MODELO BASADO

EN TÉCNICAS DE MACHINE LEARNING PARA LA CLASIFICACIÓN DE IMÁGENES MÉDICAS DEL CÁNCER PULMONAR

MODEL BASED ON MACHINE LEARNING TECHNIQUES FOR THE CLASSIFICATION OF MEDICAL IMAGES OF LUNG CANCER

Tonysé de la Rosa-Martín[1]
E-mail: severus.trm@gmail.com
ORCID: https://orcid.org/0000-0002-0881-6034
María Lucía Brito-Vallina[2]
E-mail: mbrito@umet.edu.ec
ORCID: https://orcid.org/0000-0002-1903-8514
[1] Universidad Iberoamericana. Ecuador.
[2] Universidad Metropolitana. Ecuador.

Cita sugerida (APA, séptima edición)

De la Rosa-Martín, T., & Brito-Vallina, M. L. (2025). Modelo basado en técnicas de Machine Learning para la clasificación de imágenes médicas del cáncer pulmonar. *Revista UGC, 3(2), 230-237.*

Fecha de presentación: 06/03/2025 **Fecha de aceptación:** 02/04/2025 **Fecha de publicación:** 01/05/2025

RESUMEN

Es importante destacar que este trabajo investigativo es realizar un modelo, mediante el uso de Machine Learning (ML) que pueda clasificar imágenes de Rayos X del pulmón, según los tipos del cáncer de pulmón: benigno, maligno y también imágenes normales para el entrenamiento del modelo. Se utiliza la metodología ágil Kanban, el instrumento que se analiza lo hace a través de la estadística descriptiva, utilizando las tablas de frecuencia.

Palabras clave:

Machine Learning, clasificar, Kanban, apoyo, diagnóstico.

ABSTRACT

It is important to highlight that this investigative work aims to develop a model using Machine Learning (ML) that can classify X-ray images of the lung according to the types of lung cancer: benign, malignant, and normal images for model training. The agile Kanban methodology is used, and the instrument analyzed is through descriptive statistics, which will be performed using frequency tables.

Keywords:

Machine Learning, classify, Kanban, support, diagnosis.

INTRODUCCIÓN

La inteligencia artificial 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. Por su parte, Rego et al. (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"*** (p. 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 optimizar en el uso de los recursos médicos disponibles.

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.

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, se ha realizado un análisis exhaustivo de diferentes modelos de ML con el objetivo de predecir el riesgo de contraer cuatro enfermedades: pulmonares, diabetes, cardiovasculares, cerebrovasculares. 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 precoz y el pronóstico de enfermedades, el fortalecimiento de las capacidades técnicas locales, y la consideración de la privacidad y la ética en la gestión de datos. Para este análisis, se utilizó la librería SKlearn, que permite dividir los datos en un 25% para prueba y un 75% para entrenamiento. Además, se resalta que el modelo de Bayes Naives mostró un desempeño sólido en la predicción del cáncer de pulmón (Avellán et al. 2022).

El problema de investigación es que el 70% de los diagnósticos son demasiado tardíos. Las posibles causas de este retraso en el reconocimiento del cáncer de pulmón pueden incluir la falta de acceso a pruebas de identificación oportuna, la insuficiente conciencia pública sobre los síntomas de esta enfermedad y la limitada aplicación de tecnologías avanzadas como el ML en el proceso. Por otro lado, las consecuencias de estos diagnósticos tardíos pueden manifestarse en el inicio de la enfermedad en fases avanzadas, lo que reduce las opciones de cuidado y disminuye las tasas de supervivencia. También puede resultar en una carga emocional y financiera más pesada para los pacientes y sus familias. 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.

Se plantea una solución al problema identificado en el año 2024 mediante el desarrollo de un modelo de algoritmos de ML diseñado específicamente para categorizar imágenes médicas relacionadas con el cáncer de pulmón. 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 identificación tardía de esta enfermedad, utilizando un dataset extranjero de la web "El Conjunto de Datos de Cáncer de Pulmón IQ-OTHNCCD". Ante esto se plantea como objetico: Diseñar un modelo de Machine Learning (ML) capaz de clasificar imágenes médicas del cáncer pulmonar, con la finalidad de apoyar el diagnóstico médico.

En esta investigación, se resalta que los profesionales de la salud enfrentan desafíos al detectar enfermedades pulmonares en etapas tempranas debido a la falta de tecnología avanzada de Inteligencia Artificial (IA). El 70% de los diagnósticos se realizan demasiado tarde (Cortes, 2019). Por su parte Rajpurkar et al.(2020), refieren que "no todos los hospitales utilizan ML para la clasificación de imágenes médicas para dar un apoyo al diagnóstico del cáncer de pulmón", con lo cual podemos afirmar que el proceso es demasiado lento en la contribución a los diagnósticos.

Además, Rajpurkar et al. (2020), señalan que *"no todos los hospitales utilizan ML para la clasificación de imágenes médicas para dar un apoyo al diagnóstico del cáncer de pulmón"*; lo que resulta en un proceso diagnóstico lento. Por tanto, se propone desarrollar un modelo de aprendizaje automático para categorizar las radiografías pulmonares como malignas, benignas o normales, con el fin de mejorar el apoyo al diagnóstico médico mediante la aplicación del modelo.

El objetivo principal de esta investigación es desarrollar un modelo de ML que pueda clasificar imágenes médicas asociadas con enfermedades pulmonares, con el fin de brindar apoyo al diagnóstico médico. Se adopta un enfoque positivista y cuantitativo, utilizando un diseño no experimental y descriptivo, centrado en el análisis de imágenes médicas relacionadas con enfermedades pulmonares mediante un modelo de ML (Hernández & Menodza, 2018).

La revisión documental se considera un componente esencial dentro de la metodología de investigación, por lo que *"permite establecer los pasos y acciones; este instrumento incluye protocolos de búsqueda, así como revisión de fuentes de información"* (Bernate & Fonseca, 2023). Por lo tanto, en este trabajo de titulación, se ha optado por emplear la técnica de investigación documental. En síntesis, la naturaleza de esta investigación, que se enfoca en el desarrollo de un modelo de ML utilizando la metodología Kanban para clasificar imágenes médicas relacionadas con enfermedades pulmonares, se fundamenta en un paradigma positivista. Se ha seleccionado un enfoque cuantitativo, con un diseño no experimental, y se ha aplicado un nivel descriptivo. Estas elecciones metodológicas se justifican por la necesidad de obtener resultados objetivos y cuantificables.

El problema de investigación se refiere según Espinoza (2018), *"la investigación parte de problemas, no hay investigación sin problema. Todo problema se da en un objeto, fenómeno o proceso, es decir en alguna parte de la realidad, en la que fue necesario profundizar para concretar la existencia de problemas".* El problema de investigación es que a los profesionales de la salud se les dificulta detectar el cáncer de pulmón en un período temprana, debido a la ausencia de alta tecnología de IA. El 70% de los diagnósticos son demasiado tardíos. También se encontró que Rajpurkar et al. (2020), indican que *"no todos los hospitales utilizan ML para la clasificación de imágenes médicas para dar un apoyo al diagnóstico del cáncer de pulmón"*; lo que implica que el proceso es demasiado lento en la contribución a los diagnósticos.

MATERIALES Y MÉTODOS

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. 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 y comprende un total de 1.097 imágenes. Estas varían en género, edad, nivel educativo, área de residencia y estado de vida (Al-Yasriy, 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.

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 et al., 2019).

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). 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.

RESULTADOS Y DISCUSIÓN

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.

En este apartado 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 (Ramírez Arévalo & Herrera Cubides, 2013).

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 1).

Tabla 1. 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
KNN	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 el mejor rendimiento se ha conseguido con un clasificador SVM. 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.

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. 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 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 para la construcción del modelo se hizo uso de distintas redes neuronales convolucionales pre-entrenadas, obteniendo el mejor resultado con la red DenseNet121. Finalmente se obtuvieron valores de precisión de 94% en el modelo final. 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.

Los métodos de vecinos más cercanos (KNN) suelen tener mejores resultados. 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 asegurar la calidad de los datos para el entrenamiento y la validación del modelo.

Para realizar el análisis del del dataset de imágenes clínicas del cáncer de pulmón, se presenta una matriz en la que se detalla el proceso de recopilación, organización y preprocesamiento de estas imágenes. En esta matriz se estructura en torno a los procesos implicados, la descripción y las técnicas utilizadas para mejorar la calidad del entrenamiento del modelo. Los procesos considerados incluyen: recopilación de datos, organización de datos, normalización, aumento de datos, segmentación de imágenes, división del conjunto de datos, visualización de aumento de datos. Algunas de las técnicas utilizadas son: bases de datos públicas, estructuración en directorios (Tabla 2).

Tabla 2. 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.

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. 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’.

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. 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 ∈ {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. Sea $D = \{(x_i, y_i)\}^{N}_{i=1}$ el conjunto de datos, donde T es la i - ésima imagen yi 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. El modelo de ML puede representarse como una función $f(x; \theta)$, donde 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 (F1):

   $y = f(x; \theta) = [P(y = 0|x; \theta), P(y = 1|x; \theta), P(y = 2|x; \theta)]$ (F1)

La clase predicha es la que tiene la mayor probabilidad (F2):

$$\hat{y} = \arg\max_{j} P(y = j|x; \theta) \quad \text{(F2)}$$

2. **Función de Pérdida:**

   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. Se refiere a que se utiliza la entropía cruzada categórica para medir la discrepancia entre las etiquetas verdaderas y las etiquetas predichas (F3):

$$\hat{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{2} 1_{[y_i = j]}\ log\ P(y = j|x_i; \theta) \quad \text{(F3)}$$

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

3. **Optimización:**

   Descenso del gradiente estocástico o SGD: optimizador con descenso de gradiente y momento. Puede incluirse la aceleración de Nesterov. Se refiere a que los parámetros del modelo se actualizan para minimizar la función de pérdida $L(\theta)$ utilizando un algoritmo de optimización como el descenso de gradiente estocástico (SGD) (F4):

   $\theta := \theta - \eta\ \nabla\theta\ L(\theta)$ (F4)

   Donde 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):**

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. Lo que demuestra que la precisión se calcula como: (F5)

$$\text{Precisión} = \frac{1}{N_i}\sum_{i=1}^{N} 1_{[\hat{y}i = yi]} \quad \text{(F5)}$$

2. **Sensibilidad (Recall):**

Es la proporción de clasificaciones positivas realizadas correctamente para todos los datos en la clase real. Lo que demuestra que para cada clase *j* (F6):

$$\text{Sensibilidad}_j = \frac{TP_j}{TP_j + FN_j} \quad \text{(F6)}$$

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**:

Es una medida que indica el número de clasificaciones negativas que fueron clasificadas correctamente como negativas. Lo que demuestra que para cada clase *j* (F7):

$$\text{Especificidadj} = \frac{TN_j}{TN_j + FP_j} \quad \text{(F7)}$$

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

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

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. Se refiere a una métrica combinada, para cada clase *j* (F8):

$$F1_j = 2 \text{ x } \frac{\text{Precisión}_j \text{ x Sensibilidad}_j}{\text{Precisión}_j + \text{Sensibilidad}_j} \quad \text{(F8)}$$

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

La AUC es la medida para evaluar el desempeño de los límites. 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,) (F9):

$$\text{TPR} = \frac{TP}{TP + FN} \quad \text{(F9)}$$

Tasa de Falsos Positivos () (F10):

$$\text{FPR} = \frac{FP}{FP + TN} \quad \text{(F10)}$$

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 (F11).

$P(y = j|x; \theta)$ (F11)

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 , se utiliza el método del trapecio, calculando el área bajo la curva ROC para obtener el AUC correspondiente a esa clase.

Validación Cruzada

La segunda/última etapa utiliza estos clasificadores y utiliza la validación cruzada estrategia para asegurar la comparación de tarifas entre clasificadores. 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.

CONCLUSIONES

La abstracción matemática del modelo de ML para la clasificación de imágenes médicas del cáncer pulmonar proporciona una estructura clara y precisa para resolver el problema de clasificación.

Al definir el problema, representar el modelo con sus parámetros, optimizar utilizando funciones de pérdida adecuadas y evaluar el desempeño con métricas cuantitativas y validación cruzada, se asegura que el modelo sea capaz de clasificar con precisión entre imágenes de cáncer pulmonar maligno, benigno o normales.

Esto no solo contribuye a mejorar el diagnóstico médico proporcionando resultados precisos y confiables, sino que también garantiza que el modelo sea sensible a las variaciones en los datos y específico en la identificación de diferentes tipos de cáncer pulmonar.

REFERENCIAS BIBLIOGRÁFICAS

Al-Yasriy, H. (2021). The IQ-OTHNCCD lung cancer dataset. https://www.kaggle.com/datasets/hamdallak/the-iqothnccd-lung-cancer-dataset

Avellán Valdés, S., Holguín Intriago, C. A., & Cruz Felipe, M. (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. https://publicaciones.uci.cu/index.php/serie/article/view/1096

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. https://doi.org/10.31876/rcs.v29i1.39748

Cortes, A. (2019). Una nueva tecnología pretende transformar el cáncer de pulmón en una enfermedad crónica. *EL PAÍS*. 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.*** https://ecancer.org/es/news/22569-desarrollan-una-herramienta-de-inteligencia-artificial-para-predecir-

Espinoza Freire, E. E. (2018). El problema de investigación. *Revista Conrado*, *14*(64), 22–32. https://conrado.ucf.edu.cu/index.php/conrado/article/view/808

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

Martín, S., & Manjarrés, S., & Martín, S. (2019). ***Aspectos metodológicos de la instrumentalización de la recogida de datos primarios y sus consideraciones éticas en la investigación clínica.*** Enfermería en Cardiología, 26(76), 21-26. https://enfermeriaencardiologia.com/media/acfupload/627a2235b9a86_Resumen-ART-1.pdf

Rajpurkar, P., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball,R., Langlotz, C., Shpanskaya, K Lungren, M., & Ng, A. (2017). Mura: Large dataset for abnormality detection in musculoskeletal radiographs. https://arxiv.org/abs/1712.06957

Ramírez Arévalo, H. H., & Herrera Cubides, J. F. (2013). Un viaje a través de bases de datos espaciales NoSQL. Redes de Ingeniería, 4(2), 57–69. https://doi.org/10.14483/2248762X.5923

Rego Rodríguez, F. A., Germán Flores, L., & Vitón-Castillo, A. A. (2022). Artificial intelligence and machine learning: present and future applications in health sciences. *Seminars in Medical Writing and Education*, *1*, 9. https://doi.org/10.56294/mw20229