INGENIO
2025, VOL 9(1), JULIO - DICIEMBRE, PP. 015-025

REVISTA INGENIO

Clasificación de Cáncer de Mama con Implementación de Técnicas de Análisis de Componente Principal

Classification Of Breast Cancer with Implementation of Principal Component Analysis Techniques

José Alberto León Alarcón | Universidad Técnica de Manabí, UTM
Roly Steeven Cedeño Menéndez | Universidad Técnica de Manabí, UTM

Recibido: 8/5/2025
Recibido tras revisión: 31/7/2025
Aceptado: 12/9/2025
Publicado: 28/1/2026

PALABRAS CLAVE
Procesamiento de datos, Cáncer, Inteligencia artificial, Aplicación informática.

RESUMEN
El cáncer de mama es una de las principales causas de mortalidad en mujeres a nivel mundial, lo que subraya la importancia de implementar herramientas de diagnóstico precisas y eficientes. Este estudio evaluó el desempeño de varios algoritmos de aprendizaje automático para la clasificación de tumores mamarios utilizando el Wisconsin Breast Cancer Dataset. Se aplicó Análisis de Componentes Principales (PCA) para reducir la dimensionalidad del conjunto de datos, mejorando la eficiencia computacional y manteniendo la información crítica para la clasificación.

Los modelos evaluados incluyeron Regresión Logística, Máquinas de Soporte Vectorial (SVM), Redes Neuronales, alcanzando valores máximos de AUC-ROC de 0.96, 0.95 y 0.99, respectivamente. Los resultados se compararon con estudios previos, evidenciando la solidez y aplicabilidad del enfoque propuesto.

Aunque los hallazgos son prometedores, el estudio reconoce limitaciones, como el uso de un único dataset, y sugiere integrar características clínicas adicionales en investigaciones futuras. Este trabajo demuestra la capacidad del aprendizaje automático para mejorar el diagnóstico temprano del cáncer de mama, con potencial para aplicaciones en entornos clínicos.

KEY WORDS
Data processing, Cancer, Artificial intelligence, Computer applications.

ABSTRACT
Breast cancer is one of the leading causes of mortality in women worldwide, underscoring the importance of implementing accurate and efficient diagnostic tools. This study evaluated the performance of several machine learning algorithms for breast tumor classification using the Wisconsin Breast Cancer Dataset. Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset, improving computational efficiency while maintaining critical information for classification.

The models evaluated included Logistic Regression, Support Vector Machines (SVM), Neural Networks, reaching maximum AUC-ROC values of 0.96, 0.95 and 0.99, respectively. The results were compared with previous studies, evidence of the robustness and applicability of the proposed approach.

Although the findings are promising, the study acknowledges limitations, such as the use of a single dataset, and suggests integrating additional clinical features in future research. This work demonstrates the ability of machine learning to improve early diagnosis of breast cancer, with potential for applications in clinical settings.

1. INTRODUCCIÓN

El cáncer de mama constituye uno de los principales problemas de salud a nivel mundial debido a su alta incidencia y mortalidad. Si bien los avances médicos han permitido mejorar los índices de supervivencia, el diagnóstico temprano continúa siendo un desafío. En particular, los métodos tradicionales de evaluación clínica pueden presentar limitaciones asociadas a la subjetividad del observador y a la variabilidad en la interpretación de resultados.

En este contexto, se ha incrementado el interés en el desarrollo de herramientas computacionales basadas en el análisis de datos biomédicos, que permitan apoyar de manera objetiva la clasificación de tumores mamarios. Estas herramientas no buscan reemplazar la labor clínica, sino complementar la toma de decisiones mediante el uso de

https://doi.org/10.29166/ingenio.v9i1.8194

pISSN 2588-0829
eISSN 2697-3243
fing.revista.ingenio@uce.edu.ec

algoritmos de aprendizaje automático capaces de procesar grandes volúmenes de datos y detectar patrones relevantes.

En la investigación [1] expresa que este tipo de cáncer es uno de los más frecuentemente diagnosticados, siendo este la quinta causa de muerte relacionada con el cáncer, con cerca de 2 millones de casos nuevos de manera anual en todo el mundo.

Aunque la tasa de mortalidad por cáncer de mama ha disminuido a lo largo de los últimos 50 años, debido a la mejoras diagnósticas y terapéuticas por parte del personal de salud, este tipo de cáncer sigue siendo un problema de salud pública mundial [2].

A medida que la tecnología avanza, se han desarrollado diversas herramientas estadísticas y de aprendizaje automático que facilitan el diagnóstico y la clasificación de esta enfermedad. Uno de estos enfoques es el uso de algoritmos de aprendizaje automático que permite la clasificación de datos sobre el paciente que podría padecer o no un cáncer mamario. Los algoritmos de clasificación, que buscan distinguir entre tumores benignos y malignos, son una pieza fundamental de este enfoque. Sin embargo, el rendimiento de estos algoritmos depende en gran medida de la calidad de los datos disponibles y de las técnicas de preprocesamiento utilizadas para extraer características relevantes.

La reducción de dimensionalidad es un paso crítico en problemas de clasificación con conjuntos de datos de alta dimensión, como los relacionados con imágenes médicas o datos genómicos. El Análisis de Componentes Principales (PCA) es una técnica consolidada para reducir la dimensionalidad de conjuntos de datos biomédicos de alta complejidad. Esta técnica se ha consolidado como una herramienta eficaz para la reducción de dimensionalidad en datos de alta complejidad. Al transformar variables originales en componentes no correlacionados que explican la mayor proporción de la varianza, PCA no solo elimina ruido y redundancia, sino que mejora la eficiencia y robustez de los modelos supervisados de clasificación. En aplicaciones específicas al cáncer de mama, esta reducción se traduce en modelos más rápidos, generalizables y clínicamente interpretables, al facilitar la diferenciación entre tumores benignos y malignos. Para lograrlo, convierte las variables que podrían estar correlacionadas en un grupo más reducido de características, conocidas como componentes principales [3]. La selección adecuada del número de componentes principales es fundamental para evitar tanto la pérdida de información valiosa como la inclusión de ruido, lo cual puede afectar el rendimiento del modelo de clasificación.

Algunos autores han desarrollado distintos modelos de predicción del protocolo de tratamiento de cáncer de mama. En el trabajo [4] se explora la utilización del aprendizaje automático haciendo uso de datos recolectados en el Hospital Mohammed VI de Marruecos, la cual contiene información de pacientes con dos objetivos (protocolo y ciclo de tratamiento). En este estudio se utilizaron modelos de clasificación como Gradient Boosting Classifier y Random Forest. Adicionalmente, los autores realizaron un análisis de importancia de características lo cual ayudaba a resaltar la importancia de las variables y mostrar la influencia positiva de algunas variables en los modelos.

Otros autores recopilaron información de la base de datos Breast Cancer Wisconsin Dataset, donde llevaron a cabo el procesamiento de datos respectivos, como el tratamiento de datos faltantes, la normalización y la división del conjunto de datos para los conjuntos de datos de entrenamiento y prueba en un porcentaje del 80% de los datos para el entrenamiento y un 20% de los datos para la prueba. En este estudio, se aplicaron dos enfoques diferentes: un enfoque se utilizó para la supervivencia del cáncer de mama, tales como Survival Random Forest y Cox; el enfoque adicional utilizado fue para la clasificación en donde se utilizaron algoritmos de clasificación como Naive Bayes y Random Forest [5].

En otra publicación se han utilizados técnicas como Naive Bayes, Regresión logística, máquina de soporte vectorial, K-Nearest Neighbor y árbol de decisión (DT), y técnicas de conjunto: Random forest (RF), Adaboost, XGBoost en el conjunto de datos de cáncer de mama y evaluados mediante el uso de diferentes medidas de rendimiento. En este trabajo la selección de características es una de las estrategias para la extracción de las características más significativas y útiles de un conjunto de datos esto ayuda al entrenamiento y la precisión del modelo [6].

Estudios previos que emplearon el Breast Cancer Wisconsin (Diagnostic) Dataset han reportado altos niveles de precisión en la clasificación de tumores. Por ejemplo, Agarap alcanzó una exactitud de ~99.0 % usando MLP con una división 70 %–30 % [7]. Entezari identificó al SVM como el clasificador más eficaz según sus métricas [8]. H. Benbrahim et al. evaluaron 11 modelos y encontraron que redes neuronales lograban hasta 96.5 % de exactitud [9]. Otros estudios mediante SVM y ANN han obtenido cerca del 97 % en exactitud [10]. Finalmente, combinaciones con preprocesamiento y selección rigurosa de variables llevaron a precisiones de hasta 99.1 % [11]various computer-aided diagnosis (CAD. Esta línea de resultados establece un benchmark claro frente al cual situamos y evaluamos nuestro enfoque con PCA + clasificadores supervisados.

El objetivo de este estudio es implementar y evaluar técnicas de aprendizaje automático, integradas con Análisis de Componentes Principales (PCA), para mejorar la eficiencia y precisión en la detección de cáncer de mama, contribuyendo a metodologías más ligeras y fácilmente integrables en entornos clínicos asistidos por computadora. Con este enfoque, se busca mejorar la precisión diagnóstica y apoyar la toma de decisiones tempranas en entornos clínicos. La elección del número óptimo de componentes principales se realizó mediante el método de codo, una técnica visual y analítica que

permite identificar el punto de inflexión donde la inclusión de más componentes no aporta mejoras significativas en la varianza explicada. Posteriormente, se evaluó el rendimiento en diferentes algoritmos de clasificación tales como la Regresión Logística, *Máquinas de Vectores de Soporte* (SVM) y Redes Neuronales en un conjunto de datos de pacientes con sospecha de cáncer de mama. Estos modelos se centrarán en identificar patrones y características específicas como el radio, la textura, el perímetro, el área del bulto mamario y entre otras características que permitan discriminar entre casos malignos y benignos, apoyando así al proceso de toma de decisiones médicas. A través del uso de estos algoritmos, se pretende optimizar el análisis de los datos disponibles, maximizando la precisión en la predicción y minimizando los errores en el diagnóstico. La investigación evaluará la eficacia de distintos enfoques de clasificación en términos de sensibilidad, especificidad y precisión, con el objetivo de aportar soluciones que puedan ser potencialmente aplicadas en entornos clínicos.

El aporte distintivo de este trabajo radica en evaluar de manera sistemática el impacto del PCA en el desempeño de tres algoritmos clásicos de clasificación (Regresión Logística, SVM y Redes Neuronales), utilizando métricas clínicas clave como sensibilidad, especificidad y AUC-ROC. A diferencia de estudios previos que aplican PCA solo como preprocesamiento, aquí se analiza cómo esta técnica contribuye al equilibrio entre reducción de dimensionalidad y rendimiento predictivo, aportando evidencia sobre su aplicabilidad en entornos clínicos con recursos limitados.

2. MÉTODOS Y MATERIALES

Este estudio se desarrolló en un entorno computacional utilizando datos secundarios de acceso público. El área de estudio corresponde a la clasificación automatizada de cáncer de mama a partir de un conjunto de datos biomédicos previamente recolectados y validados, proveniente del UCI Machine Learning Repository, específicamente el Breast Cancer Wisconsin (Diagnostic) Dataset [12].

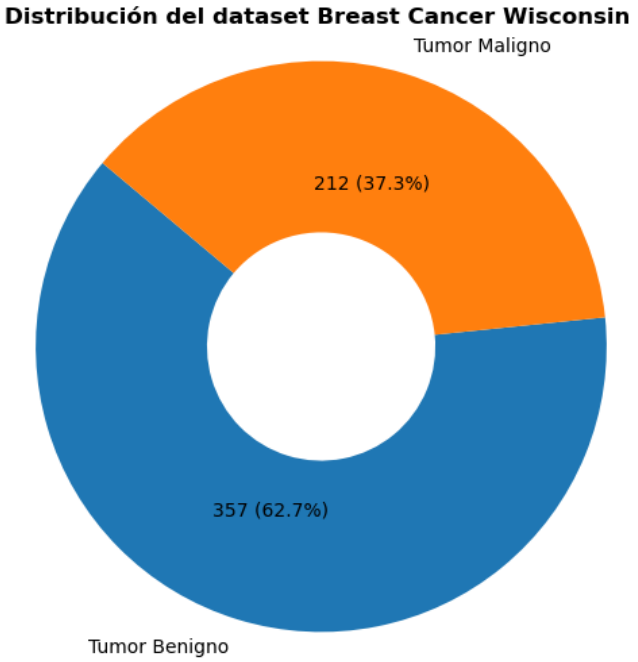
El diseño corresponde a un estudio observacional, retrospectivo y de carácter experimental-computacional, dado que se basa en registros existentes y no implicó interacción directa con pacientes.

Los hechos observacionales incluyen 569 muestras de tejido mamario (357 benignas y 212 malignas) caracterizadas mediante 30 atributos cuantitativos derivados de imágenes obtenidas por aspiración con aguja fina. Las variables comprendían de mediciones morfométricas como radio, textura, perímetro, área, suavidad, compacidad, concavidad, puntos cóncavos, simetría y dimensión fractal. La distribución de las instancias entre tumores benignos y malignos se muestra en la Fig. 1, donde se observa gráficamente la proporción de casos en cada categoría.

Las etiquetas binarias (“M” para maligno y “B” para benigno) se consideraron como la variable objetivo en las tareas de clasificación supervisada.

Fig. 1.

Distribución de instancias del Dataset



Nota. La figura muestra la distribución de clases en el conjunto de datos Breast Cancer Wisconsin [12].

Para el análisis, se procesaron los datos, se verificó que no hubiera datos ausentes. En este caso, el conjunto de datos estaba completo, por lo que no fue necesario imputar valores. Sin embargo, las etiquetas originales correspondientes a los tipos de tumores fueron transformadas a valores numéricos binarios para facilitar su manejo en los modelos de aprendizaje automático. Específicamente, la etiqueta **M** (maligno) fue codificada como **1**, mientras que la etiqueta **B** (benigno) fue codificada como **0**. Esta transformación permitió que los algoritmos de clasificación procesen las etiquetas como valores numéricos, simplificando el cálculo de métricas y optimizando el rendimiento computacional en las etapas de entrenamiento y evaluación. A partir de aquí dado que las características originales tenían magnitudes variables, se realizó un proceso de estandarización de las características para garantizar una correcta interpretación de los algoritmos de clasificación.

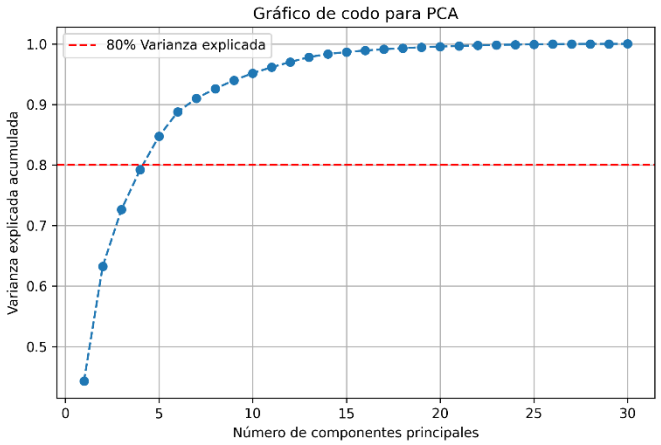
Dicho conjunto de datos fue dividido en un 80% conjunto de datos de validación y el 20% en un conjunto de datos de prueba o de testeo. Este procedimiento se llevó a cabo antes de aplicar cualquier tipo de transformación o preprocesamiento, con el fin de evitar el riesgo de fuga de datos y asegurar que los modelos a implementar aprendan únicamente a partir de la información disponible en el conjunto de entrenamiento, sin verse influenciado por datos que estarán en la evaluación final. Así, se preserva

la integridad de la validación y se evita que el modelo obtenga una ventaja indebida en la fase de pruebas.

Con el fin de reducir la dimensionalidad y mejorar la eficiencia de los modelos, se aplicó el Análisis de Componentes Principales (PCA). Esta técnica permitió reducir las 30 características iniciales a un número de 5 componentes principales, valor el cual fue escogido por medio del *método del codo* capturando cerca del 83% de la varianza explicada, preservando al máximo la variabilidad de los datos. Este punto representa un equilibrio adecuado, ya que permite reducir la dimensionalidad original de 30 a 5 componentes sin comprometer de forma significativa la información contenida en los datos (**Fig. 2**).

Fig. 2.

Gráfica del método del Codo



Nota. La figura muestra la gráfica del codo donde se observa el 80% de la varianza explicada para seleccionar el número de componentes principales.

Posteriormente, se implementaron varios algoritmos de aprendizaje automático, incluyendo Regresión Logística, *Máquinas de Vectores de Soporte* (SVM) y Redes Neuronales seleccionados por su efectividad en problemas de clasificación y su capacidad para identificar patrones complejos en datos multidimensionales.

Desde un enfoque metodológico, el aporte de este trabajo consiste en integrar PCA con tres algoritmos de clasificación bajo un esquema de validación cruzada, permitiendo comparar de forma objetiva la eficiencia computacional y el rendimiento predictivo en un problema biomédico crítico.

2.1. REGRESIÓN LOGÍSTICA

La regresión logística es un algoritmo de aprendizaje supervisado ampliamente utilizado para problemas de clasificación binaria, donde la variable dependiente toma valores discretos (0 o 1). El modelo de regresión logística estima la probabilidad de que se produzca un suceso frente a la probabilidad de que no se produzca [13]. En este caso, se ha configurado el modelo con parámetros específicos para mejorar su desempeño en conjuntos de datos con posible desbalance de clases. Estos parámetros son explicados en la Tabla 1.

TABLA I.

Parámetros de regresión logística

Parámetro	Valor	
Nivel de regularización (C)	0.01	Establece una regularización alta, lo que ayuda a prevenir el sobreajuste.
Pesos de clases (class_weight)	Balanced	Ajusta automáticamente los pesos de cada clase en función de su frecuencia en los datos de entrenamiento
Semilla (random_state)	42	Establece una semilla aleatoria para garantizar la reproducibilidad de los resultados

La implementación de la regresión logística con los parámetros ajustados garantiza una clasificación robusta en presencia de datos desbalanceados, mejora la generalización gracias a la regularización, y permite reproducir los experimentos. Su correcto ajuste y evaluación en validación cruzada son esenciales para maximizar su desempeño en aplicaciones prácticas.

2.2. MÁQUINAS DE SOPORTE VECTORIAL

Una máquina de vectores soporte (SVM) es un modelo de aprendizaje automático supervisado utilizado para la clasificación de datos. Su objetivo es identificar un hiperplano óptimo que separe las distintas clases en un espacio N-dimensional, maximizando la distancia entre ellas para mejorar la precisión del modelo [14]. En esta configuración, el modelo ha sido ajustado con parámetros específicos para mejorar su estabilidad y capacidad de generalización.

La Máquina de Soporte Vectorial con C=0.01 y kernel=”linear” garantiza un modelo interpretable, con un balance adecuado entre regularización y precisión. Además, la activación de probabilidades facilita la interpretación de los resultados.

2.3. REDES NEURONALES

A diferencia de los modelos de Regresión Logística y Máquinas de Soporte Vectorial, la Red Neuronal tuvo una configuración específica. Dicha Red Neuronal, fue creada mediante la clase Sequential [15] del Framework Keras de Python.

La red neuronal implementada sigue una estructura multicapa con activaciones no lineales. Se compone de tres

TABLA II

Parámetros de máquinas de soporte vectorial

Parámetro	**Valor**	
Nivel de regularización (C)	0.01	Establece una regularización alta, lo que ayuda a prevenir el sobreajuste.
Núcleo (kernel)	Linear	Un kernel **lineal** busca una separación directa sin introducir dimensiones adicionales.
Estimación de probabilidades (probability)	True	Habilita la estimación de probabilidades, permitiendo calcular la confianza de una clasificación. Esto es útil para interpretar los resultados y realizar análisis posteriores.

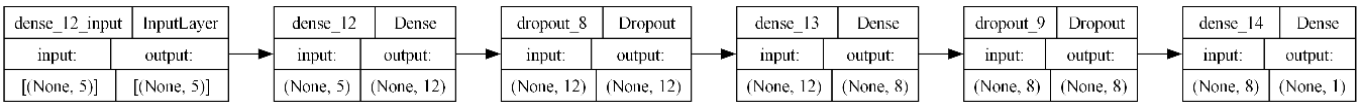
TABLA III.

Capas de la red neuronal

Capa	**Parámetro**	**Valor**	
Dense [16]	Unidades	12	Neuronas de la capa.
	Input shape	5	Características principales de entrada resultantes del PCA.
	Activación	ReLU	Función adecuada para capturar relaciones no lineales en los datos.
	Regularización	L2 (λ=0.001)	Reduce la magnitud de los pesos y evita el sobreajuste.
Dropout [17]	Tasa	0.3	Desactiva el 30% de las neuronas en cada iteración, obligando al modelo a distribuir mejor la relevancia de las características.
Dense [16]	Unidades	8	Neuronas de la capa.
	Activación	ReLU	Función adecuada para capturar relaciones no lineales en los datos.
	Regularización	L2 (λ=0.001)	Se refuerza la regularización L2 para estabilizar el aprendizaje.
Dropout [17]	Tasa	0.3	Añade otra fase de dropout, promoviendo una mejor generalización del modelo.
Dense [16]	Unidades	1	Neurona única para clasificación binaria.
	Activación	Sigmoid	Función que transforma los valores en probabilidades de pertenencia a una clase.

Fig. 3.

Diseño de la Red Neuronal



Nota. La figura presenta las capas de la red neuronal implementada.

capas densas, donde las primeras dos actúan como capas ocultas con funciones de activación ReLU, y la última corresponde a la capa de salida con activación sigmoide, adecuada para tareas de clasificación binaria. La configuración de esta red neuronal se presentó en la Tabla 3, siguiendo el mismo orden en el que se establecieron las capas.

En la Fig. 3. Se presenta la arquitectura de la red neuronal proporcionada por la función *plot_model* [18] de *Tensorflow Keras*.

Además de su arquitectura, la red neuronal fue implementada con funciones de callback que desempeñan un papel clave en la prevención del sobreajuste. Estas funciones monitorean el comportamiento del modelo durante el entrenamiento, ajustando dinámicamente parámetros como la tasa de aprendizaje o deteniendo el proceso cuando se detecta una mejora marginal en la validación. Gracias a esta estrategia, se logra una mejor generalización, evitando que el modelo memorice los datos de entrenamiento y, en cambio, aprenda patrones representativos de la información.

La Tabla 4 muestra los parámetros configurados para el callback EarlyStopping [19], el cual contribuye a evitar el sobreajuste y optimizar el uso de tiempo y recursos. Su función es interrumpir el entrenamiento cuando el

rendimiento del algoritmo deja de mejorar, evitando cómputos innecesarios y favoreciendo una mejor generalización del modelo.

TABLA IV.

Argumentos para earlystopping

Argumento	Valor	Función
Monitor	val_loss	Métrica para monitorear
Patience	10	Cantidad de épocas consecutivas sin progreso.
Restore_ best_weights	True	Restablece los pesos del modelo utilizando el valor más alto de la métrica que se está monitoreando

La Tabla 5 presenta los parámetros configurados para el callback ReduceLROnPlateau [20], el cual permite ajustar dinámicamente la tasa de aprendizaje. Gracias a esta estrategia, el modelo puede mejorar su convergencia al reducir progresivamente la tasa cuando el progreso se ralentiza, facilitando una optimización más eficiente del entrenamiento.

TABLA V.

Argumentos de reducelronplateau

Argumento	Valor	Función
Monitor	val_loss	Métrica para monitorear
Factor	0.2	Coeficiente que disminuirá la velocidad de aprendizaje (nueva_lr = lr * factor)
Patience	5	La cantidad de épocas sin mejoras tras los cuales se disminuirá la velocidad de aprendizaje
Min_lr	0.001	Tasa de aprendizaje mínima

En la fase de entrenamiento de la red neuronal, esta fue configurada inicialmente con 100 épocas, pero gracias a las funciones callback implementadas no se completaron en su totalidad las épocas configuradas con ello se completan solo 66/100 épocas.

La evaluación del rendimiento de los modelos de clasificación es fundamental en cualquier estudio de aprendizaje automático, ya que permite medir su capacidad predictiva y adecuación para la tarea en cuestión. Para una comparación precisa, se utilizaron métricas clave como exactitud, precisión, Recall, F1-Score y el área bajo la curva ROC (AUC-ROC), determinando el modelo más eficaz en la clasificación de muestras de cáncer de mama. Además, se implementó validación cruzada con 10 pliegues para garantizar una evaluación robusta y minimizar posibles sesgos en los resultados. Este enfoque divide el conjunto de entrenamiento en 10 subconjuntos, utilizando iterativamente 9 para entrenamiento y 1 para prueba, de modo que cada muestra se utiliza en validación exactamente una vez. Los resultados reportados corresponden al promedio y desviación estándar de las métricas obtenidas en los 10 pliegues.

La implementación computacional de este estudio se realizó con herramientas de código abierto ampliamente utilizadas en aprendizaje automático, garantizando un flujo de trabajo reproducible y eficiente. El análisis se llevó a cabo en Jupyter Notebook, un entorno que permite integrar código, gráficos y documentación para mejorar la trazabilidad y comprensión del proceso. Se empleó Python por su flexibilidad y amplia disponibilidad de bibliotecas especializadas en ciencia de datos.

3. RESULTADOS Y DISCUSIÓN

3.1. RESULTADOS.

Como se mencionó anteriormente, en este estudio se utilizó el Breast Cancer Wisconsin (Diagnostic) Dataset [12]. Disponible en el UCI Machine Learning Repository. Los resultados obtenidos para cada modelo en el que se implementó dicho conjunto de datos se resumen en Tabla 6, y se presentan en gráficos de desempeño en términos de la matriz de confusión y área bajo la curva ROC (AUC-ROC).

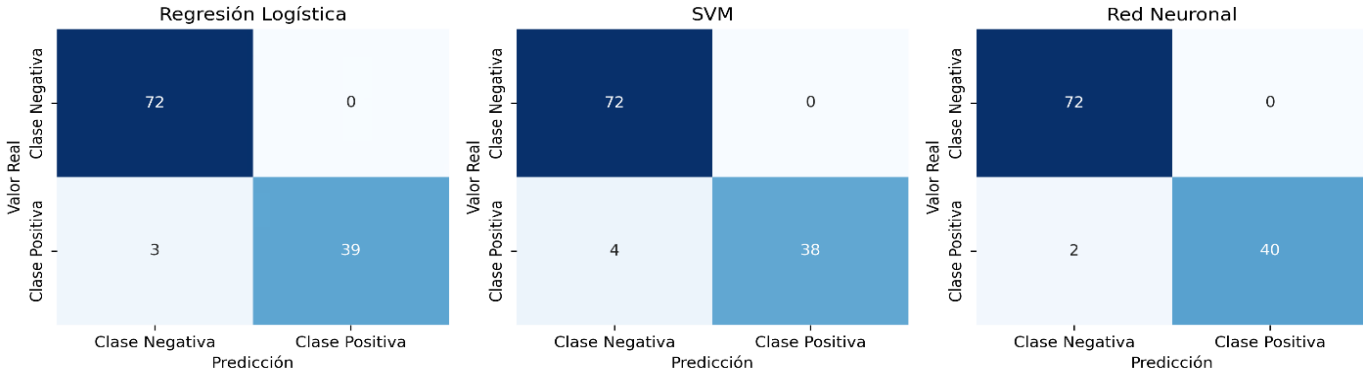
- **Regresión Logística:** El modelo alcanzó una exactitud del 97.36% y una AUC-ROC de 0.96, lo que evidencia una alta capacidad para discriminar entre casos benignos y malignos.
- **Máquinas de Soporte Vectorial (SVM):** Obtuvo una exactitud del 96.49% y una AUC-ROC de 0.95, mostrando un rendimiento ligeramente inferior al de la regresión logística, pero manteniendo una precisión competitiva.
- **Redes Neuronales:** Este modelo logró una exactitud del 98.24% y una AUC-ROC de 0.99, indicando el mejor desempeño entre los algoritmos evaluados y una notable capacidad predictiva.

Se implementó validación cruzada con 10 pliegues, permitiendo estimar la capacidad de generalización de cada algoritmo. Se calcularon exactitud, precisión, Recall y F1-score, obteniendo los valores promedio y desviación estándar para cada modelo como se muestra en la Tabla 7.

Los resultados de la validación cruzada con 10 pliegues confirman la estabilidad de los algoritmos evaluados. La Red Neuronal obtuvo el mejor desempeño promedio (Exactitud = 0.9802 ± 0.02), seguida de SVM (0.9714 ± 0.02) y Regresión Logística (0.9603 ± 0.03). Estos valores refuerzan la consistencia de los hallazgos presentados con la matriz de confusión y la curva ROC, demostrando que la reducción de dimensionalidad mediante PCA no compromete la capacidad predictiva de los modelos.

Fig. 4.

Matrices de Confusión de los Modelos Implementados



Nota. En la figura se puede observar las diferentes matrices obtenidas en base a la evaluación de los algoritmos propuestos.

De igual forma, se obtuvieron gráficos de desempeño. En la Fig. 4, se presentan las matrices de confusión para los tres modelos de clasificación evaluados: Regresión Logística, Máquinas de Vectores de Soporte (SVM), Redes Neuronales. Cada matriz muestra el número de predicciones realizadas por el modelo frente a las clases reales de los datos, proporcionando información detallada sobre los aciertos y errores en las clasificaciones.

En la Fig. 4 se puede observar para cada uno de los modelos un alto número para los verdaderos positivos y los verdaderos negativos, es decir, pacientes que verdaderamente tienen cáncer de mama fueron predichos como Maligno, en contraste con los pacientes que estaban etiquetados como Benigno fueron predichos como Benignos.

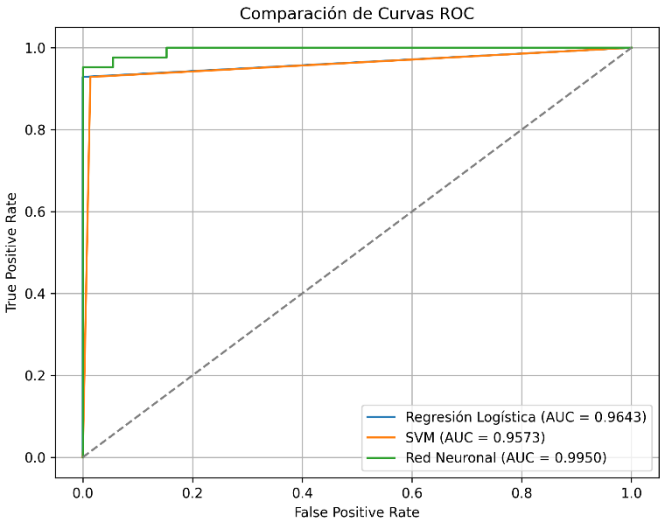
A continuación, se describe el rendimiento de la matriz de confusión en los modelos implementados:

- **Regresión logística:** Según la matriz de confusión, 72 pacientes fueron predichos como tumor benigno y realmente eran pacientes con tumor benignos; 39 pacientes fueron predichos como tumor maligno y realmente eran pacientes con tumor maligno; 3 pacientes fueron predichos como tumor benigno, pero en realidad eran tumor maligno.
- **Máquinas de Soporte Vectorial:** Según la matriz de confusión, 72 pacientes fueron predichos como tumor benigno y realmente eran pacientes con tumor benigno; 38 pacientes fueron predichos como tumor maligno y realmente eran pacientes con tumor maligno; 4 pacientes fueron predichos como tumor benigno, pero en realidad eran tumor maligno.
- **Redes Neuronales:** Según la matriz de confusión, 72 pacientes fueron predichos como tumor benigno y realmente eran pacientes con tumor benignos; 40 pacientes fueron predichos como tumor maligno y realmente eran pacientes con tumor maligno; 2 pacientes fueron predichos como tumor maligno, pero en realidad eran tumor benigno.

Asimismo, en la Fig. 5, se presentan las curvas AUC-ROC para los tres modelos de clasificación analizados. Cada curva representa la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos a diferentes umbrales de decisión.

Fig. 5.

Curvas AUC-ROC de los Modelos Implementados



Nota. En la figura se presenta las curvas ROC obtenidas en base a la evaluación de los modelos propuestos.

El análisis de las curvas AUC-ROC para los modelos evaluados muestra un rendimiento excepcional en la clasificación de cáncer de mama. El modelo implementado con una red neuronal obtuvo un área bajo la curva de 0.99, ligeramente superior a los otros dos algoritmos implementados, siendo así el modelo de regresión logística y las máquinas de soporte vectorial obtuvieron un 0.96 y 0.95, respectivamente. Este nivel de rendimiento sugiere que estos modelos son altamente efectivos y confiables para la detección de cáncer de mama.

Un hallazgo relevante es que, pese a la reducción del 83% en la dimensionalidad de los datos, los tres algoritmos mantienen métricas competitivas. En particular, la regresión logística alcanza resultados cercanos a los de la red neuronal, evidenciando que modelos más simples pueden ser adecuados en contextos con limitaciones de cómputo, mientras que las redes neuronales, aunque más complejas, ofrecen la mayor capacidad predictiva.

3.2. DISCUSIÓN

La clasificación del cáncer de mama es un área de investigación crítica, dado su impacto en la salud pública y la importancia de un diagnóstico preciso y oportuno. En este estudio, hemos evaluado la eficacia de varios algoritmos de clasificación aplicados a un conjunto de datos de pacientes con posible cáncer de mama, centrándonos en el rendimiento de modelos como Regresión Logística, Máquinas de Vectores de Soporte (SVM) y Redes Neuronales. Los resultados obtenidos a través de métricas como las curvas AUC-ROC y las matrices de confusión han proporcionado una visión profunda de la capacidad de estos modelos para discriminar entre clases.

Los resultados obtenidos en este estudio muestran que los algoritmos de aprendizaje automático, particularmente Regresión Logística y Red Neuronal, son herramientas altamente efectivas para la clasificación de tumores mamarios en pacientes con posible cáncer de mama. Ambos modelos lograron una AUC-ROC de 0.96 y 0.99, respectivamente, lo que evidencia su alta capacidad para discriminar entre tumores malignos y benignos en todos los umbrales de probabilidad evaluados. Este rendimiento destaca su potencial para ser implementados en entornos clínicos donde la precisión en el diagnóstico es crítica.

La reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA) jugó un papel fundamental en el éxito de los modelos. El método del codo, utilizado para determinar el número óptimo de componentes principales, permitió identificar un equilibrio adecuado entre la retención de información y la simplificación del modelo. Este enfoque no sólo redujo el tiempo de procesamiento, sino que también mejoró la precisión de los algoritmos al eliminar características redundantes o irrelevantes. Estudios anteriores han reportado beneficios similares al implementar PCA, lo que refuerza su validez como herramienta de preprocesamiento en problemas de clasificación de datos médicos.

Numerosos estudios han explorado el uso de algoritmos de aprendizaje automático para la clasificación de tumores mamarios, proporcionando un marco sólido para contextualizar los resultados de esta investigación. Por ejemplo, en [21] se utilizó el conjunto de datos Asia-Pacific Metaplastic Breast Cancer (AP-MBC) Consortium que consta de 347 casos de cáncer de mama metaplásico de 17 hospitales de Australia y el sudeste asiático. En este estudio fue utilizado algoritmos como: Árboles de decisión, regresión logística, naïve Bayes, perceptrón multicapa y bosque aleatorio, donde dichos algoritmos alcanzaron exactitudes y precisiones de hasta 0.838 y 0.822 respectivamente en el algoritmo de árboles aleatorios.

Asimismo, [6] utilizó varias técnicas de clasificación ML: Naïve Bayes (NB), regresión logística (LR), máquina de vectores de soporte (SVM), K-Nearest Neighbor (KNN), árbol de decisión (DT), y técnicas de conjunto: Random forest (RF), Adaboost, XGBoost en el conjunto de datos de cáncer de mama similar proveniente del Wisconsin Breast Cancer Dataset (WBCD). El autor encontró que tanto el árbol de decisión como el clasificador XGBoost tienen la mayor precisión del 97% entre todos los modelos y el mayor AUC 0,999 obtenido para el clasificador XGBoost.

No obstante, existen diferencias que merecen ser discutidas, en la investigación [22] utiliza algoritmos como Spectral clustering, DBSCAN y k-means, junto con modelos de predicción como Support Vector Machines (SVM), árboles de decisión y Random Forest. Los resultados demuestran la capacidad del modelo para predecir el tiempo que tardará el tumor en reaparecer o el tiempo que tardará el paciente en recuperarse por completo con la mejor precisión del 78,7% utilizando SVM.

A diferencia de estudios previos que reportan únicamente métricas puntuales en particiones de entrenamiento/prueba, este trabajo incorpora validación cruzada con 10 pliegues, lo que asegura que los resultados no dependan de una división específica de los datos. Este procedimiento evidencia que los tres algoritmos mantienen un rendimiento estable en diferentes subconjuntos, con desviaciones estándar reducidas en todas las métricas, lo que refuerza la robustez del enfoque propuesto.

Aunque en este trabajo no se realizó una comparación directa entre modelos con y sin PCA, existen estudios que sí la han realizado y cuyos hallazgos ofrecen un punto de referencia útil. Por ejemplo, en *RF-PCA* combinan PCA con selección de atributos y muestran que el modelo con PCA tiene un desempeño superior en precisión y tiempos de entrenamiento comparados con el modelo original sin reducción dimensional [23]which can effectively solve the problems of insufficient recognition accuracy and long time-consuming in traditional breast cancer diagnosis methods. To solve these problems, we proposed a method of attribute selection and feature extraction based on random forest (RF. También, en el trabajo [24] analiza resultados obtenidos aplicando PCA/KPCA, mostrando que estos métodos mejoran métricas como precisión y sensibilidad frente al uso de todas las variables. En *A Study Using PCA and LDA on Wisconsin Breast Cancer* [25], se evidencia que combinar PCA y LDA permite mantener un rendimiento alto, comparable al que se obtiene con modelos sin reducción, aunque no todos los trabajos informan tiempos de entrenamiento u otros

costos computacionales. Estos estudios apoyan la idea de que la reducción de dimensionalidad puede conservar precisión alta, lo que coincide con nuestros resultados: en nuestros modelos con PCA también obtenemos exactitud, precisión y sensibilidad elevados, lo que sugiere que la pérdida de información no es crítica bajo las condiciones de dimensión reducida empleadas.

Los resultados tienen implicaciones importantes para el desarrollo de sistemas automatizados de apoyo al diagnóstico médico. La alta exactitud y precisión de los modelos sugiere que podrían integrarse en flujos de trabajo clínicos para asistir a los profesionales de la salud en la detección temprana de tumores malignos, reduciendo potencialmente los tiempos de diagnóstico y mejorando las tasas de supervivencia. Aunque ambos modelos muestran un rendimiento excepcional, la variabilidad inherente a los datos clínicos y las características individuales de los pacientes pueden influir en la generalización de estos hallazgos. Por lo tanto, se recomienda realizar estudios adicionales con conjuntos de datos más amplios y diversos para validar la robustez de estos modelos en la práctica clínica real.

Además, desde una perspectiva teórica, este trabajo contribuye al creciente cuerpo de literatura que explora el uso de técnicas de reducción de dimensionalidad y algoritmos de clasificación en datos médicos. El enfoque metodológico empleado puede servir como referencia para futuras investigaciones que busquen optimizar modelos predictivos en dominios similares.

Este trabajo complementa la literatura al mostrar que PCA no solo facilita la reducción de variables, sino que además permite que modelos tradicionalmente menos potentes, como la regresión logística, alcancen niveles de precisión comparables a modelos más complejos. Este hallazgo tiene implicaciones prácticas en el diseño de sistemas de apoyo al diagnóstico, donde la eficiencia y la interpretabilidad son tan relevantes como la exactitud.

A pesar de los resultados prometedores, este estudio presenta ciertas limitaciones que deben ser consideradas. En primer lugar, el conjunto de datos utilizado, aunque ampliamente empleado en investigaciones, no representa necesariamente la diversidad de poblaciones globales. Factores como la etnicidad, la edad y las comorbilidades podrían influir en el desempeño de los modelos cuando se aplican a otras poblaciones.

En segundo lugar, el uso de PCA, aunque beneficioso en términos de rendimiento, implica una pérdida de interpretabilidad de las características originales. En contextos clínicos, esta limitación puede dificultar la aceptación de los modelos por parte de los médicos, quienes podrían preferir enfoques que ofrezcan explicaciones más claras sobre las decisiones del modelo.

Finalmente, los modelos fueron evaluados en un entorno controlado con un conjunto de datos bien definido. Su desempeño en entornos reales podría verse afectado por factores como ruido en los datos, desequilibrio de clases y variabilidad en la calidad de las imágenes o datos clínicos recolectados.

Este estudio presenta un avance significativo en el uso de algoritmos de aprendizaje automático para la clasificación de tumores mamarios. Sin embargo, el desarrollo de sistemas robustos y ampliamente aplicables para este propósito requiere una exploración más profunda y diversa. A partir de aquí, se plantean recomendaciones que pueden orientar futuras investigaciones y aplicaciones en este campo.

El conjunto de datos utilizado en este estudio, aunque reconocido en la comunidad científica, tiene limitaciones inherentes, como su tamaño y representatividad demográfica. Se recomienda incorporar datos de distintas regiones geográficas y demográficas para garantizar que los modelos sean generalizables y efectivos en poblaciones diversas. Esto incluye datos de pacientes con diferentes antecedentes genéticos, edades, y condiciones médicas previas. Además de aumentar la variedad de características clínicas y genéticas para explorar la capacidad de los modelos de identificar patrones complejos relacionados con la aparición de tumores mamarios.

El uso de PCA en este estudio resultó ser altamente beneficioso para la reducción de dimensionalidad, pero existen otras técnicas que podrían complementar o mejorar este enfoque, por ejemplo, el uso de técnicas híbridas combinando PCA con métodos de selección de características como el algoritmo de fuerza bruta análisis de relevancia mutua o métodos basados en entropía. Esto permitiría retener características más relevantes mientras se reduce la complejidad computacional. Adicional a esto, se recomienda evaluar el impacto del número de componentes principales seleccionados en la precisión de los modelos, utilizando enfoques más dinámicos para ajustar este parámetro en función de los datos.

Si bien este estudio se enfocó en modelos tradicionales de aprendizaje automático. Una recomendación de suma importancia sería el uso de redes neuronales profundas, las cuales podrían traer beneficios adicionales. En el caso del uso de redes neuronales convolucionales CNN para datos de imágenes como mamografías lo cual permite integrar automáticamente las características más relevantes en el proceso de clasificación. Sin dejar de lado la implementación de transferencia por aprendizaje, haciendo uso de modelos preentrenados con grandes conjuntos de datos médicos para mejorar la precisión en los conjuntos de datos más pequeños.

La evaluación en escenarios clínicos reales puede conllevar a que los modelos sean validados y probados en entornos reales, esto ayuda a garantizar su aplicabilidad y aceptación. Siendo así, la implementación de estudios piloto en hospitales o clínicas proporciona información valiosa sobre su viabilidad práctica. Asimismo, el desarrollo de interfaces amigables para médicos, donde se podría integrar los modelos en sistemas de apoyo al diagnóstico existentes.

4. CONCLUSIONES

El aporte central de este estudio consiste en demostrar que la reducción de dimensionalidad mediante PCA no compromete la calidad de la clasificación del cáncer de mama, sino que optimiza el uso de recursos computacionales y facilita la interpretación de modelos supervisados. Este demuestra un enfoque reproducible para la clasificación automática de cáncer de mama, combinando reducción de dimensionalidad mediante PCA con algoritmos de aprendizaje automático. Este procedimiento permitió disminuir en un 83% las variables iniciales, manteniendo una exactitud superior al 98%. Los resultados evidencian que técnicas computacionales clásicas, como la regresión logística y las redes neuronales, pueden alcanzar un rendimiento competitivo sin necesidad de arquitecturas de mayor complejidad, lo que las hace viables en escenarios con recursos computacionales limitados.

La inclusión de validación cruzada aporta solidez estadística a los resultados, confirmando que la reducción de dimensionalidad con PCA no solo mejora la eficiencia computacional, sino que mantiene la consistencia de las métricas de clasificación en múltiples particiones de los datos.

Además, la comparación con estudios previos confirma que las métricas obtenidas se ubican en el rango superior de lo reportado en la literatura, lo que respalda la validez metodológica del enfoque. Estos hallazgos refuerzan la importancia de integrar técnicas de preprocesamiento robustas con modelos de clasificación supervisada en problemas de datos biomédicos de alta dimensionalidad.

Como limitación, los experimentos se realizaron sobre un único conjunto de datos de referencia (Breast Cancer Wisconsin), lo que restringe la generalización de los resultados. Futuras investigaciones podrían ampliar este análisis incorporando otros conjuntos de datos biomédicos y explorando la integración de información heterogénea, como imágenes médicas o datos genómicos, para evaluar la escalabilidad y robustez del enfoque en diferentes contextos.

En términos prácticos, este análisis comparativo evidencia que tanto la regresión logística como las redes neuronales pueden integrarse en sistemas de apoyo clínico, siendo la primera más adecuada en entornos con recursos limitados y la segunda en escenarios donde la precisión diagnóstica es prioritaria.

REFERENCIAS

[1] M. M. Cedeño Cedeño *et al.*, «Impact of primary prevention in the early diagnosis and mortality of breast cancer in Ecuador», *Rev. Latinoam. Hipertens.*, vol. 19, n.º 3, abr. 2024, http://doi.org/10.5281/zenodo.10980345.

[2] J. Álvarez Fernández, P. Palacios Ozores, V. Cebey López, A. Cortegoso Mosquera, y R. López López, «Cáncer de mama», *Medicine (Baltimore)*, vol. 13, n.º 27, pp. 1506-1517, mar. 2021, https://10.1016/j.med.2021.03.002.

[3] IBM, «¿Qué es el análisis de componentes principales (PCA)?» [En línea]. Disponible en: https://www.ibm.com/es-es/think/topics/principal-component-analysis

[4] H. Ait Brahim, S. El-Hadaj, y A. Metrane, «Machine learning analysis of breast cancer treatment protocols and cycle counts: A case study at Mohammed vi hospital, Morocco», *Syst. Soft Comput.*, vol. 6, p. 200097, dic. 2024, https://doi.org/10.1016/j.sasc.2024.200097.

[5] M. Emily, F. Meidioktaviana, G. Z. Nabiilah, y J. V. Moniaga, «Comparative analysis of machine learning and survival analysis for breast cancer prediction», *Procedia Comput. Sci.*, vol. 245, pp. 759-767, nov. 2024, https://doi.org/10.1016/j.procs.2024.10.302.

[6] V. Nemade y V. Fegade, «Machine Learning Techniques for Breast Cancer Prediction», *Procedia Comput. Sci.*, vol. 218, pp. 1314-1320, ene. 2023, https://doi.org/10.1016/j.procs.2023.01.110.

[7] A. F. Agarap, «On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset», en *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*, feb. 2018, pp. 5-9. https://doi.org/10.48550/arXiv.1711.07831.

[8] R. Entezari, «Breast Cancer Diagnosis via Classification Algorithms», 3 de julio de 2018, *arXiv, Toronto, Canadá*: 1807.01334. https://doi.org/10.48550/arXiv.1807.01334.

[9] H. Benbrahim, H. Hachimi, y A. Amine, «Comparative Study of Machine Learning Algorithms Using the Breast Cancer Dataset», en *AI2SD 2019*, Cham, Springer: Springer International Publishing, feb. 2020, pp. 83-91. https://doi.org/10.1007/978-3-030-36664-3_10.

[10] E. S. Simant Prakoonwit, «Effective Feature Engineering and Classification of Breast Cancer Diagnosis: A Comparative Study», *BioMedInformatics*, n.º 3, pp. 616-631, agosto de 2023.

[11] S. Aamir *et al.*, «Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques», *Comput. Math. Methods Med.*, vol. 2022, n.º 1, p. 5869529, ago. 2022, https://doi.org/10.1155/2022/5869529.

[12] W. W. Olvi Mangasarian, «Breast Cancer Wisconsin (Diagnostic)». UCI Machine Learning Repository, 1993. https://doi.org/10.24432/C5DW2B.

[13] H.-Y. Kim, «Statistical notes for clinical researchers: logistic regression», *Restor. Dent. Endod.*, vol. 42, n.º 4, pp. 342-348, sep. 2017, https://doi.org/10.5395/rde.2017.42.4.342.

[14] E. Kavlakoglu, «What Is Support Vector Machine? », IBM. [En línea]. Disponible en: https://www.ibm.com/think/topics/support-vector-machine

[15] F. Chollet, «Keras documentation: The Sequential model». [En línea]. Disponible en: https://keras.io/guides/sequential_model/
[16] Keras Team, «Keras documentation: Dense layer». [En línea]. Disponible en: https://keras.io/api/layers/core_layers/dense/
[17] Keras Team, «Keras documentation: Dropout layer». [En línea]. Disponible en: https://keras.io/api/layers/regularization_layers/dropout/
[18] Keras Team, «Keras documentation: Model plotting utilities». [En línea]. Disponible en: https://keras.io/api/utils/model_plotting_utils/
[19] Keras Team, «Keras documentation: EarlyStopping». [En línea]. Disponible en: https://keras.io/api/callbacks/early_stopping/
[20] Keras Team, «Keras documentation: ReduceLROnPlateau». [En línea]. Disponible en: https://keras.io/api/callbacks/reduce_lr_on_plateau/
[21] Y. Feng *et al.*, «Predicting breast cancer-specific survival in metaplastic breast cancer patients using machine learning algorithms», *J. Pathol. Inform.*, vol. 14, p. 100329, ago. 2023, https://doi.org/10.1016/j.jpi.2023.100329.
[22] S. R. Gupta, «Prediction time of breast cancer tumor recurrence using Machine Learning», *Cancer Treat. Res. Commun.*, vol. 32, p. 100602, jul. 2022, https://doi.org/10.1016/j.ctarc.2022.100602.
[23] K. Bian, M. Zhou, F. Hu, y W. Lai, «RF-PCA: A New Solution for Rapid Identification of Breast Cancer Categorical Data Based on Attribute Selection and Feature Extraction», *Front. Genet.*, vol. 11, sep. 2020, https://doi.org/10.3389/fgene.2020.566057.
[24] R. Pirchio, «Clasificación de cáncer de mama con técnicas de análisis de la componente principal-Kernel PCA, algoritmos de máquina de vectores de soporte y regresión logística», *MediSur*, vol. 20, n.º 2, pp. 199-209, abr. 2022.
[25] G. Esen, A. Altaibek, J. Amankulov, B. Matkerim, y M. Nurtas, «Enhancing Breast Cancer Detection with Dimensionality Reduction Techniques: A Study Using PCA and LDA on Wisconsin Breast Cancer Data», *Procedia Comput. Sci.*, vol. 251, pp. 414-421, dic. 2024, https://doi.org/10.1016/j.procs.2024.11.128.

ANEXOS

TABLA VI.

Métricas de evaluación

Modelo	Exactitud	Precisión	Recall	F1-Score
Regresión Logística	0.97	1.00	0.92	0.96
Máquinas de Soporte Virtual	0.96	1.00	0.90	0.95
Redes Neuronales	0.98	1.00	0.95	0.97

TABLA VII.

Métricas de la validación cruzada

Modelo	Exactitud	Precisión	Recall	F1-Score
Regresión Logística	0.96 ± 0.03	0.96 ± 0.03	0.92 ± 0.06	0.94 ± 0.04
Máquinas de Soporte Virtual	0.97 ± 0.02	0.99 ± 0.02	0.92 ± 0.06	0.95 ± 0.03
Redes Neuronales	0.98 ± 0.02	0.97 ± 0.04	0.96 ± 0.05	0.96 ± 0.04