eficientemente sus tareas y recursos. Implementada a través de tableros visuales, esta metodología ayuda a visualizar los flujos de trabajo y la carga de trabajo. Además, fomenta una cultura de mejora constante, donde los equipos revisan y ajustan regularmente sus procesos para incrementar su eficiencia y efectividad (Martins,2024).

Continuando con el desarrollo del presente proyecto denominado *“Diseño de un Modelo basado en Técnicas de Machine Learning para la Clasificación de Imágenes Médicas del Cáncer Pulmonar: Contribuciones al Diagnóstico Médico”*, se utiliza la metodología Ágil Kanban a través de fases: la primera visualización del flujo de trabajo: en un tablero Kanban se muestra todas las etapas del software, las columnas del tablero pueden incluir "Por hacer", "En progreso", "Listo para revisión" y "Completado". La segunda definición de elementos de trabajo: preprocesamiento de datos, la selección y entrenamiento de algoritmos de ML, la evaluación del modelo y la realización de la abstracción matemática de todo lo que tenga el modelo. La tercera limitación del trabajo en curso (WIP): establecer un límite de tres elementos en la columna "En progreso". La cuarta es cuando la tarea está casi terminada y la quinta es cuando la tarea está completada. La sexta mejora continua: se identifica áreas de oportunidad y aplica cambios incrementales permitiendo optimizar el flujo de trabajo y aumentar la eficiencia.

CAPÍTULO IV

ANÁLISIS E INTERPRETACIÓN DE LOS RESULTADOS

En el ámbito del análisis e interpretación de los resultados se procesa estadísticamente toda la información obtenida, auxiliándose de gráficos, tablas, diagramas que le permitan analizar e interpretar los resultados obtenidos con mayor facilidad para poder realizar generalizaciones y arribar a conclusiones y recomendaciones basadas en los resultados obtenidos a partir de la contrastación con la teoría que se parte (Calderón y Piñeiro, 2003). En este capítulo se define un modelo de ML para clasificar imágenes médicas del cáncer pulmonar, durante este proceso, se utiliza una matriz comparativa de modelos de ML para identificar cuál es el más apropiado para su implementación práctica. Además, se realiza un análisis detallado del dataset de imágenes clínicas del cáncer de pulmón mediante otra matriz. Asimismo, se lleva a cabo la formulación matemática del modelo de ML utilizado para clasificar estas imágenes médicas particulares.

Definición del modelo de ML para la clasificación de imágenes radiológicas del cáncer pulmonar

Para definir el modelo de ML, se presenta una matriz comparativa de diversos enfoques de ML, lo cual permite elegir el más adecuado para la clasificación de imágenes médicas del cáncer pulmonar. Esta matriz se basa en una evaluación exhaustiva de varias métricas de rendimiento, tales como precisión, sensibilidad, especificidad, puntaje F1, AUC-ROC, tiempo de entrenamiento e interpretabilidad. Entre los métodos considerados se incluyen Random Forest, Red Neuronal CNN, SVM y K-NN. Finalmente, se concluye que el modelo más adecuado es SVM con el tiempo de entrenamiento de 1.5 horas y la interpretabilidad alta. Los resultados obtenidos son a través del instrumento de revisión documental.

Tabla 3. *Definición del modelo de ML para la clasificación de imágenes médicas del cáncer pulmonar*

Modelo	Precisión	Sensibilidad	Especificidad	Puntaje F1	AUC-ROC	Tiempo de Entrenamiento	Interpretabilidad
Random Forest	0.92	0.89	0.94	0.91	0.96	2 horas	Moderada
Red Neuronal CNN	0.94	0.92	0.95	0.93	0.97	4 horas	Baja
SVM	0.88	0.85	0.91	0.87	0.94	1.5 horas	Alta
K-NN	0.86	0.83	0.89	0.85	0.92	1 hora	Moderada

Para comprender que es la sensibilidad se debe tener en cuenta que según Fos (2016) “El mejor rendimiento se ha conseguido con un clasificador SVM” (pág. 92). Lo que demuestra que el modelo SVM (Support Vector Machine) se ha destacado al ofrecer el mejor rendimiento, superando significativamente a otros clasificadores en términos de interpretabilidad y en un tiempo de entrenamiento más corto en comparación con la Red Neuronal CNN y Random Forest. Sin embargo, su desempeño es menor en métricas como AUC-ROC, puntaje F1, especificidad, sensibilidad y precisión en comparación con la Red Neuronal CNN y Random Forest.

Según Salvat (2023) “Una vez comprobada la capacidad predictiva del modelo, se procede a entrenar el Random Forest Classifier con sus hiperparámetros predeterminados. Una vez entrenado el modelo, se realiza un estudio de los resultados brindados por este (…)” (p. 53). Se refiere que Random Forest es más bajo que en el desempeño de métricas como AUC-ROC, puntaje F1, especificidad, sensibilidad y precisión. Pero es mejor que la Red Neuronal CNN en el tiempo de entrenamiento y en la interpretabilidad. Sin embargo, es más mejor que los modelos de SVM y K-NN, en el desempeño de métricas como AUC-ROC, puntaje F1, especificidad, sensibilidad y precisión.

Para entender cómo se construyó el modelo, se debe tener en cuenta que según Guerrero (2022) “Para la construcción del modelo se hizo uso de distintas redes neuronales convolucionales preentrenadas, obteniendo el mejor resultado con la red DenseNet121. Finalmente se obtuvieron valores de precisión de 94% en el modelo final” (…) (pág. 6). Lo que demuestra que la Red Neuronal CNN sobresale en métricas de rendimiento, mostrando una AUC-ROC más alta, así como en puntaje F1, especificidad, sensibilidad y precisión. Sin embargo, su entrenamiento requiere más tiempo y su nivel de interpretabilidad es bajo.

Según Prieto (2022) “Los métodos de vecinos más cercanos (KNN) suelen tener mejores resultados” (p. 32). Se refiere que el modelo KNN se destaca por su tiempo de entrenamiento de 1 hora, por lo que es más eficiente en comparación con otros modelos como SVM, la Red Neuronal CNN y Random Forest. Sin embargo, en el desempeño de métricas como AUC-ROC de 0.92, puntaje F1 de 0.85, especificidad de 0.89, sensibilidad de 0.83 y precisión de 0.86 es inferior al del SVM. A pesar de

ello, su nivel de interpretabilidad es moderado en comparación con el método de Random Forest.

Análisis del dataset de imágenes clínicas del cáncer de pulmón, para el aseguramiento de la calidad de los datos previo al entrenamiento y a la validación del modelo

Para realizar el análisis del del dataset de imágenes clínicas del cáncer de pulmón, para el aseguramiento de la calidad de los datos previo al entrenamiento y a la validación del modelo, se presenta la siguiente matriz en la que se detalla el proceso, la descripción y las técnicas utilizadas. Los procesos considerados incluyen: recopilación de datos, organización de datos, normalización, aumento de datos, segmentación de imágenes, división del conjunto de datos y visualización de aumento de datos. Algunas de las técnicas utilizadas son: bases de datos públicas, estructuración en directorios, entre otras.

Tabla 4. *Análisis del dataset de imágenes clínicas del cáncer de pulmón*

Proceso	Descripción	Técnicas Utilizadas
Recopilación de Datos	Obtener imágenes médicas provenientes de fuentes confiables, tales como bases de datos públicas, hospitales o estudios científicos.	• Bases de datos públicas. • Colaboración con entidades médicas.
Organización de Datos	Estructurar las imágenes en directorios de acuerdo con sus etiquetas, como "maligno", "benigno" y "normal".	• Estructuración en directorios. • Etiquetado adecuado.

Normalización	Ajustar los valores de píxeles de las imágenes para que se encuentren dentro del rango [0, 1], mejorando así la estabilidad del modelo durante el entrenamiento.	Uso de ‘ImageDataGenerator’ en Keras con el parámetro 'rescale=1/255’.
Aumento de Datos	Generar variaciones de las imágenes para ampliar el tamaño del conjunto de datos y mejorar la robustez del modelo.	• Rotación • Traslación • Escalado • Espejado horizontal. • Corte
Segmentación de Imágenes	Resaltar áreas específicas de las imágenes, como nódulos pulmonares, como los nódulos pulmonares, para centrar la atención del modelo en las áreas más significativas.	Uso de técnicas de segmentación como: • Umbralización • Operaciones morfológicas • Contornos con OpenCV.
División del Conjunto de Datos	Dividir el conjunto de datos en conjuntos de entrenamiento y validación para evaluar el rendimiento del modelo.	Uso de ‘ImageDataGenerator’ en Keras con el parámetro ‘validation_split=0.2’.
Visualización de Aumento de Datos	Visualizar las imágenes aumentadas para verificar la diversidad y calidad de las transformaciones realizadas.	Plotting con Matplotlib para revisar las imágenes aumentadas, verificar que el aumento sea significativo y variado.

Conforme a Guerrero (2022) “El preprocesamiento previo al entrenamiento de las redes se llevó a cabo utilizando la librería de Keras v.2.8, se realizó una normalización de los datos redimensionando todas las imágenes a una dimensión de 128X128 píxeles y un reescalado de 1/255” (p. 38). Se refiere a que en el proceso de Normalización se ajusta los valores de píxeles de las imágenes para que se encuentren dentro del rango [0, 1] y se utiliza de ‘ImageDataGenerator’ en Keras con el parámetro 'rescale=1/255’. Entre otros procesos son: Recopilación de Datos obtener imágenes médicas provenientes de fuentes confiables a través de bases de datos públicas; Organización de Datos se utiliza la técnica de estructuración en directorios de acuerdo con sus etiquetas, como "maligno", "benigno" y "normal"; División del Conjunto de Datos para evaluar el rendimiento del modelo con el uso de ‘ImageDataGenerator’ en Keras con el parámetro ‘validation_split=0.2’.

Según Guerrero (2022) “Al tratarse de un conjunto de datos con un número limitado de elementos, se aplicaron distintas técnicas de aumentación de datos como el estiramiento, rotación, translación, corte y otras deformaciones de forma aleatoria” (p. 38). Se refiere a que el proceso de Aumento de Datos genera variaciones de las imágenes de acuerdo a técnicas de aumentación como rotación, traslación, escalado, espejado horizontal y corte. Entre otros procesos son: Segmentación de Imágenes para centrar la atención del modelo en las áreas más significativas con técnicas de segmentación como: umbralización, operaciones morfológicas, contornos con OpenCV; Visualización de Aumento de Datos para verificar la diversidad y calidad de las transformaciones realizadas, utilizando Plotting con Matplotlib.

Desarrollo de una abstracción matemática para el modelo de ML y evaluación del desempeño

El problema de clasificación es de multiclase donde el objetivo es clasificar imágenes médicas del cáncer pulmonar en tres categorías: benigno (1), maligno (2) y normal (0), contribuir al diagnóstico médico. Para esto, se utilizará una representación matemática clara y precisa del problema. Se a evaluar el modelo mediante métricas cuantitativas como precisión, sensibilidad, especificidad, AUC-ROC y F1-score, y aplicar validación cruzada para asegurar su robustez y generalización. A

continuación, se presenta la abstracción matemática para el modelo de ML para la representación del problema de investigación:

Representación del Problema

1. **Espacio de Entrada:**

   Cada imagen médica se representa como un tensor de dimensiones (H, W, C) donde H es la altura, W es la anchura y C es el número de canales (por ejemplo, 3 para imágenes RGB).

2. **Espacio de Salida:**

   Las etiquetas de las imágenes se representan como un vector de clase $y \in \{0, 1, 2\}$, donde:
   0 corresponde a imágenes normales.
   1 corresponde a imágenes de cáncer benigno.
   2 corresponde a imágenes de cáncer maligno.

3. **Conjunto de Datos:**

   El conjunto organizado de datos es utilizado para análisis o para alimentar modelos de aprendizaje automático (KeepCoding, 2023). Sea $D = \{(\boldsymbol{x}_i, y_i)\}^{N}_{i=1}$ el conjunto de datos, donde $\boldsymbol{x}$ es la i - ésima imagen y y_i es la etiqueta correspondiente.

Modelo de Clasificación

1. **Función de Hipótesis:**

   La técnica de Bagging mejora la precisión y robustez al combinar múltiples modelos con diferentes subconjuntos de datos, reduciendo así el sobreajuste y aumentando la estabilidad y capacidad de generalización (Parra, 2019). El modelo de ML puede representarse como una función $f(x; \theta)$, donde $\boldsymbol{x}$ es la imagen de entrada y θ son los parámetros del modelo. La función de hipótesis $f(x; \theta)$ produce una probabilidad para cada clase:

$$\hat{y} = f(x; \theta) = [P(y = 0|x; \theta), P(y = 1|x; \theta), P(y = 2|x; \theta)]$$

La clase predicha y es la que tiene la mayor probabilidad:

$$\hat{y} = \underset{j}{arg\ max}\ P(y = j|x;\ \theta)$$

2. **Función de Pérdida:**

   De acuerdo con Toquero (2021) “Representa la suma del error: la diferencia entre el valor predicho y el real. Se emplea en problemas supervisados, es decir, con la variable respuesta conocida” (p. 34). Se refiere a que se utiliza la entropía cruzada categórica para medir la discrepancia entre las etiquetas verdaderas y y las etiquetas predichas $\hat{y}$:

$$L(\theta) = -\frac{1}{N}\sum_{i\,=\,1}^{N}\sum_{j\,=\,0}^{2} 1_{[y_{i\,=\,j]}}\ log\ P(y = j|x_{i;}\ \theta)$$

   Aquí, $1_{[y_{i\,=\,j]}}$ es un indicador que vale 1 si $y_i = j$ y 0 en caso contrario.

3. **Optimización:**

   Dependiendo de Toquero (2021) “Descenso del gradiente estocástico o SGD: optimizador con descenso de gradiente y momento. Puede incluirse la aceleración de Nesterov” (p. 34). Se refiere a que los parámetros del modelo θ se actualizan para minimizar la función de pérdida $L(\theta)$ utilizando un algoritmo de optimización como el descenso de gradiente estocástico (SGD):

$$\theta :=\ \theta - \eta\ \nabla_\theta\ L(\theta)$$

   Donde η es la tasa de aprendizaje y $\nabla_\theta\ L(\theta)$ es el gradiente de la función de pérdida con respecto a θ.

Evaluación del Modelo

1. **Precisión (Accuracy):**

   De acuerdo a Guerrero (2022) “Es la relación entre el número de clasificaciones positivas realizadas correctamente y el total de clasificaciones positivas

realizadas. Una alta precisión se relaciona con la baja tasa de falsos positivos" (p. 50). Lo que demuestra que la precisión se calcula como:

$$\text{Precisión} = \frac{1}{N} \sum_{i=1}^{N} 1_{[\hat{y}i = yi]}$$

2. **Sensibilidad (Recall):**

   De acorde a Guerrero (2022) "Es la proporción de clasificaciones positivas realizadas correctamente para todos los datos en la clase real" (p. 51). Lo que demuestra que para cada clase j:

$$\text{Sensibilidad}_j = \frac{TP_j}{TP_j + FN_j}$$

   Donde TP_j es el número de verdaderos positivos para la clase j y FN_j es el número de falsos negativos para la clase j.

3. **Especificidad**:

   Según Guerrero (2022) "Es una medida que indica el número de clasificaciones negativas que fueron clasificadas correctamente como negativas" (p. 51). Lo que demuestra que para cada clase j:

$$\text{Especificidad}_j = \frac{TN_j}{TN_j + FP_j}$$

   Donde TN_j es el número de verdaderos negativos para la clase j y FP_j es el número de falsos positivos para la clase j.

4. **Puntaje F1 (F1-score)**:

   Conforme a Rivas (2023) "F1 score que representa una forma de medición entre la presión y sensibilidad de manera similar, se opta los tres primeros añadiendo a su estudio la exactitud que representa la cantidad de aciertos correctos" (p. 101). Se refiere a una métrica combinada, para cada clase j:

$$F1_j = 2 \text{ x } \frac{\text{Precisión}_j \text{ x Sensibilidad}_j}{\text{Precisión}_j + \text{Sensibilidad}_j}$$

5. **Área bajo la curva ROC (AUC-ROC)**:

De acuerdo a Rivas (2023) “La AUC es la medida para evaluar el desempeño de los límites” (p. 101). Se refiere a que la curva ROC (Receiver Operating Characteristic) se traza representando la tasa de verdaderos positivos (TPR o Sensibilidad) frente a la tasa de falsos positivos (FPR) a diferentes umbrales de decisión.

Definiciones:

- **Verdaderos Positivos (TP):** Número de instancias correctamente clasificadas como la clase positiva.
- **Falsos Positivos (FP):** Número de instancias incorrectamente clasificadas como la clase positiva.
- **Verdaderos Negativos (TN):** Número de instancias correctamente clasificadas como la clase negativa.
- **Falsos Negativos (FN):** Número de instancias incorrectamente clasificadas como la clase negativa.

Tasa de Verdaderos Positivos (Sensibilidad, TPR):

$$\text{TPR} = \frac{TP}{TP + FN}$$

Tasa de Falsos Positivos (FPR):

$$\text{FPR} = \frac{FP}{FP + TN}$$

Procedimiento

Se convierte el Problema Multiclase en Problemas Binarios, para cada clase j, se realiza un etiquetado binario donde la clase j es etiquetada como positiva y todas las demás clases como negativas. Se calculan TPR y FPR a diferentes umbrales (de 0 a 1) en incrementos pequeños, para determinar cómo clasificar las instancias de cada clase.

$$P(y = j|x; \theta)$$

Para cada umbral, se calculan la Tasa de Verdaderos Positivos (TPR) y la Tasa de Falsos Positivos (FPR). La TPR, también conocida como Sensibilidad, es la proporción de casos positivos correctamente identificados por el modelo entre todos los casos que son realmente positivos. La FPR, en cambio, es la proporción de casos negativos incorrectamente clasificados como positivos por el modelo entre todos los casos realmente negativos. Luego, se traza la curva ROC representando la TPR en el eje y y la FPR en el eje x. Para encontrar el Área Bajo la Curva (AUC) para cada clase j, se utiliza el método del trapecio, calculando el área bajo la curva ROC para obtener el AUC correspondiente a esa clase.

Validación Cruzada

Dependiendo de Rivas (2023) “La segunda/última etapa utiliza estos clasificadores y utiliza la validación cruzada estrategia para asegurar la comparación de tarifas entre clasificadores” (p. 27). Se refiere a que La validación cruzada es una técnica que se emplea para estimar cómo se comportará un modelo en un conjunto de datos independiente, dividiendo los datos en subconjuntos de entrenamiento y prueba repetidamente. A continuación, se define los pasos del procedimiento de k-pliegues.

Procedimiento de k-pliegues:

1. Se divide el conjunto de datos en k pliegues.
2. Se entrena el modelo en k−1 pliegues, validando en el pliegue restante.
3. Se repite el proceso k veces, cada vez con un pliegue diferente como conjunto de validación.
4. Se promedia las métricas obtenidas en cada pliegue para obtener una estimación robusta del desempeño del modelo.

Desarrollo del modelo de ML para la clasificación de imágenes médicas del cáncer pulmonar

En el desarrollo del modelo de ML para la clasificación de imágenes médicas del cáncer pulmonar, se ha realizado en Google Colab, utilizando TensorFlow y Keras, integrando bibliotecas y módulos de Python esenciales para el procesamiento de imágenes; el entrenamiento, la evaluación y la visualización de resultados del modelo SVM, incluyendo gráficos para la evaluación del modelo y la exactitud es de 99%, utilizando el dataset "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD". El código está optimizado y estructurado para una mejor legibilidad, e incluye la funcionalidad para guardar y cargar el modelo con las bibliotecas de HDF5 y Joblib.

Estas son las herramientas que se han implementado:

- **Numpy:** Para la realización de arrays y operaciones numéricas.
- **Pandas:** Se ha implementado para la manipulación y análisis de datos.
- **Matplotlib:** Para la creación de gráficos y visualizaciones.
- **OpenCV:** Para el procesamiento de imágenes.
- **OS:** Para la manipulación de directorios y archivos.
- **Sklearn (Scikit-learn):** Para el modelado, entrenamiento, evaluación del modelo de SVM, y procesamiento de datos.
- **Seaborn:** Para la visualización de la matriz de confusión.
- **Scikit-image:** Para la extracción de características HOG (Histogram of Oriented Gradients) de las imágenes.

Pasos a seguir:

1. **Carga y Preprocesamiento de Datos:** Se ha cargado las imágenes según las 3 categorías, utilizando OpenCV para el procesamiento de imágenes. Se ha redimensionado el tamaño y convertido a la escala de grises.

```
Categorías de las imágenes: ['Malignant cases', 'Bengin cases', 'Normal cases']
```

Figura 5. *Categorías de las imágenes*

Figura 6. *Casos de las imágenes*

```
Número de archivos en Malignant cases: 561
Número de archivos en Bengin cases: 120
Número de archivos en Normal cases: 416
```

Figura 7. *Número de los casos*

```
Número de muestras totales: 1097
```

Figura 8. *Número total de los casos*

2. **Extracción de Características:** Se ha utilizado el Histograma de Gradientes Orientados (HOG) para clasificar imágenes según la forma y estructura de los objetos.

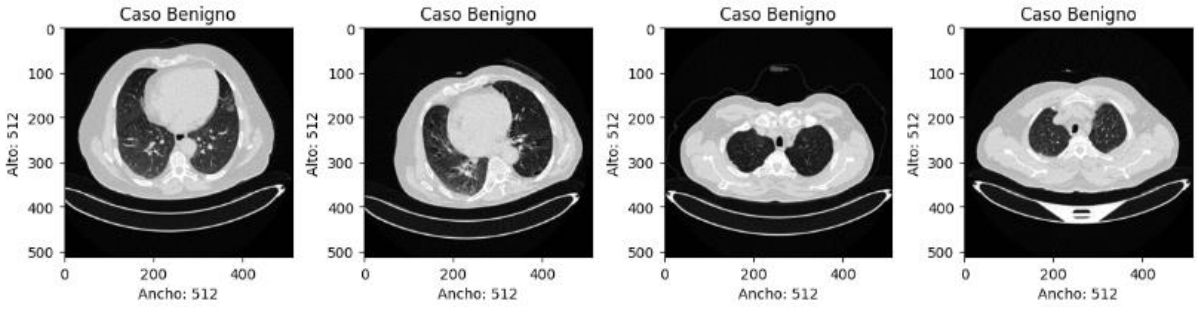

Figura 9. *Casos Benignos*

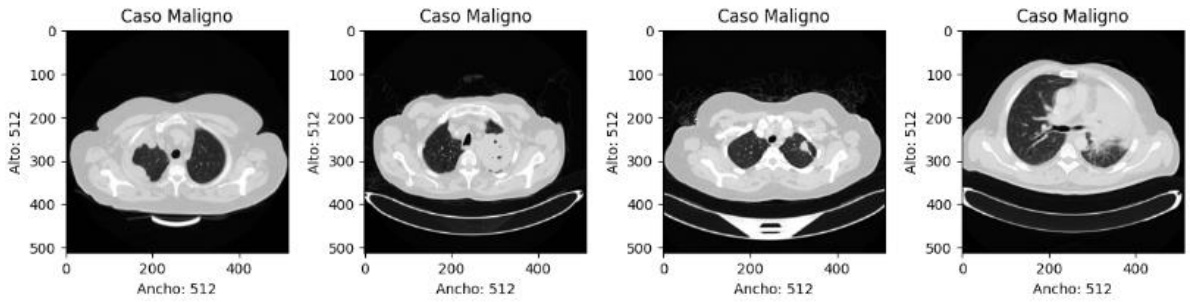

Figura 10. *Casos Malignos*

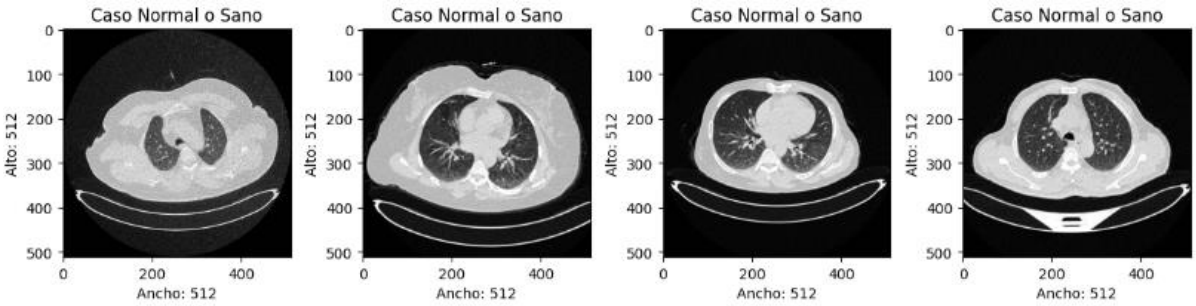

Figura 11. *Casos Normales o Sanos*

3. **División de Datos:** Se ha divido los datos, el 80% de entrenamiento y el 20% de prueba. Las imágenes del entrenamiento son 877 y de la prueba 220.

```
Número de muestras de entrenamiento: 877
Número de muestras de prueba: 220
------------------------------------
```

Figura 12. *División de Datos*

4. **Entrenamiento del Modelo SVM:** Se ha entrenado un modelo SVM con un kernel lineal, utilizando los datos y evaluando el tiempo de entrenamiento.

```
Tiempo de entrenamiento: 12.889176607131958 segundos
```

Figura 13. *Tiempo de entrenamiento*

5. **Evaluación del Modelo:** Se ha evaluado el modelo SVM utilizando métricas como la matriz de confusión, el informe de clasificación, la precisión del modelo, la sensibilidad (recall), la especificidad, el puntaje F1, AUC-ROC.

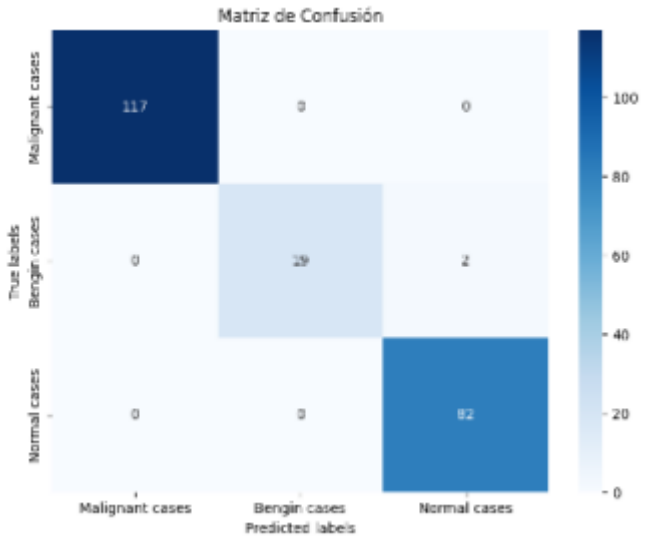

Figura 14. *Matriz de confusión*

```
Informe de clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       117
           1       1.00      0.90      0.95        21
           2       0.98      1.00      0.99        82

    accuracy                           0.99       220
   macro avg       0.99      0.97      0.98       220
weighted avg       0.99      0.99      0.99       220

Precisión del modelo: 0.990909090909091
```

Figura 15. *El informe de clasificación y la precisión del modelo*

```
Sensibilidad (Recall) por clase: [1.0, 0.9047619047619048, 1.0]
Especificidad por clase: [1.0, 1.0, 0.9855072463768116]
Sensibilidad promedio: 0.9682539682539683
Especificidad promedio: 0.9951690821256038
```

Figura 16. *La sensibilidad (recall), la especificidad*

```
Puntaje F1 por clase: [1.0, 0.9500000000000001, 0.9927007299270074]
```

Figura 17. *Puntaje F1 por clase*

```
Puntaje F1 promedio: 0.9809002433090025
```

Figura 18. *Puntaje F1 promedio*

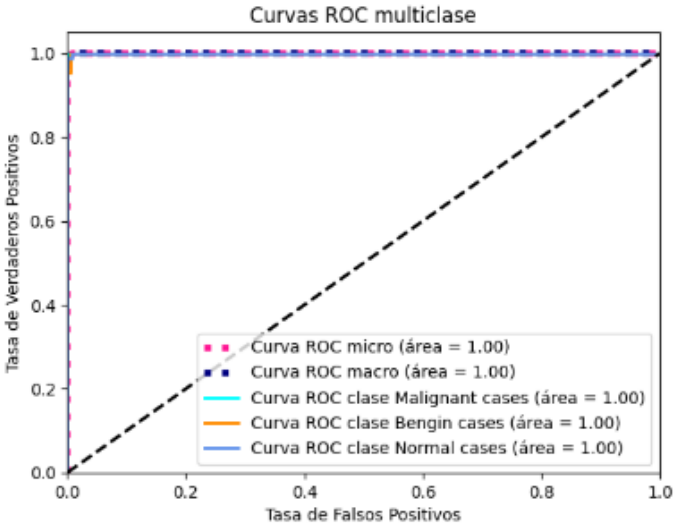

Figura 19. *Curva AUC-ROC*

6. **Validación Cruzada:** Se ha realizado Validación Cruzada y la Historia de la Precisión del Modelo SVM para evaluar la robustez del modelo SVM.

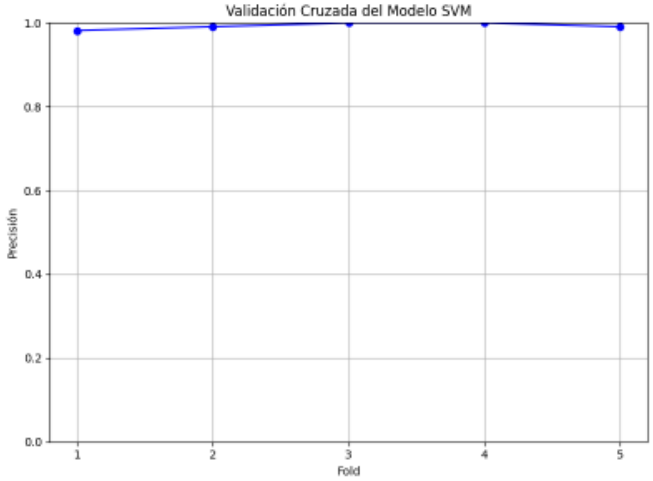

Figura 20. *Validación Cruzada*

```
Promedio de precisión de la validación cruzada: 0.9927189705271896
```

Figura 21. *Promedio de precisión de la validación cruzada*

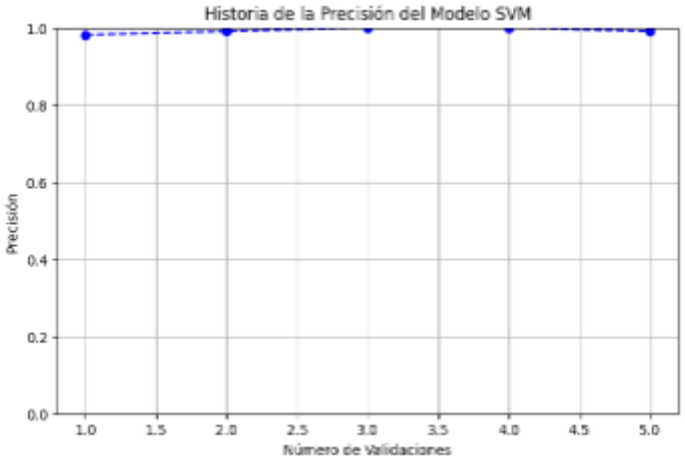

Figura 22. *Historia de la Precisión del Modelo SVM*

7. **Visualización y Análisis:** Se ha visualizado resultados importantes como el AUC-ROC por clase y el promedio.

```
AUC-ROC por clase:
Clase 0: 1.0
Clase 1: 0.9523809523809523
Clase 2: 0.9927536231884059
```

Figura 23. *AUC-ROC por clase*

```
AUC-ROC promedio: 0.9998907792848334
```

Figura 24. *AUC-ROC promedio*

8. **Guardado del Modelo y Métricas:** Finalmente, se ha guardado el modelo SVM entrenado y las métricas en un archivo HDF5 (svm_model.h5) utilizando la biblioteca h5py.

```
['metrics', 'svm_model']
Coeficientes: [[ 2.53826521e-02  3.00102161e-02  3.88911109e-02 ...  3.11460946e-04
  -1.04728142e-05  2.83380043e-04]
 [ 2.94395149e-02  2.15224335e-02  3.83107196e-02 ...  2.69806419e-04
  -2.18937785e-04 -7.91588515e-04]
 [-8.85173426e-03 -1.92519883e-02 -2.01201356e-02 ... -8.90382755e-04
  -1.83883900e-03 -3.15024499e-03]]
Intercepto: [ 2.93106871  3.0737139  -0.65840195]
Precisión: 0.990909090909091
Sensibilidad promedio: 0.9682539682539683
Especificidad promedio: 0.9951690821256038
AUC-ROC promedio: 0.9998907792848334
Tiempo de entrenamiento: 12.889176607131958 segundos
Descripción: Modelo SVM entrenado para clasificación de imágenes de cáncer de pulmón
Modelo y métricas guardadas en svm_model.h5
```

Figura 25. *svm_model.h5*

9. **Guardado del Modelo y Métricas:** Finalmente, se ha guardado el modelo SVM entrenado y las métricas en un archivo PKL (svm_cancer_pulmon_model.pkl), utilizando la biblioteca Joblib.

```
Modelo y métricas guardadas en svm_cancer_pulmon_model.pkl y svm_cancer_pulmon_metrics.pkl
```

Figura 26. *Modelo y métricas guardadas*

```
Modelo cargado:
SVC(kernel='linear', probability=True, random_state=42)
Métricas cargadas:
accuracy: 0.990909090909091
sensitivity: 1.0
specificity: 0.9855072463768116
f1_score: [1.0, 0.9500000000000001, 0.9927007299270074]
auc_roc: 0.9998907792848334
training_time: 12.889176607131958
```

Figura 27. *Modelo y métricas cargadas*

Figura 28. *Resultados obtenidos*

La exactitud del modelo es de 99%, significa que ha clasificado correctamente para 220 imágenes en total.

El modelo con la extensión .ipynb, desarrollado en Google Colab, junto con la licencia, la descripción del proyecto detallada en el archivo README.md y los archivos del modelo y las métricas entrenadas, se ha almacenado en el repositorio de GitHub titulado Modelo-SVM-para-la-clasificacion-de-imagenes-del-cancer-del-pulmon.

Figura 29. *Repositorio de GitHub*

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

Tomando en cuenta lo mencionado por Alba y Córdoba (2004), “Las conclusiones se deducen del análisis de los resultados de la investigación y contribuyen al conocimiento especializado”. Esto quiere decir que las conclusiones son el producto final de un proceso de investigación meticuloso. Por su parte Rubio, Rivera, Murillo, Gómez, & Ramírez (2021) refieren que “Las recomendaciones incluyen temas que, normalmente, están ligados a las conclusiones; el investigador condensa aquellas sugerencias que se originaron durante el proceso de realización del estudio” (p. 9), con lo cual se puede entender que las recomendaciones están directamente vinculadas a los hallazgos y conclusiones del estudio, proporcionando orientación práctica basada en los resultados obtenidos.

Conclusiones

A pesar de las dificultades comunes en el desarrollo del modelo de Machine Learning (ML) para clasificar imágenes médicas del cáncer pulmonar, como la calidad y cantidad de datos, el preprocesamiento de imágenes y la optimización de parámetros, el modelo de Máquina de Vectores de Soporte (SVM) ha demostrado ser altamente efectivo. Usando el dataset "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD", este modelo clasifica con precisión imágenes de rayos X en benigno, maligno y normal. Las métricas de rendimiento, incluyendo precisión, sensibilidad, especificidad, puntaje F1 y AUC-ROC, destacan su capacidad para detectar patologías pulmonares con exactitud. La validación cruzada refuerza su robustez, la alta interpretabilidad garantiza resultados claros y útiles para radiólogos y médicos, mejorando significativamente los resultados para los pacientes.

Como consecuencia, el modelo de ML elegido para la clasificación de imágenes radiológicas del cáncer pulmonar fue la SVM. Esta decisión se fundamentó en la

capacidad del SVM para gestionar datos de alta dimensionalidad y su eficacia en la clasificación tanto binaria como multiclase, lo que es crucial para diferenciar entre imágenes benignas, malignas y normales. El modelo SVM fue sometido a un exhaustivo proceso del 20% prueba y 80% entrenamiento, utilizando el dataset "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD", que incluye una extensa variedad de casos. Además, se realizó una validación cruzada para garantizar su robustez y capacidad de generalización, logrando resultados sobresalientes en precisión, sensibilidad, especificidad, AUC-ROC y puntaje F1.

En este sentido, se considera que en el análisis del dataset de imágenes clínicas para el cáncer de pulmón se realizó para garantizar la calidad de los datos utilizados en el entrenamiento del modelo de ML. El dataset "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD", que se utilizó consta de 1.097 imágenes de tomografías computarizadas, fue sometido a un exhaustivo proceso de preprocesamiento, que incluyó normalización, eliminación de duplicados y corrección de imágenes, con un enfoque riguroso. Se aplicaron técnicas avanzadas de aumento de datos y segmentación para fortalecer el modelo. La adecuada organización en categorías y la validación cruzada garantizan que el modelo SVM se entrene con datos precisos, lo que contribuye a un alto desempeño en la clasificación y asegura resultados fiables en la detección de patologías pulmonares. La calidad del dataset es fundamental para la precisión, sensibilidad y especificidad del modelo, asegurando resultados robustos.

De esta forma, se ha desarrollado una abstracción matemática para el modelo de ML, que resulta fundamental para comprender y evaluar su funcionamiento. En la clasificación de imágenes del cáncer pulmonar, se utiliza un modelo de SVM, que define una función de decisión mediante un hiperplano que separa las clases en el espacio de características, maximizando el margen entre ellas. Esta estructura facilita la evaluación del rendimiento del modelo mediante métricas como precisión, sensibilidad, especificidad, puntaje F1 y AUC-ROC, obtenidas a partir de la matriz de confusión y la validación cruzada. La abstracción matemática proporciona una base sólida para definir, representar y optimizar el modelo, garantizando una clasificación precisa de los tipos de cáncer pulmonar y casos sin cáncer pulmonar, mejorando así el diagnóstico médico.

Finalmente, se desarrolló el modelo de ML para clasificar imágenes médicas de cáncer pulmonar y representa un avance significativo en el diagnóstico de esta enfermedad. Se ha empleado un modelo de SVM debido a la eficacia para manejar y clasificar grandes volúmenes de datos, este proceso abarcó la recolección y preprocesamiento de datos, selección de características y ajuste del modelo, el cual fue evaluado utilizando métricas clave como precisión, sensibilidad, especificidad, puntaje F1, AUC-ROC y validación cruzada. A pesar de los desafíos relacionados con la calidad y cantidad de datos, el modelo SVM ha demostrado ser altamente efectivo en la detección precisa de patologías pulmonares, proporcionando resultados claros y útiles que optimizan los diagnósticos médicos.

Recomendaciones

Para asegurar la efectividad del modelo SVM, es necesario realizar una verificación completa en un entorno de prueba, empleando el dataset "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD". Este procedimiento permitirá evaluar la capacidad de generalización y la robustez del modelo, y generará un informe detallado que incluirá métricas de rendimiento tales como precisión, sensibilidad, especificidad, puntaje F1 y AUC-ROC. La validación debe ser realizada por el equipo de desarrollo del modelo junto con expertos en análisis de datos y debe completarse en un plazo de 1 a 3 meses. Este enfoque garantizará que el modelo sea sólido antes de su aplicación en entornos reales, beneficiando a investigadores y desarrolladores al proporcionar una base sólida para futuras aplicaciones y mejorando el proceso de desarrollo de modelos de ML.

Para realizar una revisión se debe detallar los resultados del modelo SVM y documentar estos hallazgos en un informe final. Este informe debe incluir un análisis de los resultados obtenidos a partir de la validación cruzada, comparar el rendimiento con otros métodos si es necesario, e incorporar métricas de rendimiento como precisión, sensibilidad, especificidad, AUC-ROC, puntaje F1; la interpretabilidad y el tiempo de entrenamiento. La revisión y documentación se realizarán en las instalaciones del equipo de desarrollo, sin necesidad de ajustes finales en un entorno de laboratorio, y deben completarse en un plazo de 1 a 5 semanas. Esto garantizará

que el modelo esté optimizado y listo para su uso en aplicaciones reales, beneficiando a investigadores y desarrolladores con un análisis detallado y a profesionales clínicos con un modelo validado y bien documentado.

Finalmente, es esencial realizar una revisión detallada del proceso de preprocesamiento y las técnicas aplicadas al dataset "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD". Esta revisión debe evaluar la eficacia de las técnicas de normalización, eliminación de duplicados, corrección de imágenes corruptas, aumento de datos y segmentación. Además, es necesario realizar la validación en las categorías con datos precisos y confiables. y la evaluación de los resultados de la validación cruzada para garantizar que el modelo SVM. El equipo de desarrollo del modelo y los expertos en análisis de datos y procesamiento de imágenes, llevará a cabo esta tarea y documentará los hallazgos en un informe final, completando el proceso en un plazo de 1 a 5 semanas. Esto proporcionará una base sólida para futuras investigaciones y asegurará un modelo robusto para una detección efectiva de patologías pulmonares.

REFERENCIAS BIBLIOGRÁFICAS

Alba, P. E. G., & Córdoba, B. R. (2004). Metodología de la investigación. Nueva Imagen.

Abelaira, C., Ruano-Ravina, A., & Fernández-Villar, A. (2020). Artificial Intelligence in Thoracic Radiology. A Challenge in COVID-19 Times? *Archivos de bronconeumologia*, *57*, 15-16.

Altuna Pérez, S. (2023). MODELO PREDICTIVO DE LA RESPUESTA AL TRATAMIENTO DE QUIMIO-RADIOTERAPIA EN CÁNCER DE PULMÓN DE CÉLULAS NO PEQUEÑAS LOCALMENTE AVANZADOS BASADO EN VARIABLES RADIÓMICAS DE CT.

Arias, F. (2012). *Proyecto de Investigación. Introducción a la metodología de la científica.* Caracas: Episteme.

Arias, F. (2016). Proyecto de Investigación. Introducción a la metodología de la científica. Caracas : Episteme.

Azuero, Á. E. A. (2019). Significatividad del marco metodológico en el desarrollo de proyectos de investigación. Revista arbitrada interdisciplinaria Koinonía, 4(8), 110-127.

Bernate, J. A., & Fonseca, I. P. (2023). Impacto de las Tecnologías de Información y Comunicación en la educación del siglo XXI: Revisión bibliométrica. *Revista de ciencias sociales*, *29*(1), 227-242.

Bernavent, D., Colomer, J., Quecedo, L., Gol-Monserrat, J., Llano Señarís, J. (2024). Inteligencia Artificial y Desiciones Clínicas. *Recuperado el 16 de enero de 2024, de* https://www.dropbox.com/sh/d6plg2q4oydvpcs/AADGtlJV1bJtG1NJF2hl6pXFa/Archivos%20Infolibros%20ES/Temas/390%20Inteligencia%20Artificial/20.%20Inteligencia%20Artificial%20y%20Decisiones%20Cl%C3%ADnicas%20autor%20Varios%20autores.pdf?dl=0

Calderón Fornaris, P. A., & Piñeiro Suárez, N. (2003). Metodología de la Investigación Científica. *Selección de lecturas. Recuperado el 23 de mayo de 2024, de*

https://gc.scalahed.com/recursos/files/r161r/w24908w/S2/metodologia_investigacion_cientifica_lecturas.pdf

Castillo Brito, Y., & Herrera Roldan, G.M. (2017). Evaluación de conocimientos mediados por la tecnología (E-EVALUACIÓN) en instituciones de educación superior. INCYT-UNIBE ISBN:978-9942-8586-4-1.

Cerón, M. C., & Cerâon, M. C. (2006). *Metodologías de la investigación social*. Santiago: LOM ediciones.

Cortes, A. (2019). *Una nueva tecnología pretende transformar el cáncer de pulmón en una enfermedad crónica*. Ediciones EL PAÍS S.L. https://elpais.com/elpais/2019/11/08/ciencia/1573214337_571170.html

Ecancer (2023). Desarrollan una herramienta de inteligencia artificial para predecir el riesgo de cáncer de pulmón. Recuperado el 27 de abr. de 2024 de https://ecancer.org/es/news/22569-desarrollan-una-herramienta-de-inteligencia-artificial-para-predecir-el-riesgo-de-cancer-de-pulmon

El estado de la Legaltech en… Ecuador. (2020, junio 25). Legaltechies | Consultoría y asesoramiento en el estudio e implementación de tecnología en materia legal; Legaltechies. https://legaltechies.es/2020/06/25/el-estado-de-la-legaltech-en-ecuador/

Enfermedades respiratorias y/o pulmonares ocasionadas y derivadas por el Covid19 (Bachelor's thesis, Universidad de Guayaquil. Facultad de Ciencias Matemáticas y Físicas. Carrera de Ingeniería en Sistemas Computacionales.).

Esquivel, D. (s/f). Machine learning vs. Deep learning - CDA informática. *Recuperado el 18 de enero de 2024, de* https://www.cdainfo.com/es/noticias/148-machine-learning-vs-deep-learning

Fos Guarinos, B. (2016). *Diseño de técnicas de inteligencia artificial aplicadas a imágenes médicas de rayos X para la detección de estructuras anatómicas de los*

pulmones y sus alteraciones (Doctoral dissertation, Universitat Politècnica de València).

Freire, A. X., Benítez, S., Briones, K., & Freire, N. V. (2003). Duración de la valoración diagnóstica del cáncer de pulmón frente a otros tumores sólidos en el Instituto Oncológico Nacional de Ecuador. Archivos de Bronconeumología, 39(4), 167-170.

Freire, C. E. E. E. (2018). El problema de investigación. Revista Conrado, 14(64), 22-32.

Fuentes, A. S. F., Cabrera, R. G., & Rodriguez, E. V. (2022). Identificación automática de cáncer de piel aplicando machine learning. REVISTA COLOMBIANA DE TECNOLOGIAS DE AVANZADA (RCTA), 2(40), 1-6.

García A. (2003). EL CUESTIONARIO COMO INSTRUMENTO DE INVESTIGACIÓN/EVALUACIÓN. Etapas del Proceso Investigador: INSTRUMENTACIÓN. Recuperado el 28 de agosto de 2024, de http://www.etpcba.com.ar/documentos/sitios/evaluacion_intitucional/8_el_cuestionario.pdf

Géron A. (s/f). Aprende Machine Learning con Scikit-Learn, Keras y TensorFlow Conceptos, herramientas y técnicas para conseguir sistemas inteligentes. (3ª ed.). Recuperado el 27 de agosto de 2024, de https://anayamultimedia.es/primer_capitulo/aprende-machine-learning-con-scikit-learn-keras-y-tensorflow-tercera-edicion.pdf

Guerrero, M. (2022). Modelo basado en deep learning para el diagnóstico de tuberculosis pulmonar utilizando radiografías de tórax y perfiles clínicos.

González, A. (2020). ¿Qué es Machine Learning? *Recuperado 27 de diciembre de 2023, de* https://cleverdata.io/que-es-machine-learning-big-data/

Gonzáles, J. L. (2020). Técnicas e instrumentos de investigación científica. Arequipa, Arequipa, Perú.

González, R. (s/f). Python PARA TODOS. Recuperado el 26 de agosto de 2024, de https://launchpadlibrarian.net/18980633/Python%20para%20todos.pdf

Grady, D. (2019). La inteligencia artificial hizo una prueba para detectar el cáncer de pulmón… y aprobó con honores. The New York times. https://www.nytimes.com/es/2019/05/22/espanol/inteligencia-artificial-cancer-pulmon.html

Hernández, S. R., & Mendoza, C. P. (2018). Metodología de la investigación. Las rutas cuantitativas, cualitativas y mixta. México: Mc Graw Hill Education.

IBM. (2023). AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the difference? Recuperado 16 de enero de 2023, de https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/

Instituto de Ingeniería del Conocimiento (2020). *Machine Learning y Deep Learning* - Expertos en IIC. *Recuperado 16 de enero de 2023, de* https://www.iic.uam.es/inteligencia-artificial/machine-learning-deep-learning/

Iriarte Erick (2023). *¿Leyes para la Inteligencia Artificial?.* Recuperado 27 de diciembre de 2023, de https://www.lahora.com.ec/editorial/columnistas-nacionales/leyes-para-la-inteligencia-artificial/

Igual R., & Medrano, C. (2023) Tutorial de OpenCV. Recuperado 27 de agosto de 2023, de file:///C:/Users/reaf/Downloads/Tutorial_de_OpenCV.pdf

Julca Villarreal, B. F. (2020). Aplicación de Deep Learning sobre imágenes topográficas para mejorar la precisión del diagnóstico de queratocono en una clínica de Lima.

KeepCoding (2023). ¿Qué son los datasets? *Recuperado 16 de enero de 2023, de* https://keepcoding.io/blog/que-son-datasets/

Lanuez, M. y Fernández, E. (2014). Metodología de la Investigación Educativa. (CDROM). IPLAC, La Habana, Cuba

Leivi, A. E. (2019). Análisis de la implementación de Machine Learning en el diagnóstico por imágenes.

López, A. R. (2023). ¿Qué es Google Colab y cómo usarlo?: La guía definitiva. Recuperado el 27 de agosto de 2024, de https://seoalex.es/blog/que-es-google-colab-y-como-usarlo/

Lopez Tumbaco, O. P., & Terranova Pihuave, J. B. (2023) Técnicas de machine learning basadas en aprendizaje supervisado para la predicción de Enfermedades respiratorias y/o pulmonares ocasionadas y derivadas por el Covid19 (Bachelor's thesis, Universidad de Guayaquil. Facultad de Ciencias Matemáticas y Físicas. Carrera de Ingeniería en Sistemas Computacionales.).

Martín, S. R., Manjarrés, S. M., & Martín, S. R. (2019). Aspectos metodológicos de la instrumentalización de la recogida de datos primarios y sus consideraciones éticas en la investigación clínica. Enfermería en cardiología: revista científica e informativa de la Asociación Española de Enfermería en Cardiología, (76), 21-26.

Martins, J. (2024, enero 19). *¿Qué es la metodología Kanban y cómo funciona?* Asana. Recuperado el 23 de mayo de 2024, de https://asana.com/es/resources/what-is-kanban

Paitán, H. Ñ., Mejía, E. M., Ramírez, E. N., & Paucar, A. V. (2014). Metodología de la investigación cuantitativa-cualitativa y redacción de la tesis. Ediciones de la U.

Parra, F. (2019). *Estadística y Machine Learning con R*. Recuperado el 5 de julio de 2024, de https://bookdown.org/content/2274/portada.html

Parsons, C. (2021, septiembre 28). ¿Qué Es un Modelo de Machine Learning? Blog oficial de NVIDIA Latino América. *Recuperado el 16 mayo del 2024, de* https://la.blogs.nvidia.com/blog/que-es-un-modelo-de-machine-learning/

Pineda, J. M. (2022). Modelos predictivos en salud basados en aprendizaje de maquina (machine learning). Revista Médica Clínica Las Condes, 33(6), 583-590.

Prieto González, L. S. (2022). Análisis de modelos de difusión por imágenes de resonancia magnética nuclear con machine learning (Doctoral dissertation, Universidad Nacional de Colombia).

Rivas Plata Casas, C. G. (2023). Clasificación de cáncer de pulmón en imágenes de tomografías mediante procesamiento de imágenes y aprendizaje automático.

Roche (2024). IA para el cáncer de pulmón. *Recuperado el 27 de abr. de 2024 de* https://www.rocheplus.es/innovacion/inteligencia-artificial/ia-para-cancer-pulmon.html

Rodríguez, F. A. R., Flores, L. G., & Vitón-Castillo, A. A. (2022, September). Artificial intelligence and machine learning: present and future applications in health sciences. In *Seminars in Medical Writing and Education* (Vol. 1, pp. 9-9).

Rodríguez Pérez, D. J. (2020). Herramienta para analizar matrices de expresión génicas con machine learning.

Rubio, D. B., Rivera, P. E. C., Murillo, P. G. G., Gómez, G. G., & Ramírez, A. J. P. (2021). Sugerencias para escribir análisis de resultados, conclusiones y recomendaciones en tesis y trabajos de grado. CITAS: Ciencia, innovación, tecnología, ambiente y sociedad, 7(1), 1.

Russell, S. J., & Norvig, P. (2016). Artificial intelligence: *A modern approach* (4th ed.). Pearson Education.

Salvat Navarro, A. (2023). *Aplicaciones del Machine Learning en el diagnóstico del cáncer de pulmón* (Bachelor's thesis, Universitat Politècnica de Catalunya).

Sarmiento-Ramos, J. L. (2020). Aplicaciones de las redes neuronales y el deep learning a la ingeniería biomédica. *Revista UIS Ingenierías*, 19(4), 1-18.

Solano, A. G. (2021). The IQ-OTHNCCD Lung Cancer Dataset. Recuperado 27 de diciembre de 2023, de https://www.kaggle.com/datasets/antonixx/the-iqothnccd-lung-cancer-dataset

Solis, D. C. (2023, mayo 1). Datasets: Qué son y cómo acceder a ellos. *Recuperado 27 de diciembre de 2023, de* https://openwebinars.net/blog/datasets-que-son-y-como-acceder-a-ellos/

Toquero Barón, M. (2021). Clasificación de imágenes médicas de Rayos-X mediante redes neuronales convolucionales.

Tuarez Vega, R. J., & Vera Pizanan, R. N. (2022). Desarrollo de software biomédico mediante modelos deep learning para la detección de tumores pulmonares en la aplicación de procesamiento de imágenes espectrales para el departamento médico de la Universidad Técnica de Cotopaxi Extensión La Maná (Bachelor's thesis, Ecuador: La Mana: Universidad Técnica de Cotopaxi (UTC)).

Valdés, S. A., Intriago, C. A. H., & Felipe, M. D. R. C. (2022). Predicción de las principales enfermedades que afectan la salud en Ecuador a partir de factores de riesgo. Serie Científica de la Universidad de las Ciencias Informáticas, 15(8), 37-50.

Vargas, M., Biggs, D., Larraín, T., Alvear, A., Pedemonte, J. C., & de Anestesiología, R. (2022). Inteligencia artificial en medicina: Métodos de modelamiento (Parte I). Revista Chilena de Anestesia, 51(5), 527-534.

Zamora-de-Ortiz, M. S., Serrano-Pastor, F. J., & Martínez-Segura, M. (2020). Validez de contenido del modelo didáctico P-VIRC (preguntar-ver, interpretar, recorrer, contar) mediante el juicio de expertos. Formación universitaria, 13(3), 43-54.

ANEXOS

ANEXO 1. ENCUESTA: EVALUACIÓN DE MODELOS DE MACHINE LEARNING PARA LA CLASIFICACIÓN DE IMÁGENES MÉDICAS DEL CÁNCER PULMONAR

Objetivo del Instrumento: Evaluar la selección, implementación, análisis y desempeño de diversos modelos de Machine Learning (ML) en la clasificación de imágenes radiológicas del cáncer pulmonar.

Instrucciones: Por favor, revise cada uno de los ítems enunciados a continuación y marque con una “X”, en la casilla de "Sí" o "No" según corresponda.

Tabla 5. *Lista de chequeo para la validación del modelo de ML*

#	ITEMS O ENUNCIADO	SI	NO
1	¿Cómo se compara la precisión del SVM con la del Random Forest? • Precisión del SVM: 0.88 • Precisión del Random Forest: 0.92	X	
2	¿Cuál es la diferencia en tiempo de entrenamiento entre el modelo CNN y SVM? • Tiempo de Entrenamiento del CNN: 4 horas • Tiempo de Entrenamiento del SVM: 1.5 horas	X	

3	¿Cuál es el modelo con el puntaje F1 más alto? • Puntaje F1 del CNN: 0.93	X	
4	¿Qué modelo es más fácil de interpretar? • Interpretabilidad del SVM: Alta	X	
5	¿Cómo se comparan los puntajes F1 de SVM y K-NN? • Puntaje F1 del SVM: 0.87 • Puntaje F1 del K-NN: 0.85	X	
6	¿Qué modelo requiere el menor tiempo de entrenamiento? • Tiempo de Entrenamiento del K-NN: 1 hora	X	

ANEXO 2. JUICIO DE EXPERTO

INSTRUCCIONES:

Coloque una “X” en la casilla correspondiente según su apreciación a los ítems y a la elección de la solución, conforme las pautas que se detallan a continuación:

Tabla 6. *Criterios del Juicio de Experto*

Ítems	Claridad en la redacción		Coherencia interna		Inducción a la respuesta (Sesgo)		Lenguaje adecuado la población		Mide lo que pretende		Valoración			Observaciones
	Si	No	Si	No	Si	No	Si	No	Si	No	Esencial	Útil pero no esencial	No importante	
1	X		X			X	X		X		X			
2	X		X			X	X		X		X			

3	X		X			X	X		X		X			
4	X		X			X	X		X		X			
5	X		X			X	X		X		X			
6	X		X			X	X		X		X			

Apreciación cualitativa:

Observaciones:

Validado por: Msc. Sandino Jaramillo

Profesión: Ingeniero en sistemas

Cargo que desempeña: Director de Carrera de Ingeniería de Software (E)

Firma:

ANEXO 3. JUICIO DE EXPERTO

INSTRUCCIONES:

Coloque una “X” en la casilla correspondiente según su apreciación a los ítems y a la elección de la solución, conforme las pautas que se detallan a continuación:

Tabla 7. *Criterios del Juicio de Experto*

Ítems	Claridad en la redacción		Coherencia interna		Inducción a la respuesta (Sesgo)		Lenguaje adecuado la población		Mide lo que pretende		Valoración			Observaciones
	Si	No	Si	No	Si	No	Si	No	Si	No	Esencial	Útil pero no esencial	No importante	
1	X		X			X	X		X		X			
2	X		X			X	X		X		X			

3	X		X			X	X		X		X			
4	X		X			X	X		X		X			
5	X		X			X	X		X		X			
6	X		X			X	X		X		X			

Apreciación cualitativa:

Observaciones:

Validado por: MSc. Carlos Antonio Ayala Tipán, (c)

Profesión: Académico/Investigador

Cargo que desempeña: Técnico de Investigación a tiempo completo

Firma:

ANEXO 4. JUICIO DE EXPERTO

INSTRUCCIONES:

Coloque una “X” en la casilla correspondiente según su apreciación a los ítems y a la elección de la solución, conforme las pautas que se detallan a continuación:

Tabla 8. *Criterios del Juicio de Experto*

Ítems	Claridad en la redacción		Coherencia interna		Inducción a la respuesta (Sesgo)		Lenguaje adecuado la población		Mide lo que pretende		Valoración			Observaciones
	Si	No	Si	No	Si	No	Si	No	Si	No	Esencial	Útil pero no esencial	No importante	
1	X		X			X	X		X		X			
2	X		X			X	X		X		X			

3	X		X			X	X		X		X			
4	X		X			X	X		X		X			
5	X		X			X	X		X		X			
6	X		X			X	X		X		X			

Apreciación cualitativa:

Observaciones:

Validado por: Msc. Miguel Ángel Fernández Marín

Profesión: Profesor

Cargo que desempeña: Profesor

Firma:

ANEXO 5. JUICIO DE EXPERTO

INSTRUCCIONES:

Coloque una “X” en la casilla correspondiente según su apreciación a los ítems y a la elección de la solución, conforme las pautas que se detallan a continuación:

Tabla 9. *Criterios del Juicio de Experto*

Ítems	Claridad en la redacción		Coherencia interna		Inducción a la respuesta (Sesgo)		Lenguaje adecuado la población		Mide lo que pretende		Valoración			Observaciones
	Si	No	Si	No	Si	No	Si	No	Si	No	Esencial	Útil pero no esencial	No importante	
1	X		X			X	X		X		X			
2	X		X			X	X		X		X			

3	X		X			X	X		X		X			
4	X		X			X		X	X		X			
5	X		X			X		X	X		X			
6	X		X			X		X	X		X			

Apreciación cualitativa:

Observaciones:

Validado por: Msc. Miguel Flores

Profesión: Doctor en Estadística e Investigación Operativa

Cargo que desempeña: Profesor Titular, Escuela Politécnica Nacional

Firma:

MIGUEL ALFONSO FLORES SANCHEZ

Firmado digitalmente por MIGUEL ALFONSO FLORES SANCHEZ
Nombre de reconocimiento (DN): c=EC, o=SECURITY DATA S.A., ou=ENTIDAD DE CERTIFICACION DE INFORMACION, serialNumber=160320192034, cn=MIGUEL ALFONSO FLORES SANCHEZ
Fecha: 2024.07.12 07:42:50 -05'00'

ANEXO 6. JUICIO DE EXPERTO

INSTRUCCIONES:

Coloque una “X” en la casilla correspondiente según su apreciación a los ítems y a la elección de la solución, conforme las pautas que se detallan a continuación:

Tabla 10. *Criterios del Juicio de Experto*

Ítems	Claridad en la redacción		Coherencia interna		Inducción a la respuesta (Sesgo)		Lenguaje adecuado la población		Mide lo que pretende		Valoración			Observaciones
	Si	No	Si	No	Si	No	Si	No	Si	No	Esencial	Útil pero no esencial	No importante	
1	X		X		X		X		X		X			
2	X		X		X		X		X		X			

3	X		X		X		X		X		X		
4	X		X		X		X		X		X		
5	X		X		X		X		X		X		
6	X		X		X		X		X		X		

Apreciación cualitativa:

Observaciones:

Validado por: PhD. Roberto Andrade

Profesión: Investigador/ Docente

Cargo que desempeña: Director de Innovación y Vinculación

Firma:

ANEXO 7. DESARROLLO DEL MODELO SVM CON LA METODOLOGÍA KANBAN

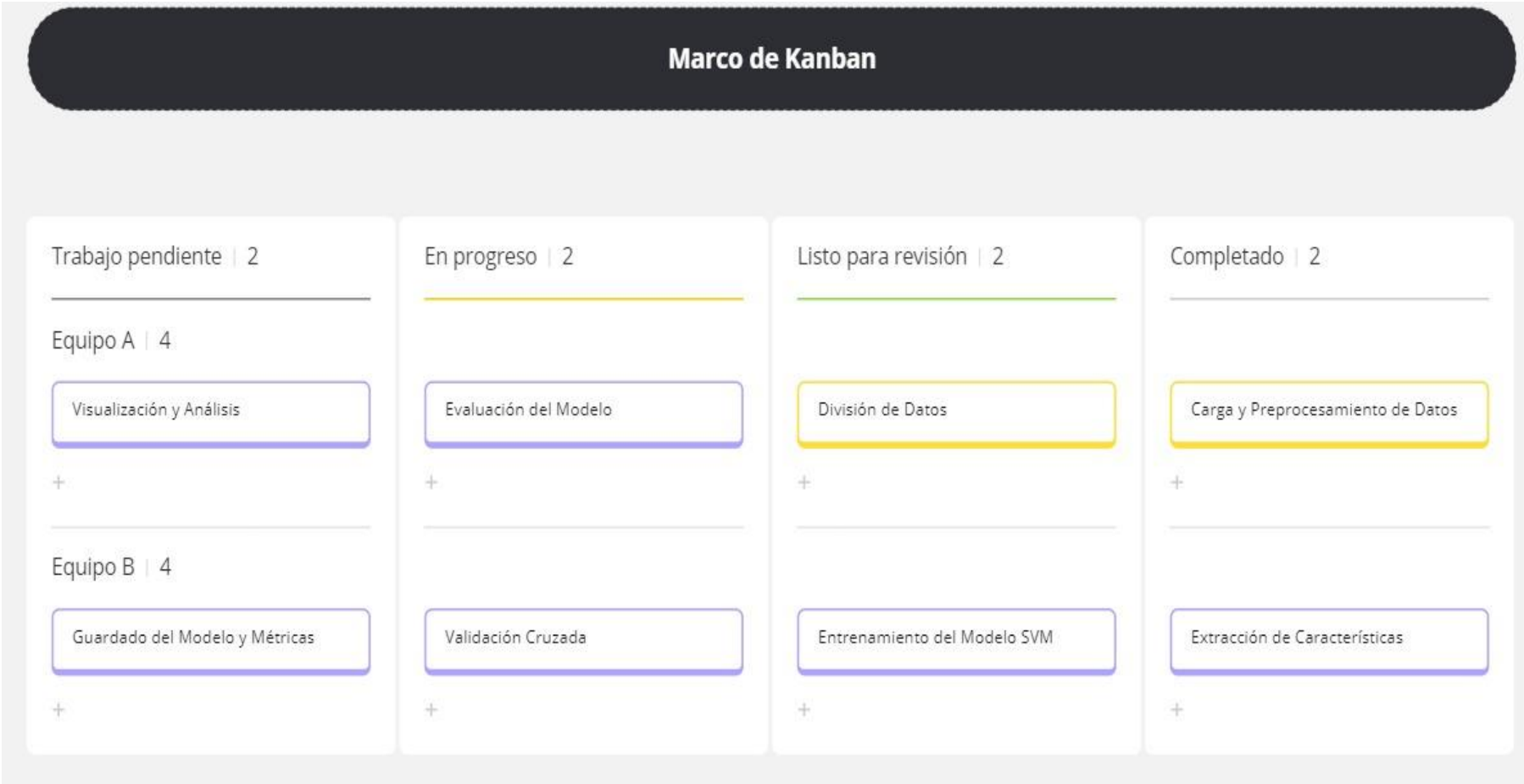

Figura 30. *Metodología Kanban para el desarrollo del modelo de SVM para la clasificación de imágenes del cáncer pulmonar*

ANEXO 8. EL LINK DEL MODELO DE MÁQUINA DE VECTORES DE SOPORTE

El modelo se encuentra guardado en el repositorio de GitHub:

https://github.com/odalis-rea-c/Modelo-SVM-para-la-clasificacion-de-imagenes-del-cancer-del-pulmon.git