ISBN: 978-628-96613-1-6. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2025.1.1.2051

Predictive Accuracy in the Detection of Breast Cancer through Machine Learning

Darwin Patiño-Pérez,Ph.D[1]; Freddy Burgos-Robalino,MSc[1]; Zynnia Reyes-Sánchez,MSc[1]; Ana Ramírez-Hecksher,MSc[2]; **Celia Munive-Mora,BS**[3]

[1]Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Física,
Grupo de Investigación de Inteligencia Artificial, Ecuador,darwin.patinop@ug.edu.ec,
[2]Universidad de Guayaquil, Facultad de Ciencias Médicas, Ecuador, freddy.burgosr@ug.edu.ec,
zynnia.reyessan@ug.edu.ec,ana.ramirezh@ug.edu.ec,
[3]St Luke's University Hospital Network, PA, United States,celia.munive@sluhn.org

Abstract*– Breast cancer is one of the most significant health problems worldwide, and its early detection is crucial to improve clinical outcomes for patients. In this context, machine learning models have become valuable tools to predict the presence of this disease with greater precision. This study performs a comparative analysis of three machine learning models: logistic regression, random forest, and support vector machines (SVM), using the Wisconsin Breast Cancer Diagnosis dataset. This data set includes features derived from fine-needle aspiration images of breast masses, with 357 benign and 212 malignant cases. The results of the study reveal that the random forest model outperforms the other two in terms of predictive accuracy. This model, which uses the top 5 predictors ("concave point mean", "area mean", "radius mean", "perimeter mean", and "concavity mean"), achieves an accuracy of about 94.15% and a cross-validation score of about 95.61% on the test data set. These findings highlight the effectiveness of random forest in identifying complex patterns in data, making it a promising tool for breast cancer prediction. In conclusion, this study demonstrates the potential of machine learning models, particularly random forest, in improving early detection of breast cancer. These advances could have a significant impact on clinical practice, facilitating more accurate and timely diagnoses, which in turn could improve patient outcomes and reduce mortality associated with this disease.

***Keywords*— Breast Cancer, Machine Learning, Early Detection, Predictive Accuracy, Support Vector Machine.**

Precisión Predictiva en la Detección de Cancer de Mama Mediante *Machine Learning*

Darwin Patiño-Pérez,Ph.D[1]; Freddy Burgos-Robalino,MSc[1]; Zynnia Reyes-Sánchez,MSc[1]; Ana Ramírez-Hecksher,MSc[2]; Celia Munive-Mora,BS[3]
[1]Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Física,
Grupo de Investigación de Inteligencia Artificial, Ecuador,darwin.patinop@ug.edu.ec,
[2]Universidad de Guayaquil, Facultad de Ciencias Médicas, Ecuador, freddy.burgosr@ug.edu.ec,
zynnia.reyessan@ug.edu.ec,ana.ramirezh@ug.edu.ec,
[3]St Luke's University Hospital Network, PA, United States,celia.munive@sluhn.org

Resumen– ***El cáncer de mama es uno de los problemas de salud más significativos a nivel mundial, y su detección temprana es crucial para mejorar los resultados clínicos de los pacientes. En este contexto, los modelos de aprendizaje automático se han convertido en herramientas valiosas para predecir la presencia de esta enfermedad con mayor precisión. Este estudio realiza un análisis comparativo de tres modelos de aprendizaje automático: regresión logística, bosque aleatorio y máquinas de soporte vectorial (SVM), utilizando el conjunto de datos de diagnóstico de cáncer de mama de Wisconsin. Este conjunto de datos incluye características derivadas de imágenes de aspiración con aguja fina de masas mamarias, con 357 casos benignos y 212 malignos. Los resultados del estudio revelan que el modelo de bosque aleatorio supera a los otros dos en términos de precisión predictiva. Este modelo, que utiliza los 5 predictores principales ("media de puntos cóncavos", "media de área", "media de radio", "media de perímetro" y "media de concavidad"), alcanza una precisión de aproximadamente el 94.15% y una puntuación de validación cruzada de alrededor del 95.61% en el conjunto de datos de prueba. Estos hallazgos resaltan la eficacia del bosque aleatorio para identificar patrones complejos en los datos, lo que lo convierte en una herramienta prometedora para la predicción del cáncer de mama. En conclusión, este estudio demuestra el potencial de los modelos de aprendizaje automático, particularmente el bosque aleatorio, en la mejora de la detección temprana del cáncer de mama. Estos avances podrían tener un impacto significativo en la práctica clínica, facilitando diagnósticos más precisos y oportunos, lo que a su vez podría mejorar los resultados de los pacientes y reducir la mortalidad asociada a esta enfermedad.***

Palabras clave- Cáncer de Mama, Aprendizaje Automático , Detección Temprana, Precisión Predictiva, Máquina de Soporte Vectorial.

I. Introducción

El cáncer de mama, una de las formas de cáncer más prevalentes entre las mujeres en todo el mundo, tiene un impacto significativo en la salud pública y el bienestar individual[1]. La detección temprana y la predicción precisa del cáncer de mama son cruciales para mejorar los resultados de los pacientes, la planificación del tratamiento y las tasas de supervivencia. Los enfoques de diagnóstico convencionales a menudo se basan en interpretaciones subjetivas y análisis manuales, que pueden consumir mucho tiempo y ser propensos a errores.

En los últimos años, las técnicas de aprendizaje automático han surgido como herramientas poderosas para la predicción del cáncer de mama, ofreciendo el potencial de mejorar la precisión del diagnóstico y facilitar estrategias de tratamiento personalizadas. Aprovechando los algoritmos computacionales, los modelos de aprendizaje automático pueden analizar patrones complejos dentro de grandes conjuntos de datos, lo que permite el descubrimiento de información valiosa para la predicción precisa del cáncer de mama[2][3].

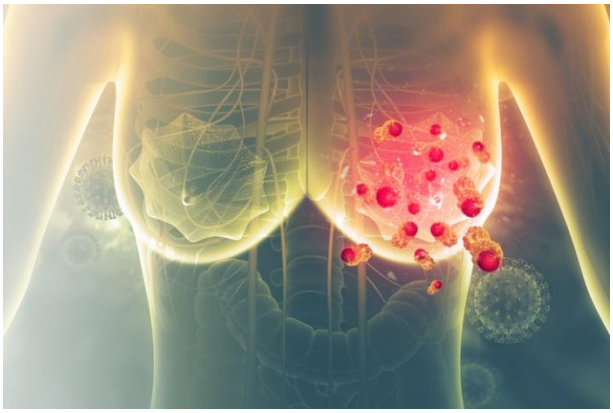

Fig. 1 Cáncer de Mama.

Los algoritmos de aprendizaje automático también desempeñaron un papel importante en el campo de la clasificación de datos genéticos del cáncer[4]. Los modelos de Regresión Logística se han investigado ampliamente para la predicción del cáncer de mama, demostrando su potencial para clasificar con precisión los casos benignos y malignos [5]. Este algoritmo clásico de clasificación binaria ofrece una buena interpretabilidad para relaciones lineales simples, proporcionando información valiosa sobre la probabilidad de aparición del cáncer de mama.

Las máquinas de soporte vectorial (SVM) son un método ampliamente utilizado en el aprendizaje automático, especialmente en aplicaciones médicas como la predicción del cáncer de mama[6]. Estos modelos buscan encontrar un hiperplano óptimo que separe las clases, como tumores benignos y malignos, maximizando el margen entre ellas para

lograr una clasificación precisa. Sin embargo, su principal limitación es la dificultad para escalar en conjuntos de datos muy grandes, lo que puede ser un desafío cuando se trabaja con grandes volúmenes de datos médicos[7].

Para manejar la complejidad de los datos no lineales, como los patrones en imágenes de mamografías o características derivadas de biopsias, se utilizan funciones de kernel que transforman los datos a un espacio de mayor dimensionalidad[8], permitiendo una separación más efectiva. Esto hace que las SVM sean una herramienta valiosa en la detección temprana del cáncer de mama, donde la precisión en la clasificación es crítica para mejorar los resultados de los pacientes.

El algoritmo denominado Bosque Aleatorio o *Random Forest*, ha demostrado ser una herramienta altamente efectiva en la predicción del cáncer de mama debido a su capacidad para manejar datos complejos y no lineales[9]. Este método combina múltiples árboles de decisión, cada uno entrenado con subconjuntos aleatorios de los datos, lo que reduce el riesgo de sobreajuste y mejora la generalización del modelo. En el contexto del cáncer de mama, el bosque aleatorio puede analizar una amplia variedad de características, como datos clínicos, biomarcadores y resultados de imágenes médicas, para identificar patrones asociados con la presencia o ausencia de la enfermedad.

Además, su capacidad para calcular la importancia de las características permite a los investigadores y médicos identificar qué variables tienen mayor impacto en el diagnóstico, lo que puede ser crucial para la toma de decisiones clínicas. Estudios recientes han mostrado que el bosque aleatorio alcanza altos niveles de precisión, recall y F1-score en la clasificación de tumores malignos y benignos, lo que lo convierte en una opción confiable para la detección temprana y el pronóstico del cáncer de mama. Su robustez y versatilidad lo posicionan como una herramienta prometedora en la lucha contra esta enfermedad.

II. Materiales y Métodos

A. *Dataset*

En este estudio se utilizó el conjunto de datos de diagnóstico de cáncer de mama de Wisconsin[10]. Este dataset según la Tabla I, contiene características calculadas a partir de imágenes digitalizadas de aspiraciones con aguja fina (FNA) de masas mamarias. Incluye un total de 569 casos, que consisten en 357 casos benignos y 212 malignos. Cada caso está representado por diez características de valor real para cada núcleo celular, incluidos radio, textura, perímetro, área, suavidad, compacidad, concavidad, puntos cóncavos, simetría y dimensión fractal.

TABLA I
MUESTRA DE DATOS

mean radius	mean texture	mean perimeter	mean area	...	worst symmetry	worst fractal dimension
17.99	10.38	122.80	1001.0	...	0.4601	0.11890
20.57	17.77	132.90	1326.0	...	0.2750	0.08902
19.69	21.25	130.00	1203.0	...	0.3613	0.08758
11.42	20.38	77.58	386.1	...	0.6638	0.17300
20.29	14.34	135.10	1297.0	...	0.2364	0.07678

Además, se calcularon la media, el error estándar y la "peor" o más grande (media de los tres valores más grandes) de estas características para cada imagen, lo que resultó en un total de 30 características. Los valores medios de radio celular, perímetro, área, compacidad, concavidad y puntos cóncavos se han identificado como características informativas para la clasificación del cáncer de mama.

B. *Modelos de Aprendizaje Automático*

Se realizará un análisis comparativo de modelos de aprendizaje supervisado —incluyendo regresión logística, máquinas de vectores de soporte (SVM) y bosque aleatorio— para la predicción de cáncer de mama a partir de variables clínicas y morfológicas. La investigación se centrará en: (1) el preprocesamiento y la caracterización de un dataset de pacientes con diagnóstico confirmado, (2) la implementación y optimización de hiperparámetros de cada algoritmo mediante validación cruzada estratificada, y (3) la evaluación comparativa del rendimiento predictivo utilizando métricas como precisión, sensibilidad y área bajo la curva ROC.

Adicionalmente, se emplearán técnicas de interpretabilidad (como análisis de importancia de características y coeficientes) para identificar los predictores clínicos más relevantes asociados al diagnóstico. Los hallazgos de este trabajo buscan establecer un marco metodológico que contribuya al desarrollo de herramientas de apoyo clínico basadas en inteligencia artificial, con potencial aplicación en la detección temprana y estratificación de riesgo en oncología mamaria.

Regresión Logística es una técnica fundamental de aprendizaje automático utilizada para problemas de clasificación binaria, donde se busca predecir una variable dependiente categórica basada en una o más variables independientes[11]. A diferencia de la regresión lineal, que predice valores continuos, la regresión logística utiliza la función sigmoide para transformar sus salidas en probabilidades entre 0 y 1, lo que la hace ideal para predecir la probabilidad de pertenencia a una clase específica ver Fig 2.

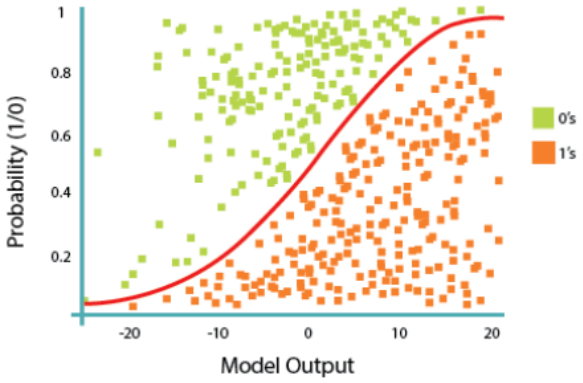

Fig. 2 Regresión Logística

Máquinas de Soporte Vectorial (SVM) representan una técnica avanzada de aprendizaje automático diseñada específicamente para resolver problemas de clasificación mediante la identificación de un hiperplano óptimo según la Fig. 3, que maximiza el margen entre diferentes clases de datos[12]. Este algoritmo se destaca por su capacidad para manejar espacios de alta dimensionalidad a través del "truco del kernel" (*kernel trick*), que permite transformar datos no linealmente separables en un espacio de características de mayor dimensión donde sí pueden separarse linealmente.

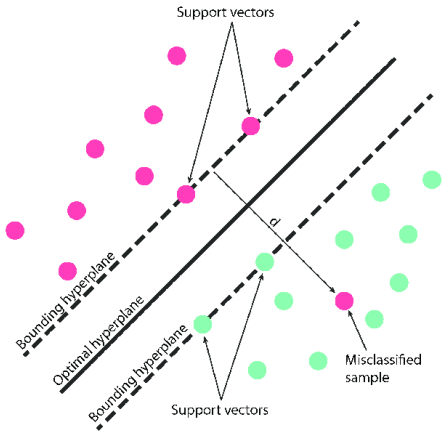

Fig. 3 Support Vector Machine - SVM

Las SVM utilizan vectores de soporte, que son los puntos de datos más cercanos al hiperplano de separación, para definir el margen máximo entre clases, lo que las hace particularmente robustas frente al ruido y eficaces en la prevención del sobreajuste. El modelo puede adaptarse a diferentes tipos de problemas mediante la selección de distintas funciones kernel (lineal, polinomial, RBF - *Radial Basis Function*), permitiendo capturar relaciones complejas en los datos. Esta versatilidad, combinada con su sólida fundamentación matemática en la teoría del aprendizaje estadístico, ha convertido a las SVM en una herramienta esencial en aplicaciones como reconocimiento de patrones, clasificación de textos, diagnóstico médico y análisis de imágenes, destacándose por su capacidad para proporcionar resultados precisos incluso con conjuntos de datos relativamente pequeños.

Bosque Aleatorio o *Random Forest* es un algoritmo de ensamble de árboles predictores según la Fig. 4, donde cada árbol se basa en los valores de un conjunto de características aleatorias extraídas de manera independiente y con la misma distribución para todos los árboles en el conjunto[13] . La principal característica de los bosques aleatorios radica en su enfoque de *ensemble learning*, que combina múltiples árboles de decisión para mejorar la precisión y robustez del modelo. A diferencia de un solo árbol de decisión, que puede ser propenso al sobreajuste (*overfitting*), este algoritmo genera una variedad de árboles entrenados con subconjuntos aleatorios de los datos y características, lo que introduce diversidad en el proceso de aprendizaje.

Esta diversidad permite que el modelo generalice mejor a nuevos datos, ya que los errores individuales de los árboles se compensan entre sí durante el proceso de votación o promediado. Además, el bosque aleatorio es capaz de manejar tanto datos numéricos como categóricos, y son particularmente útiles para identificar la importancia de las características, lo que ayuda a entender qué variables tienen mayor impacto en las predicciones. Su capacidad para equilibrar precisión, interpretabilidad y resistencia al ruido lo convierte en una de las técnicas más populares y confiables en el ámbito del aprendizaje automático.

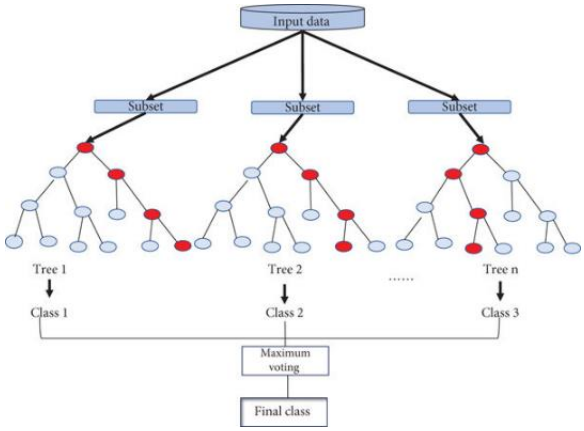

Fig. 4 Random Forest

Todas las tareas de preprocesamiento, modelado y evaluación se llevaron a cabo en el entorno de programación de Google Colab utilizando Python. Se hizo uso de la biblioteca Scikit-learn para acceder a su amplia gama de herramientas y funcionalidades. Scikit-learn ofrece una variedad de algoritmos de aprendizaje automático, técnicas de preprocesamiento de datos, métodos de evaluación de modelos y más, facilitando así la implementación de modelos de clasificación y la evaluación de su rendimiento en este entorno específico.

C. Métricas de Evaluación

Matriz de Confusión es una matriz que resume el rendimiento de un modelo de ML en un conjunto de datos de prueba según la Tabla II. Su función principal es evaluar el rendimiento de los modelos de clasificación, cuyo objetivo es predecir una etiqueta categórica para cada instancia de entrada. La matriz muestra el número de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN) producidos por el modelo en los datos de la

prueba [13] de ella se obtienen diferentes indicadores como:

Precisión: La precisión mide la exactitud general de las predicciones y se calcula como la relación entre las instancias clasificadas correctamente y el número total de instancias.

TABLA II
MATRIZ DE CONFUSION

		Predicción	
		Positivos	Negativos
Observación	Positivos	True Positive (TP)	False Negative (FN)
	Negativos	False Positive (FP)	True Negative (TN)

Accuracy o precisión: Es una métrica definida como la proporción de casos verdaderos, positivos y negativos, identificados correctamente y recuperados entre el total de instancias recuperadas por el modelo. La precisión es una media aritmética ponderada de precisión y precisión inversa según (1).

$$Accuracy: A = \frac{tp + tn}{tp + tn + fp + fn} \qquad (1)$$

Donde, tp son los verdaderos positivos, tn son los verdaderos negativos, fp son los falsos positivos y fn los falsos negativos. Esta métrica mide la exactitud general de las predicciones y se calcula como la relación entre las instancias clasificadas correctamente y el número total de instancias.

Validación cruzada: La validación cruzada es una técnica utilizada para evaluar el rendimiento de generalización de los modelos[14] En este estudio, se realizó una validación cruzada de k pliegues, dividiendo el conjunto de datos en k pliegues de igual tamaño. Los modelos se entrenaron y evaluaron k veces, y cada pliegue sirvió como conjunto de prueba una vez, mientras que los pliegues restantes se usaron para el entrenamiento.

D. Implementación

Los modelos de aprendizaje automático y las métricas de evaluación se implementaron utilizando el lenguaje de programación Python y la biblioteca scikit-learn, un conjunto de herramientas de aprendizaje automático ampliamente utilizado[15]. En las secciones siguientes, presentaremos los resultados de nuestro análisis utilizando el conjunto de datos, los modelos y las métricas de evaluación descritos. Los hallazgos arrojarán luz sobre el rendimiento predictivo de los modelos de regresión logística, árbol de decisiones y bosque aleatorio para el diagnóstico de cáncer de mama.

III. RESULTADOS

El rendimiento de cada modelo se evaluó utilizando las métricas de precisión y la validación cruzada.

La matriz de confusión para el modelo de Regresión Logística Fig. 5, muestra un rendimiento general bueno en la clasificación de casos de cáncer de mama. El modelo clasificó correctamente 99 casos benignos (Verdaderos Negativos) y 100 casos malignos (Verdaderos Positivos), lo que indica una alta precisión en la identificación de ambos tipos de casos. Sin embargo, hubo 5 falsos positivos, es decir, casos benignos que fueron incorrectamente clasificados como malignos, y 7 falsos negativos, que son casos malignos clasificados incorrectamente como benignos.

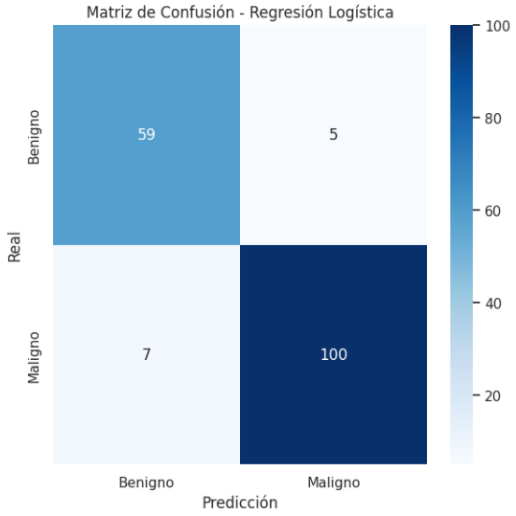

Fig. 5 Regresión Logística

Los falsos negativos son particularmente preocupantes en el contexto médico, ya que podrían retrasar el tratamiento necesario para los pacientes. Aunque el número de falsos negativos es relativamente bajo (7), es importante minimizarlo aún más para mejorar la confiabilidad del modelo. Por otro lado, los falsos positivos (5) pueden generar ansiedad innecesaria en los pacientes, pero son menos críticos que los falsos negativos. En resumen, el modelo de Regresión Logística demuestra un buen equilibrio entre la identificación de casos benignos y malignos, con una alta tasa de aciertos en ambas categorías. No obstante, hay margen de mejora, especialmente en la reducción de falsos negativos, para garantizar que todos los casos de cáncer sean detectados de manera oportuna.

La matriz de confusión para el modelo de Bosque Aleatorio Fig. 6, muestra un rendimiento general sólido en la clasificación de casos de cáncer de mama. El modelo clasificó correctamente 60 casos benignos (Verdaderos Negativos) y 101 casos malignos (Verdaderos Positivos), lo que indica una alta precisión en la identificación de ambos tipos de casos.

Sin embargo, hubo 4 falsos positivos, es decir, casos benignos que fueron incorrectamente clasificados como malignos, y 6 falsos negativos, que son casos malignos clasificados incorrectamente como benignos. Los falsos negativos son particularmente preocupantes en el contexto médico, ya que podrían retrasar el tratamiento necesario para

los pacientes. Aunque el número de falsos negativos es relativamente bajo (6), es importante minimizarlo aún más para mejorar la confiabilidad del modelo. Por otro lado, los falsos positivos (4) pueden generar ansiedad innecesaria en los pacientes, pero son menos críticos que los falsos negativos.

En resumen, el modelo basado en bosque aleatorio demuestra un buen equilibrio entre la identificación de casos benignos y malignos, con una alta tasa de aciertos en ambas categorías. No obstante, hay margen de mejora, especialmente en la reducción de falsos negativos, para garantizar que todos los casos de cáncer sean detectados de manera oportuna.

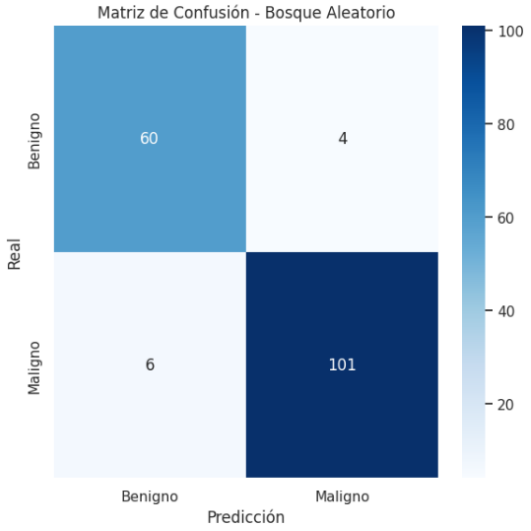

Fig. 6 Bosque Aleatorio

La matriz de confusión proporcionada para el modelo Maquina de Soporte Vectorial o SVM según Fig. 7, muestra la distribución de las predicciones en comparación con los valores reales. En este caso, se observa que el modelo tiene una alta capacidad para clasificar correctamente los casos de cáncer de mama. La matriz indica que la mayoría de los casos benignos y malignos fueron clasificados correctamente, lo que sugiere un bajo número de falsos positivos y falsos negativos.

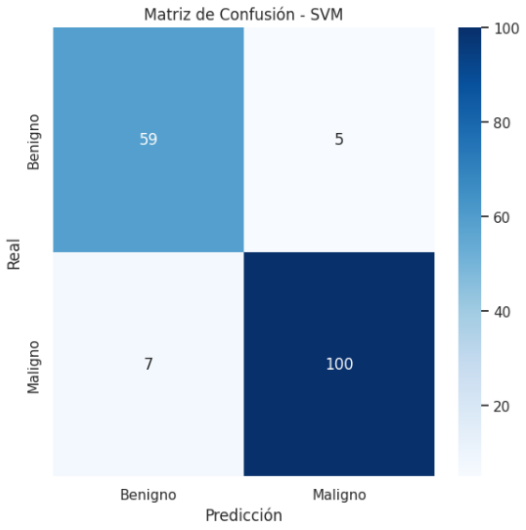

Fig. 7 Maquina de Soporte Vectorial

Esto es crucial en aplicaciones médicas, donde los falsos negativos (casos malignos clasificados como benignos) pueden tener consecuencias graves. La precisión en la clasificación de casos malignos es particularmente importante, ya que garantiza que los pacientes reciban el tratamiento adecuado de manera oportuna. En general, la matriz de confusión refleja un rendimiento sólido del modelo SVM, respaldando su utilidad en la detección temprana y precisa del cáncer de mama.

Según la Tabla III, los resultados obtenidos de los tres modelos (Regresión Logística, Bosque Aleatorio y Maquina de Soporte Vectorial) muestran un rendimiento general alto en la tarea de detección de cáncer de mama. El Bosque Aleatorio destaca con la mayor precisión (Accuracy) de 94.15%, seguido de cerca por el SVM con 93.56% y la Regresión Logística con 91.81%. Esto indica que el Bosque Aleatorio es ligeramente superior en términos de clasificación correcta de casos.

Además, el Bosque Aleatorio también tiene el mayor F1-Score (94.17%), una métrica que combina precisión y recall, lo que sugiere que este modelo equilibra mejor la identificación de casos positivos y negativos.

TABLA III
RENDIMIENTO DE LOS MODELOS

Modelo	Accuracy	Cross-Validation	F1-Score	Recall	Kappa	ROC AUC
Regresión Logística	0.918129	0.9508	0.91878	0.9181	0.8284	0.982769
Bosque Aleatorio	0.94152	0.9561	0.94169	0.9415	0.8759	0.987807
SVM	0.935673	0.9455	0.93577	0.9357	0.8631	0.984813

En cuanto a la validación cruzada (Cross-Validation), el Bosque Aleatorio obtiene la puntuación más alta (95.61%), lo que indica una mayor robustez y generalización en comparación con los otros dos modelos. Sin embargo, los otros dos modelos siguen siendo competitivos en esta métrica con 95.08% y 94.55% respectivamente.

En términos de Recall, que mide la capacidad del modelo para identificar correctamente los casos positivos, el Bosque Aleatorio nuevamente lidera con 94.15%, seguido por el SVM (93.57%) y la Regresión Logística (91.81%). Esto es crucial en aplicaciones médicas, donde es vital minimizar los falsos negativos.

El Índice Kappa, que mide la concordancia entre las predicciones y las etiquetas reales ajustada por el azar, también favorece al Bosque Aleatorio con 87.59%, indicando un acuerdo casi perfecto

La Curva ROC (Receiver Operating Characteristic) es una herramienta gráfica utilizada para evaluar el rendimiento de un modelo de clasificación binaria. Esta curva representa la relación entre la Tasa de Verdaderos Positivos (TPR) y la Tasa de Falsos Positivos (FPR) para diferentes umbrales de clasificación.

Tasa de Verdaderos Positivos (TPR): También conocida como sensibilidad, mide la proporción de casos positivos que son correctamente identificados por el modelo. Se calcula como TPR=TPTP+FNTPR=TP+FNTP, donde TP son los verdaderos positivos y FN los falsos negativos.

Tasa de Falsos Positivos (FPR): Mide la proporción de casos negativos que son incorrectamente clasificados como positivos. Se calcula como FPR=FPFP+TNFPR=FP+TNFP, donde FP son los falsos positivos y TN los verdaderos negativos.

Observando la Fig 8 se tiene que:

La Curva Celeste: Representa la curva ROC de un modelo de Regresión Logística. El área bajo la curva (AUC) es alta, lo que indica un buen rendimiento.

La Curva Naranja: Representa la curva ROC del modelo basado en Bosque Aleatorio. Este modelo tiene un AUC aún más alto, lo que sugiere un mejor rendimiento que la Regresión Logística.

La Curvva Verde: Representa la curva ROC del modelo Maquina de Soporte Vectorial. Este modelo tiene un AUC aún más bajo, lo que sugiere un mejor rendimiento que la Regresión Logística.

La Curva Azul: Representa la línea diagonal que corresponde a un modelo aleatorio, con un AUC de 0.5, lo que indica que no tiene capacidad de discriminación.

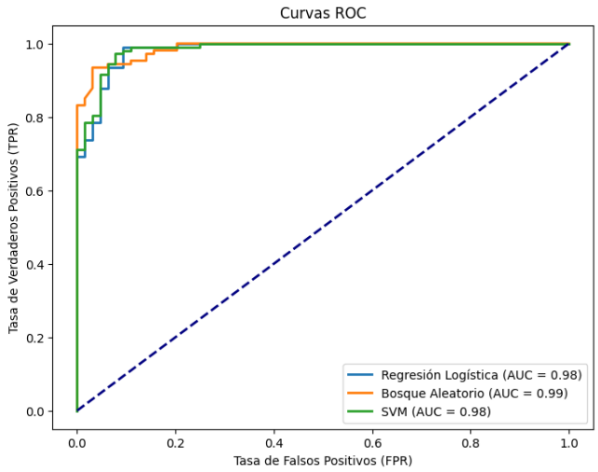

Fig. 8 Curva ROC

Finalmente, en la métrica ROC AUC, que evalúa la capacidad del modelo para distinguir entre clases, el Bosque Aleatorio nuevamente tiene el mejor desempeño con 98.78%, seguido por el SVM (98.48%) y la Regresión Logística (98.28%).

En conclusión, aunque los tres modelos muestran un rendimiento excelente, el Bosque Aleatorio se posiciona como el más efectivo en la mayoría de las métricas clave para la detección de cáncer de mama.

IV. Discusión

Los resultados de esta investigación contribuyen al creciente corpus de conocimientos en el campo de la predicción del cáncer de mama y destacan el potencial de las técnicas de aprendizaje automático para mejorar la precisión del diagnóstico. El uso de algoritmos de aprendizaje automático puede ayudar a los profesionales sanitarios a tomar decisiones informadas, lo que podría conducir a una detección más temprana del cáncer de mama y a una mejora de los resultados de los pacientes. La alta precisión alcanzada por el modelo bosque aleatorio sugiere su idoneidad para su integración en la práctica clínica como una herramienta adicional para ayudar en el diagnóstico del cáncer de mama.

A pesar de los prometedores resultados, es esencial reconocer ciertas limitaciones de este estudio. En primer lugar, el análisis se realizó únicamente en el conjunto de datos de diagnóstico de cáncer de mama de Wisconsin, lo que puede limitar la generalización de los hallazgos a otras poblaciones o conjuntos de datos. Las investigaciones futuras deberían apuntar a validar el desempeño de estos modelos en conjuntos de datos diversos y más grandes para garantizar su solidez y confiabilidad.

Además, la interpretación de las predicciones de los modelos de aprendizaje automático puede plantear desafíos debido a su complejidad inherente. Si bien el modelo de bosque aleatorio demostró un rendimiento superior, comprender el proceso específico de toma de decisiones y la importancia biológica subyacente de los predictores identificados justifica una mayor investigación.

Finalmente, este estudio demuestra la eficacia de los modelos de aprendizaje automático, en particular el algoritmo bosque aleatorio, en la predicción del cáncer de mama utilizando el conjunto de datos de diagnóstico de cáncer de mama de Wisconsin. La identificación de los principales predictores y la alta precisión predictiva del modelo bosque aleatorio enfatizan el potencial de las técnicas de aprendizaje automático para ayudar a los profesionales de la salud a realizar diagnósticos precisos y oportunos. Es necesario realizar más investigaciones para validar estos hallazgos en diversos conjuntos de datos y explorar formas de mejorar la interpretabilidad de los modelos de aprendizaje automático en el contexto del diagnóstico del cáncer de mama.

V. Conclusiones

En este estudio, comparamos tres algoritmos de aprendizaje automático para la predicción del cáncer de mama utilizando el conjunto de datos de diagnóstico de cáncer de mama de Wisconsin. Los modelos de regresión logística, la máquina de soporte vectorial y bosque aleatorio se evaluaron en función de la precisión y la importancia de las características. Los resultados destacan el potencial de las técnicas de aprendizaje automático para predecir con precisión el diagnóstico de cáncer de mama. Entre los modelos probados, el algoritmo de bosque aleatorio demostró ser el más eficaz, logrando la mayor precisión en el conjunto de datos de prueba. Su enfoque de conjunto, que combina múltiples árboles de decisiones, mejora las capacidades predictivas y la solidez. Además, identificamos predictores

claves, incluidos "media de puntos cóncavos", "media de área", "media de radio", "media de perímetro" y "media de concavidad", que ofrecen información valiosa sobre las características cruciales para la predicción del cáncer de mama. Esta información proporciona a los profesionales de la salud conocimientos diagnósticos críticos.

Los resultados obtenidos demuestran que los tres modelos (Regresión Logística, Bosque Aleatorio y SVM) son altamente efectivos para la detección de cáncer de mama, con métricas que reflejan un rendimiento sobresaliente en términos de precisión, recall y capacidad de generalización. Sin embargo, el Bosque Aleatorio se destaca como el modelo más robusto y confiable, al obtener la mayor precisión (0.94152), el mayor F1-Score (0.941694) y el mejor Índice Kappa (0.875925). Estas métricas indican que el Bosque Aleatorio no solo clasifica correctamente la mayoría de los casos, sino que también logra un equilibrio óptimo entre la identificación de casos positivos y negativos, lo que es fundamental en el ámbito médico. Además, su alto valor de ROC-AUC (0.987807) confirma su excelente capacidad para distinguir entre clases, lo que lo convierte en una herramienta prometedora para la detección temprana de esta enfermedad.

Por otro lado, aunque el SVM muestra un rendimiento ligeramente inferior al Bosque Aleatorio, su puntuación de validación cruzada (0.944715) es la más alta entre los tres modelos, lo que sugiere una mayor estabilidad y capacidad de generalización en diferentes conjuntos de datos. La Regresión Logística, aunque menos precisa en comparación, sigue siendo un modelo competitivo, especialmente por su simplicidad e interpretabilidad, lo que puede ser útil en contextos donde se requiere una explicación clara de las decisiones del modelo.

En conclusión, mientras que el Bosque Aleatorio es el modelo más efectivo en la mayoría de las métricas clave, la elección final del modelo podría depender de factores adicionales, como la interpretabilidad, la velocidad de entrenamiento o el contexto específico de aplicación. Estos avances en el uso de modelos de aprendizaje automático tienen el potencial de transformar la práctica clínica, mejorando la precisión y oportunidad de los diagnósticos, lo que podría salvar vidas y reducir la mortalidad asociada al cáncer de mama.

Nuestra investigación contribuye al diagnóstico del cáncer de mama al mostrar el potencial del aprendizaje automático para mejorar la detección temprana y las estrategias de tratamiento personalizadas. La alta precisión y la selección de características informativas del modelo Bosque Aleatorio lo hace adecuado para su integración en la práctica clínica. Sin embargo, reconocemos limitaciones, como la dependencia del conjunto de datos de Wisconsin. Es esencial una mayor validación en conjuntos de datos más grandes y diversos.

Además, abordar el desafío de la interpretabilidad de los modelos de aprendizaje automático es vital para mejorar la transparencia y la toma de decisiones. Para abordar estos desafíos, la investigación futura puede aprovechar tecnologías como Kafka para capturar una gama más amplia de datos, lo que facilita la investigación a mayor escala[16], los conocimientos del campo del reconocimiento de imágenes en el aprendizaje automático pueden inspirar avances en los métodos de detección del cáncer de mama, mejorando potencialmente la precisión y la eficiencia[17]. El desarrollo de la tecnología de aprendizaje profundo y seguimiento visual en el campo de la biología traerá más iluminación a los posteriores [18]–[20] .

Finalmente, este estudio destaca la importancia de los algoritmos de aprendizaje automático en la predicción del cáncer de mama. Los hallazgos subrayan la eficacia del modelo bosque aleatorio para clasificar con precisión las masas mamarias y brindan información valiosa sobre los predictores claves asociados con la malignidad. La investigación continua puede mejorar aún más el diagnóstico del cáncer de mama, lo que contribuye a mejores resultados para los pacientes. Las consideraciones éticas, incluida la privacidad de los datos y los modelos interpretables, refuerzan el uso responsable del aprendizaje automático en este campo crucial.

REFERENCIAS

[1] L.- Barranco, A. Rob, and C. Arquitectura, “Clasificación De Tumores En Cáncer De Mama Basado En Redes Neuronales De Convolución,” *V Jorn. Investig. y postgrado la Esc. Politécnica Super. Sevilla (2019), pp. 87-94.*, 2019.

[2] J. Galarza, “Reducción de dimensionalidad en Machine Learning Diagnóstico de cáncer de mama basado en datos genómicos y de imagen.,” *Univ. Politec. Val.*, 2017.

[3] C. A. Madrigal-González, R. Prada-Vásquez, and D. S. Fernández-McCann, “Detección Automática de Microcalcificaciones en una Mamografía Digital, Usando Técnicas de Inteligencia Artificial,” *TecnoLógicas*, 2013.

[4] Y. Wei, M. Gao, J. Xiao, C. Liu, Y. Tian, and Y. He, “Research and Implementation of Cancer Gene Data Classification Based on Deep Learning,” *J. Softw. Eng. Appl.*, vol. 16, no. 06, 2023.

[5] D. Vetrithangam, P. Shruti, B. Arunadevi, R. Himabindu, P. Naresh Kumar, and A. Ramesh Kumar, “OPTIMUM FEATURE SELECTION BASED BREAST CANCER PREDICTION USING MODIFIED LOGISTIC REGRESSION MODEL,” *J. Theor. Appl. Inf. Technol.*, vol. 101, no. 8, 2023.

[6] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, “A support vector machine-based ensemble algorithm for breast cancer diagnosis,” *Eur. J. Oper. Res.*, vol. 267, no. 2, 2018.

[7] Y. C. Cujilan *et al.*, “Supervised Learning Techniques for the Optimization of Diagnosis Processes of Diabetes in Public Health Centers,” in *Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology*, 2022, vol. 2022-July.

[8] D. Patiño Perez and R. Silva Bustillo, “X-Ray Images Analysis by Medium Artificial Neural Network,” *ECUADORIAN Sci. J. VOL. 5*, 2021.

[9] D. P. Pérez, R. S. Bustillos, C. M. Mora, and M. Botto-Tobar, “Prediction of Covid19 with the use of Random Forests Algorithm and Artificial Neural Networks,” *Ecuadorian Sci. J.*, vol. 4, no. 2, pp. 101–110, Sep. 2020.

[10] N. Wu *et al.*, “Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle,” *Kaggle*, vol. 4, no. November, 2019.

[11] S. Castrillon Osorio, L. M. Giraldo Marín, H. H. Jaramillo Villegas, and C. C. Piedrahita Escobar, “Machine learning aplicado en la clasificación y predicción de la depresión: Una revisión sistemática,” *Rev. Ibérica Sist. e Tecnol. Informação*, 2022.

[12] S. Xu and Z. Pan, “A novel ensemble of random forest for assisting diagnosis of Parkinson’s disease on small handwritten dynamics dataset,” *Int. J. Med. Inform.*, vol. 144, 2020.

[13] D. Patiño Pérez, R. Silva Bustillos, C. Munive Mora, and M. Botto-Tobar, "Predicción de Covid19 con el uso del Algoritmo Random Forest y Redes Neuronales Artificiales," *Ecuadorian Sci. J.*, vol. 4, no. 2, 2020.

[14] I. Tougui, A. Jilbab, and J. El Mhamdi, "Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications," *Healthc. Inform. Res.*, vol. 27, no. 3, 2021.

[15] J. Díaz-Ramírez, "Aprendizaje Automático y Aprendizaje Profundo," *Ingeniare. Rev. Chil. Ing.*, vol. 29, no. 2, 2021.

[16] Y. Wei, M. Li, and B. Xu, "Research on Establish an Efficient Log Analysis System with Kafka and Elastic Search," *J. Softw. Eng. Appl.*, vol. 10, no. 11, 2017.

[17] D. P. Perez, R. S. Bustillos, M. Botto-Tobar, and C. M. Mora, "X-Ray Images Analysis by Medium Artificial Neural Network," *Ecuadorian Sci. J.*, vol. 5, no. 1, pp. 55–60, Mar. 2021.

[18] X. Yang, R. Bist, S. Subedi, and L. Chai, "A deep learning method for monitoring spatial distribution of cage-free hens," *Artif. Intell. Agric.*, vol. 8, 2023.

[19] S. Subedi, R. Bist, X. Yang, and L. Chai, "Tracking floor eggs with machine vision in cage-free hen houses," *Poult. Sci.*, vol. 102, no. 6, 2023.

[20] X. Yang, L. Chai, R. B. Bist, S. Subedi, and Z. Wu, "A Deep Learning Model for Detecting Cage-Free Hens on the Litter Floor," *Animals*, vol. 12, no. 15, 2022.