Ciencia Latina Revista Científica Multidisciplinar, Ciudad de México, México.
ISSN 2707-2207 / ISSN 2707-2215 (en línea), septiembre-octubre 2025,
Volumen 9, Número 5.

https://doi.org/10.37811/cl_rcm.v9i5

APLICACIÓN DE MACHINE LEARNING EN EL DIAGNÓSTICO DEL CÁNCER DE MAMA: UN ENFOQUE BASADO EN BUENAS PRÁCTICAS

APPLICATION OF MACHINE LEARNING IN BREAST CANCER DIAGNOSIS: A BEST PRACTICES APPROACH

Carmen Liliana Rodriguez Paez
Universidad Autónoma del Estado de México

Ricardo Rico Molina
Universidad Autónoma del Estado de México

Mariam Juárez González
Universidad Autónoma del Estado de México

Jesus Dario Botello Jaime
Universidad Autónoma del Estado de México

DOI: https://doi.org/10.37811/cl_rcm.v9i5.20705

Aplicación de Machine Learning en el Diagnóstico del Cáncer de Mama: Un Enfoque Basado en Buenas Prácticas

Carmen Liliana Rodriguez Paez[1]
clrodriguezp@uaemex.mx
https://orcid.org/0000-0002-3856-0797
Universidad Autónoma del Estado de México
México

Ricardo Rico Molina
rricom@uaemex.mx
https://orcid.org/0000-0001-9586-8758
Universidad Autónoma del Estado de México
México

Mariam Juárez González
mjuarezg007@alumno.uaemex.mx
https://orcid.org/0009-0002-3766-2094
Universidad Autónoma del Estado de México
México

Jesus Dario Botello Jaime
Jbotelloj2200@alumno.ipn.mx
https://orcid.org/0009-0003-6375-7166
Instituto Politecnico Nacional
México

RESUMEN

El presente artículo tiene como objetivo identificar y describir buenas prácticas en la construcción de modelos de clasificación aplicados al diagnóstico de cáncer de mama, utilizando el conjunto de datos *Breast Cancer Wisconsin (Diagnostic)*. Basado en la revisión de estudios recientes y en la aplicación práctica de técnicas de modelado, se busca ofrecer una guía comprensible para personas que inician en el aprendizaje automático. La metodología se estructuró siguiendo el enfoque SEMMA (Sample, Explore, Modify, Model, Assess), que orientó la selección de variables, la exploración de datos y la validación de modelos. Se implementaron estrategias metodológicas como SelectKBest para la selección de características, validación cruzada anidada para asegurar una evaluación rigurosa, y optimización de hiperparámetros mediante Optuna. Además, se aplicaron procesos de calibración y ajuste de umbral para mejorar la confiabilidad de las predicciones.Los algoritmos analizados incluyeron modelos lineales, basados en árboles, máquinas de soporte vectorial (SVM), K-vecinos más cercanos (KNN), redes neuronales y métodos de ensamble, evaluados con métricas como MCC, AUC-ROC y Brier Score. Los resultados destacaron CatBoost por su discriminación y calibración (AUC-ROC y AUC-PR cercanas a 1, Brier bajo), SVM con ponderación de clases por su equilibrio (F1 y MCC elevados) y XGBoost por su robustez general.

Palabras clave: machine learning, cáncer de mama, diagnóstico asistido por computadora, modelos de clasificación, metodología SEMMA

[1] Autor principal
Correspondencia: clrodriguezp@uaemex.mx

Application of Machine Learning in Breast Cancer Diagnosis: A Best Practices Approach

ABSTRACT

The purpose of this article is to identify and describe best practices in the construction of classification models applied to breast cancer diagnosis, using the Breast Cancer Wisconsin (Diagnostic) dataset. Based on a review of recent studies and the practical application of modeling techniques, it seeks to provide a comprehensive guide for those new to machine learning. The methodology was structured following the SEMMA (Sample, Explore, Modify, Model, Assess) approach, which guided the selection of variables, data exploration, and model validation. Methodological strategies such as SelectKBest were implemented for feature selection, nested cross-validation to ensure rigorous evaluation, and hyperparameter optimization using Optuna. In addition, calibration and threshold adjustment processes were applied to improve the reliability of predictions. The algorithms analyzed included linear models, tree-based models, support vector machines (SVM), K-nearest neighbors (KNN), neural networks, and ensemble methods, evaluated with metrics such as MCC, AUC-ROC, and Brier Score. The results highlighted CatBoost for its discrimination and calibration (AUC-ROC and AUC-PR close to 1, low Brier), SVM with class weighting for its balance (high F1 and MCC), and XGBoost for its overall robustness.

***Keywords*:** machine learning, breast cancer, computer-assisted diagnosis, classification models, SEMMA methodology

Artículo recibido 26 setiembre 2025
Aceptado para publicación: 29 octubre 2025

INTRODUCCIÓN

A nivel mundial el cáncer representa uno de los principales problemas de salud pública, con un aumento constante de acuerdo con la Organización Panamericana de la Salud (2025), se proyecta en un 60% para el 2045 en la región de la Américas, pasando de 4.2 millones de casos registrados en 2022 a una estimación de 6.7 millones de casos, de igual forma el cáncer de mama es la neoplasia más común en mujeres (Palmero et al., 2021; Pérez-Herrero et al., 2023; Sung et al., 2021), con más de 2.3 millones de casos registrados a nivel mundial desde el 2022, representando el 25% de todos los tipos de cáncer que existen de acuerdo con cifras de la Organización Mundial de la Salud (2025). En México, las estadísticas son igualmente altas, durante el año 2024 el Instituto Nacional de Estadística y Geografía, da a conocer que la mortalidad a causa de tumores malignos es 89,633, siendo el 9% equivalente a 9,034 muertes ocasionados por cáncer de mama. Por consiguiente, la detección temprana es clave, ya que aumenta las posibilidades de éxito con tratamientos menos invasivos (Díaz et al., 2024). En este campo, los sistemas de diagnóstico asistido por computadora (CAD) han avanzado de forma notable ya que permiten analizar volúmenes de información, extraer patrones ocultos y formular modelos predictivos, lo que ha conllevado a la exploración de interacciones entre múltiples variables y la predicción de enfermedades a partir de datos históricos, como resultado pueden ser instrumentos muy útiles en ciertas partes del proceso de detección y diagnóstico del cáncer donde se requiere clasificar información y encontrar conocimiento en grandes volúmenes de datos (Aljuaid et al., 2022; Zaalouk et al., 2022). En este punto destaca el Breast Cancer Wisconsin (Diagnostic), uno de los conjuntos de datos más utilizados para evaluar algoritmos de clasificación (Wolberg et al., 1993), empleado en investigaciones que van desde modelos simples como la regresión logística y las máquinas de soporte vectorial, hasta enfoques complejos como árboles de decisión, técnicas de ensemble learning y redes neuronales profundas (Battineni et al., 2020; Darapureddy & Suman, 2024). A partir de estas investigaciones, se ha observado que el rendimiento de un modelo depende no solo del algoritmo, sino también de factores como el tamaño de la muestra, la normalización de datos, la selección de características relevantes y el ajuste de hiperparámetros. Aunque algunos trabajos reportan precisiones superiores al 90% (Yan et al., 2020), suelen lograrse bajo condiciones muy específicas, como el uso de un único modelo entrenado de forma intensiva o preprocesamientos diseñados a la medida.

A diferencia de investigaciones que se enfocan en un único modelo, este trabajo busca ofrecer una guía dirigida a personas de distintas áreas que inician en el análisis de datos y en machine learning, explicando conceptos y técnicas de manera que cualquier lector pueda comprenderlos, reproducir el proceso y adaptarlo a sus propios contextos. Para ello, se presenta un comparativo de múltiples clasificadores aplicados al data set señalado, siguiendo un marco metodológico unificado y aplicando buenas prácticas identificadas en la literatura reciente**.**

MATERIALES Y METODOS

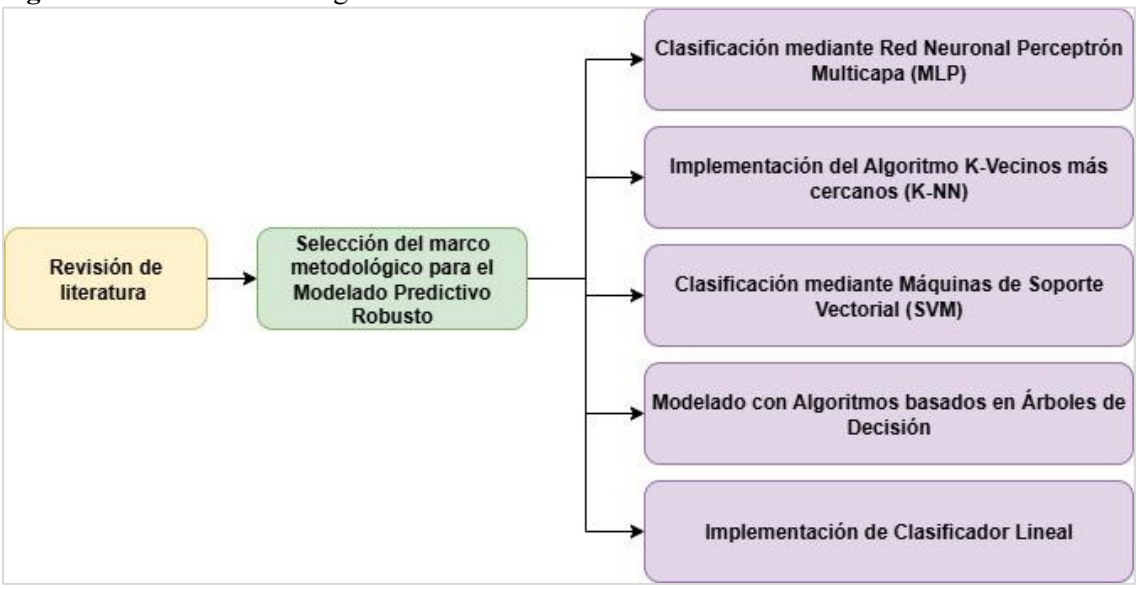
La investigación se desarrolló siguiendo un enfoque cualitativo y comparativo. En una primera etapa se realizó una revisión de la literatura reciente en bases de datos cientificas, a partir de la cual se seleccionaron artículos relevantes para la temática. Estos trabajos permitieron identificar buenas prácticas metodológicas aplicables al desarrollo de modelos clasificatorios en datos estructurados y conjuntos de tamaño reducido, como el Breast Cancer Wisconsin (Diagnostic) (WDBC) obtenido de la plataforma electrónica Kaggle. Se empleó el conjunto WDBC. La variable objetivo distingue tumores benignos (B) y malignos (M). Para preservar la proporción de clases (≈63% B / 37% M), se aplicó una partición estratificada en entrenamiento (80%) y prueba (20%).

Para el desarrollo, de los modelos con estas buenas prácticas se utilizó SEMMA, metodología desarrollada por SAS Institute con el objetivo de estructurar proyectos de minería de datos, cuenta con cinco fases: Muestreo, Exploración, Modificación, Modelado y Evaluación (Gómez et al., 2017). Cada una de estas cumplen un papel específico dentro del ciclo de análisis, asegurando que los datos sean preparados, transformados, modelados y evaluados de forma congruente.

En la segunda etapa, las prácticas identificadas se integraron en un flujo de trabajo reproducible implementado en Python, que incluyó la normalización de los datos posterior a la división en entrenamiento y prueba para evitar data leakage, la selección de características relevantes mediante SelectKBest (criterio univariado), la validación cruzada repetida, anidada para mejorar la estabilidad de las estimaciones y reducir el sesgo optimista, la optimización de hiperparámetros se realizó con la herramienta Optuna, acotando espacios plausibles y usando AUC -ROC como objeto interno. El ajuste de umbral de decisión con base en métricas objetivo y la calibración de probabilidades para modelos seleccionados.

Siguiendo este marco metodológico, se entrenaron y evaluaron diferentes tipos de algoritmos clasificadores, entre ellos modelos lineales como Regresión Logística (con y sin balanceo de clases); modelos basados en árboles como Árbol de Decisión (criterio entropy, con y sin balanceo), Random Forest, Gradient Boosting, XGBoost, LightGBM y CatBoost; máquinas de soporte vectorial (SVM, con núcleo no lineal, con y sin balanceo de clases); métodos basados en instancias como k-vecinos más cercanos (KNN); redes neuronales como Perceptrón Multicapa (MLP) de arquitectura poco profunda; y modelos de ensamble como combinaciones Voting Classifier y Stacking Classifier (Figura 1). Este diseño aseguró que todos los modelos fueran evaluados bajo las mismas condiciones experimentales, garantizando la comparabilidad y la posibilidad de reproducir los resultados por parte de otros investigadores.

Figura 1. Método de investigación



RESULTADOS Y DISCUSIÓN

Se identificaron varias prácticas metodológicas para lograr resultados sólidos y reproducibles en el análisis del conjunto de datos Breast Cancer Wisconsin (Diagnostic) en la clasificación de cáncer de mama. La primera de ellas, es la normalización posterior a la partición de los datos, aplicada únicamente sobre el conjunto de entrenamiento. Esto significa que los valores de las variables se ajustan a la misma escala, pero solo después de separar los datos en los grupos de entrenamiento y prueba. El objetivo es evitar algo que se mirara en varios espacios, data leakage o “fuga de datos”, esto ocurre cuando la

información del grupo de prueba se filtra hacia el proceso de entrenamiento, lo que puede llevar a que el modelo parezca más preciso de lo que realmente es cuando se aplique en nuevos casos (Battineni et al., 2020; Sung et al., 2021; Zaalouk et al., 2022).

Otra práctica clave implementada fue la validación cruzada anidada y repetida (rnCV), método que consiste en dividir el conjunto de entrenamiento en múltiples partes y entrenar el modelo varias veces con distintas combinaciones, de manera que cada parte actúe en algún momento como validación. En esta hay dos niveles: uno interno para ajustar los parámetros del modelo y otro externo para evaluar su rendimiento. Al repetir este ciclo varias veces, se reduce el sesgo y se obtiene una estimación más estable y confiable del rendimiento del modelo (Díaz et al., 2024; Wolberg et al., 1993; Hussain et al., 2024).

La tercera característica señalada son técnicas como SelectKBest, método que se encarga de analizar todas las variables del conjunto de datos y asigna una puntuación a cada una según su relación con la variable objetivo, en este caso si un tumor es benigno o maligno. A partir de estas puntuaciones, conserva únicamente las variables con mejor capacidad predictiva y descarta aquellas cuya aportación al modelo es mínima o poco relevante. Utilizar el mismo subconjunto de variables para todos los modelos permite que las comparaciones sean justas y que el rendimiento se evalúe bajo condiciones idénticas, evitando que un modelo tenga ventaja por haber usado información distinta (Aljuaidet al., 2022; Kumar et al., 2022).

En cuanto al ajuste de los modelos, el uso de Optuna para la búsqueda de hiperparámetros se mostró especialmente útil. Estos últimos son las configuraciones internas de cada modelo y que tienen un rol en los resultados obtenidos. Lo que hace Optuna es automatizar la búsqueda de las mejores combinaciones posibles, explorando de forma inteligente un espacio muy amplio de opciones y maximizando métricas como el AUC-ROC en un número reducido de intentos. Esta métrica AUC-ROC, cuyo nombre es Área Bajo la Curva de Característica Operativa del Receptor, evalúa la capacidad de un modelo para diferenciar entre clases, en conjunto de datos que se está manejando si el tumor es benigno o llega a ser maligno, esto lo hace en distintos umbrales de decisión: un valor de 1 indica una discriminación entre las perfecta, mientras que un valor de 0.5 refleja un desempeño equivalente al azar (Wolberg et al., 1993; Boddu & Jan, 2025; Yan et al., 2020).

Para la evaluación de resultados se recomienda el uso de métricas robustas como el Matthews Correlation Coefficient (MCC), particularmente valiosa cuando las clases no están perfectamente equilibradas, ya que considera tanto los aciertos como los errores de cada clase, ofreciendo una visión más justa que métricas como la exactitud (Accuracy) o el F1-score para el balance (Díaz et al., 2024). Esto resulta especialmente funcional en conjunto de datos como el Breast Cancer Wisconsin (Diagnostic) donde el 62,74 % son Benigno y 37,26 % el Maligno. Tener en cuenta que técnicas como el sobre muestreo artificial (oversampling). En conjuntos equilibrados, este tipo de técnicas generan datos sintéticos basados en los datos ya existentes, pero que no representan patrones reales. En lugar de mejorar el aprendizaje, pueden llevar al modelo a aprender información irrelevante o engañosa es decir "introducir ruido", incrementando el riesgo de sobreajuste y reduciendo su capacidad de generalizar a datos nuevos (Battineni et al., 2020; Darapureddy & Suman, 2024; Hussain et al 2024).

La literatura subraya la importancia de probar múltiples paradigmas de aprendizaje automático desde modelos lineales hasta redes neuronales y algoritmos basados en árboles de decisión para capturar patrones complementarios y evitar depender de un único enfoque (SAS Institute Inc, 2023; Sung et al., 2021; Tuerhong et al., 2023).

En el muestreo, el código carga el conjunto de datos, después se eliminan columnas irrelevantes, fijándose el índice y transformando la variable objetivo (B o M) en formato numérico. Después se imprime el tamaño del conjunto de datos y la distribución de clases, para verificar la representatividad de la muestra, luego se realiza una partición estratificada en entrenamiento y prueba (80/20) para conservar la proporción de casos B y M para asegurarse que ambos subconjuntos mantengan la estructura de clases del conjunto original. El primer conjunto se utiliza para ajustar los modelos y que aprendan los patrones de los datos, mientras que el segundo se utiliza para evaluar su desempeño con información nueva y que el modelo no ha visto antes, permitiendo simular de esta forma comportamiento en un escenario real.

En la fase de exploración, el script revisa rápidamente las dimensiones del conjunto de datos y la distribución de clases, con el fin de confirmar que los tamaños y proporciones son adecuados antes de iniciar cualquier transformación o modelado. Esta inspección inicial permite detectar posibles problemas y comprender mejor las características de los datos con los que se trabajará.

La fase de modificación se desarrolla a través de un pipeline (secuencia automatizada de pasos que procesa los datos y entrena el modelo de forma ordenada y reproducible) de preprocesamiento que incluye la imputación de valores faltantes con la mediana y la normalización de los datos (Kumar, 2018), esto último es un proceso que ajusta los valores de las variables a la misma escala para que ninguna tenga más influencia que otra en el entrenamiento del modelo y se puede aplicar solo a variables numéricas. Es relevante que el ajuste de este pipeline se realiza solo con el conjunto de entrenamiento y posteriormente se aplica al de prueba, evitando el data leakage. Además, dentro de cada modelo se incorpora un paso de selección de características mediante la técnica SelectKBest.

En la fase de modelado, el código implementa un conjunto variado de clasificadores, cada uno integrado en un pipeline con su correspondiente selector de características. Se incluyen modelos lineales (Regresión Logística, con y sin balanceo), basados en árboles (Árbol de Decisión, Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost), máquinas de soporte vectorial con núcleo no lineal (con y sin balanceo), métodos por instancias (KNN), redes neuronales (Perceptrón Multicapa) y ensambles (Voting y Stacking Classifier). Para cada modelo se define un espacio de hiperparámetros y se optimiza automáticamente con Optuna, empleando validación cruzada estratificada y AUC-ROC como métrica objetivo, dentro de un esquema de validación cruzada anidada. Los mejores modelos se calibran y, cuando es posible, se combinan en un ensamble suave tipo Voting, que consiste en integrar las predicciones de varios clasificadores y tomar la decisión final según la votación conjunta, ya sea por mayoría de clases predichas o promediando las probabilidades, con el fin de aprovechar las fortalezas de cada modelo y mejorar el rendimiento global.

En la evaluación, se revisa qué tambien funciona el modelo usando el grupo de datos que se reservó para la prueba, antes de hacerlo se ajusta el “punto de corte” o umbral de decisión, que es el valor a partir del cual el modelo decide si un caso es positivo o negativo. Para comprender mejor las métricas calculadas, es fundamental entender la matriz de confusión la cual es una herramienta que permite evlauar el desempeño de un modelo y predecir cada clase del conjunto de datos pruebas (Cotrina et al., 2024) con sus cuatro componentes: los verdaderos positivos (TP) son casos que son evaluados adecuadamente como positivos; los verdadero negativos (TN) son casos que son evaluados correctamente como negativos; los falsos positivos (FP) son casos que fueron evaluados como positivos

pero en realidad son negativos; y los falsos negativos (FN) son casos que fueron evaluados como negativos pero en realidad son positivos. Estas categorias forman la base para el cálculo de todas las métricas de evaluación utilizadas en este estudio.

La exactitud (accuracy), mide el porcentaje total de predicciones correctas sobre el conjunto evaluado, es decir, de todas las prediciones realizadas por los diferentes modelos cuántas fueron correctass. Este valor va de 0 a 1 (0%-100%), entre mas cercano este al 1, mejor sera el modelo. La precisión (precision), que indica de todas las predicciones positivas cuántas son realmente correctas, ayudando a controlar los falsos positivos. Se enfoca en los TP y los FP, indicando cuantos casos feuron realmente positivos. La sensibilidad (recall), que refleja la capacidad del modelo para detectar correctamente los casos positivos reales, reduciendo falsos negativos. El F1-score es la media armónica entre precisión y recall, útil cuando se busca un equilibrio entre ambas. Se considera que entre más cercano este a 1 es mejor y al 0 es peor.

También se calcula el Matthews Correlation Coefficient (MCC), que evalúa la correlación entre las predicciones y los valores reales considerando todas las categorías de la matriz de confusión, siendo más robusto ante desbalances. Su rango de valores va de -1 a 1. Entre mas cercano a 1 se dice que está mas correlacionado o que su valor es más cercano a la realida; si tiende más a 0 quiere decir que su predicción es más aleatoria. La especificidad (specificity) mide la capacidad de identificar correctamente los casos negativos, evitando falsos positivos.

El Area Under the Precision-Recall Curve (AUC-PR) resume la relación entre precisión y recall a distintos umbrales de decisión, siendo especialmente relevante en conjuntos desbalanceados. Finalmente, el Brier Score evalúa la calidad de las probabilidades predichas mediante el cálculo del error cuadrático medio entre estas y los valores reales, de modo que valores más bajos indican una mejor calibración del modelo (Tabla 1).

Tabla 1. Comparación de los modelos

Modelo	Accuracy	Precision	Recall	F1	MCC	Specificity	AUC-ROC	AUC-PR	Brier
LogisticRegression_sinCW	0,9663	0,9683	0,9789	0,9732	0,9294	0,9451	0,9931	0,9957	0,0259
LogisticRegression_conCW	0,97	0,9693	0,9836	0,9762	0,9364	0,9471	0,9931	0,9957	0,0297
RandomForest_sinCW	0,9538	0,9578	0,9696	0,9633	0,9025	0,9275	0,9921	0,9954	0,0341
RandomForest_conCW	0,9502	0,9578	0,9637	0,9604	0,8946	0,9275	0,9924	0,9955	0,0338
XGBoost	0,9714	0,9714	0,9836	0,9773	0,9394	0,951	0,9927	0,9957	0,027
LightGBM	0,9626	0,9603	0,9813	0,9705	0,9208	0,9314	0,9914	0,9947	0,0298
SVM_sinCW	0,967	0,9728	0,9754	0,9737	0,9307	0,9529	0,9957	0,9974	0,0302
SVM_conCW	0,9714	0,9706	0,9848	0,9774	0,9394	0,949	0,9927	0,9956	0,0303
KNN	0,9634	0,9659	0,9766	0,971	0,9223	0,9412	0,9881	0,9925	0,0323
MLP	0,9707	0,9664	0,9883	0,977	0,9383	0,9412	0,9914	0,9943	0,033
GradientBoosting	0,9575	0,9552	0,9789	0,9665	0,9103	0,9216	0,9957	0,9975	0,0307
DecisionTree_sinCW	0,9297	0,9343	0,9579	0,9449	0,8521	0,8824	0,9802	0,9865	0,0628
DecisionTree_conCW	0,9377	0,9362	0,9673	0,9512	0,8671	0,8882	0,9785	0,9789	0,0362
CatBoost	0, 9678	0,9664	0,9836	0,9746	0,9319	0,9412	0,9977	0,9986	0,0175

La Tabla 1 muestra los resultados de aplicar buenas prácticas metodológicas, donde la mayoria de los modelos presentan un desempeño sobresaliente. En términos de exactitud (Accuracy), todos los modelos alcanzaron valores superiores a 0.9, lo que indica que más del 90% de las predicciones fueron correctas. Los modelos con mejor desempeño superaron el 0.96, mientras que incluso aquellos con menor rendimiento se mantuvieron por encima de 0.90, evidenciando que la metodología implementada garantiza una alta capacidad predictiva general independientemente del algoritmo seleccionado. Esta consistencia en el desempeño superior contrasta con estudios previos donde la falta de buenas prácticas genera una mayor variabilidad entre modelos.

En términos de discriminación, CatBoost fue el más destacado (AUC-ROC y AUC-PR cercanas a 1) y, además, presentó el Brier score más bajo del conjunto, lo que indica excelente calibración. El SVM con ponderación de clases (class weight) y XGBoost alcanzaron los valores más altos de MCC, reflejando una correlación muy fuerte entre predicciones. El SVM sin ponderación obtuvo la mayor especificidad, útil para minimizar falsos positivos. En contraste, el árbol de decisión simple (con o sin ponderación) mostró los valores relativos más bajos dentro de la cohorte y sirve como línea base.

Lectura de métricas clave

Exactitud (Accuracy). Aun cuando la mayoría de modelos superaron el 90%, accuracy por sí sola puede ser engañosa bajo desbalance moderado; se privilegió su lectura en conjunto con MCC y AUC-PR.

Precisión y sensibilidad. La primera controla falsos positivos; la segunda, falsos negativos. El equilibrio se sintetizó en F1, en el que MLP y SVM con ponderación se ubicaron entre los mejores. Los colores en esta columna indica en verde más intenso cual es mejor y en rojo cual es menos mejor.

MCC. El coeficiente osciló en valores altos para los mejores modelos; XGBoost y SVM con ponderación alcanzaron los máximos observados con 0.9394 en verde intenso y el de menor valor fue el DecisionTree_conCW con 0.8521 en rojo.

Especificidad (specificity). El SVM sin ponderación lideró con respecto a los demás métodos. En la tabla se observa que el mejor modelo evaluado fue SVM_sinCW con 0.9529 en verde intenso y el más bajo fue DecissionTree_sinCW con 0.8824 en rojo.

AUC-ROC y AUC-PR. Ambas cercanas a 1 para CatBoost; AUC-PR. En este caso el mejor modelo fue el CatBoost con 0.9977 en verde intenso y el que obtuvo el menos valor fue el DecisionTree_conCW con 0.9785 aún muy bueno.

Brier score. CatBoost obtuvo el valor mínimo, lo que respalda la confiabilidad de sus probabilidades. Un Brier bajo en conjunto con AUC alta sugiere que el modelo no solo separa bien, sino que «sabe cuán seguro está». En el caso de la tabla a diferencia de las demás la mejor es el modelo CatBoost con 0.0175 y el más bajo es el modelo DecisionTree_sinCW con 0.0628.

En la descripción se citaron los mejores y peores modelos de acuerdo a las diferentes evaluaciones, es de aclarar que todos los resultados de las evaluaciones son buenos, solo hay algunas con valores más bajos. Lo anterior se debe a el objetivo del artículo que es la aplicación de modelos usando buenas prácticas. Lo cual queda demostrado en esta tabla en donde todos los resultados son buenos en mayor o menor medida.

CONCLUSIONES

Lo anterior permite mostrar una serie de buenas prácticas para la generación de modelos, como es la normalización posterior a la partición de datos, la selección de características, la validación cruzada anidada, la optimización de hiperparámetros y la calibración de probabilidades, que permite generar acompañada de lo observación de diversas métricas una visión más rica y equilibrada de los rendimientos de los modelos, que basarlos en una solo métrica, permitiendo analizar el comportamiento de las diferentes clases de modelos desde múltiples perspectivas, identificando fortalezas y limitaciones. En el análisis realizado, CatBoost ofreció la mejor combinación de discriminación y calibración; SVM con ponderación y XGBoost mostraron el mayor equilibrio global (MCC/F1); y los árboles simples quedaron como referencia de base.

REFERENCIAS BIBLIOGRAFICAS

Aljuaid, H., Alturki, N., Alsubaie, N., Cavallaro, L., & Liotta, A. (2022). Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. *Computer Methods and Programs in Biomedicine*, *223*, 106951. https://doi.org/10.1016/j.cmpb.2022.106951

Battineni, G., Chintalapudi, N., & Amenta, F. (2020) Performance analysis of different machine learning algorithms in breast cancer predictions. *EAI Endorsed Transactions on Pervasive Health and Technology* 6(23), e4. https://doi.org/10.4108/eai.28-5-2020.166010

Boddu, A. S., & Jan, A. (2025). A systematic review of machine learning algorithms for breast cancer detection. *Tissue & cell*, *95*, 102929. https://doi.org/10.1016/j.tice.2025.102929

Cotrina-Teatino, M. A., Riquelme, A. I., Guartan, J. A., & Marquina, J. J. (2025). Machine Learning aplicado a la exploración minera usando matriz de confusión. *Sciéndo Ingenium*, *21*(1), 63-74. https://doi.org/10.17268/rev.cyt.2025.01.06

Darapureddy, N., & Suman, K. (2024). Performance Analysis and Comparison of Machine Learning Algorithms for Breast Cancer Dataset. *Contemporary Perspective on Science, Technology and Research 6*, 89–99. https://doi.org/10.9734/bpi/cpstr/v6/7561E

Díaz, O., Rodríguez-Ruíz, A., & Sechopoulos I. (2024). Artificial Intelligence for breast cancer detection: Technology, challenges, and prospects. *European Journal of Radiology, 175*, 111457.

https://doi.org/10.1016/j.ejrad.2024.111457

Gomez, H., Jiménez, R., Hernández, G., & Martinez, Á. (2017). A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change. *Advances in Science, Technology and Engineering Systems Journal*, *2*(3), 598-604. https://doi.org/10.25046/aj020376

Hussain, S., Ali, M., Naseem, U., Nezhadmoghadam, F., Jatoi, M. A., Gulliver, T. A., & Tamez-Peña, J. G. (2024). Breast cancer risk prediction using machine learning: a systematic review. *Frontiers in oncology*, *14*, 1343627. https://doi.org/10.3389/fonc.2024.1343627

Instituto Nacional de Estadística y Geografía. (2024). *Estadísticas a propósito del día internacional de la lucha contra el cáncer de mama*. https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2024/EAP_LuchaCMama24.pdf

Kumar, M., Singhal, S., Shekhar, S., Sharma, B., & Srivastava, G. (2022). Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning. *Sustainability*, *14*(21), 13998. https://doi.org/10.3390/su142113998

Kumar, V. H. (2018). Python libraries, development frameworks and algorithms for machine learning applications. *International Journal of Engineering Research & Technology (IJERT)*, *7*(04). https://scholar.google.com/scholar?as_q=Python+Libraries%2C+Development+Frameworks+and+Algorithms+for+Machine+Learning+Applications&as_occt=title&hl=en&as_sdt=0%2C31

Organización Mundial de la Salud. (2025*). Cáncer de mama*. https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer

Organización Panamericana de la Salud. (2025). *Cáncer.* https://www.paho.org/es/temas/cancer

Palmero, J., Lasar Rosenthal, J., Juárez, L., & Medina, C. (2021). Cáncer de mama: una visión general. *Acta médica Grupo Ángeles, 19*(3), 354-360. https://www.scielo.org.mx/scielo.php?pid=s1870-72032021000300354&script=sci_arttext

Pérez-Herrero, M., López-Alvarez, S., & Nebril, B. A. (2023). Factores perioperatorios en el cancer de mama. Revisión sistemática de su influencia en el pronóstico. *Revista de Senología y Patología Mamaria*, *36*(1), 100413. https://doi.org/10.1016/j.senol.2022.03.001

SAS Institute Inc. (2023). Introduction to SEMMA. SAS Enterprise Miner Documentation. https://documentation.sas.com/doc/en/emref/15.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209-249. https://doi.org/10.3322/caac.21660

Tuerhong, A., Silamujiang, M., Xianmuxiding, Y., Wu, L., & Mojarad, M. (2023). An ensemble classifier method based on teaching-learning-based optimization for breast cancer diagnosis. *Journal of cancer research and clinical oncology*, *149*(11), 9337–9348. https://doi.org/10.1007/s00432-023-04861-5

Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). Breast Cancer Wisconsin (Diagnostic) [Dataset]. *UCI Irvine Machine Learning Repository.* https://doi.org/10.24432/C5DW2B

Yan, R.., Ren, F., Wang, Z., Wang, L., Zhang, T., Liu, Y., Rao, X., Zheng, C., & Zhang, F. (2020). Breast cancer histopathological image classification using a hybrid deep neural network. *Methods*, *173*, 52-60. https://doi.org/10.1016/j.ymeth.2019.06.014

Zaalouk, A., Ebrahim, G., Mohamed, H., Hassan, H., & Zaalouk, M. (2022). A Deep Learning Computer-Aided Diagnosis Approach for Breast Cancer. *Bioengineering*, *9*(8), 391. https://doi.org/10.3390/bioengineering9080391