



Amanda Razera – Universidad Estatal del Medio Oeste

Maiara Luiza Biava Miri – Centro Universitario Campo Real

Eduardo de Almeida Ravarena – Centro Universitario Campo Real

Gabryela Paulista Mateucci – Centro Universitario Campo Real

Camila Padilha Duda – Centro Universitario Campo Real

## Resumen

La aplicación de técnicas de aprendizaje automático en oncología ha permitido la integración de datos genómicos complejos para la predicción de resultados clínicos. En el cáncer de mama subtipo luminal B, la alta heterogeneidad tumoral representa un desafío para la estratificación pronóstica y la definición terapéutica. Este estudio tuvo como objetivo desarrollar modelos predictivos basados en variantes somáticas para la evaluación de la agresividad tumoral. Se implementó un proceso de aprendizaje automático utilizando algoritmos supervisados, incluyendo XGBoost, Máquinas de Vectores de Soporte y Redes Neuronales Artificiales. La selección de variables se realizó en función de su importancia predictiva, priorizando los genes con relevancia biológica. El rendimiento de los modelos se evaluó utilizando métricas como el área bajo la curva ROC, la sensibilidad, la especificidad y la puntuación F1. Los resultados demostraron una alta capacidad predictiva, destacando el modelo XGBoost (AUC = 0,88), seguido de las Redes Neuronales (AUC = 0,87) y las SVM (AUC = 0,85). El análisis de interpretabilidad indicó que genes como PIK3CA, TP53 y ERBB2 fueron los principales determinantes de las predicciones.

Los resultados refuerzan el potencial del aprendizaje automático para identificar patrones genómicos asociados con la agresividad tumoral, lo que contribuye a las estrategias de medicina de precisión.

Palabras clave: Aprendizaje automático. Cáncer de mama. Variantes somáticas. Pronóstico. Genómica.

## Abstracto

La aplicación de técnicas de aprendizaje automático en oncología ha permitido la integración de datos genómicos complejos para predecir resultados clínicos. En el cáncer de mama luminal B, la alta heterogeneidad tumoral representa un desafío importante para la estratificación pronóstica y la toma de decisiones terapéuticas. Este estudio tuvo como objetivo desarrollar modelos predictivos basados en variantes somáticas para evaluar la agresividad tumoral. Se implementó un proceso de aprendizaje automático utilizando algoritmos supervisados, incluyendo XGBoost, Máquinas de Vectores de Soporte y Redes Neuronales Artificiales. La selección de características se basó en la importancia predictiva, priorizando los genes biológicamente relevantes. El rendimiento del modelo se evaluó utilizando métricas como el área bajo la curva ROC, la sensibilidad, la especificidad y la puntuación F1. Los resultados demostraron un alto rendimiento predictivo, con XGBoost logrando los mejores resultados (AUC = 0,88), seguido de las Redes Neuronales (AUC = 0,87) y SVM (AUC = 0,85). El análisis de interpretabilidad reveló que genes como PIK3CA, TP53 y ERBB2 fueron los principales contribuyentes a las predicciones del modelo. Estos hallazgos resaltan el potencial de los enfoques de aprendizaje automático para identificar patrones genómicos asociados con la agresividad tumoral, lo que respalda las estrategias de medicina de precisión.

Palabras clave: aprendizaje automático, cáncer de mama, variantes somáticas, pronóstico, genómica

## 1. Introducción

El cáncer de mama sigue siendo una de las principales causas de morbilidad y mortalidad entre las mujeres. mujeres a nivel mundial, caracterizadas por una alta heterogeneidad biológica y clínica. molecular (SUNG et al., 2021; BRAZIL, 2024). La evolución del conocimiento en biología molecular Esto permitió la clasificación de los tumores en subtipos intrínsecos, como luminal A, luminal B, HER2-enriquecido y triple negativo, contribuyendo significativamente al refinamiento del pronóstico y terapéutico (PEROU et al., 2000; SØRLIE et al., 2001).

Entre estos, el subtipo Luminal B presenta una mayor agresividad clínica y un índice más elevado. proliferativo y con mayor riesgo de recurrencia en comparación con el subtipo Luminal A, lo que refleja su complejidad molecular y comportamiento clínico heterogéneo (PRAT et al., 2015; BURSTEIN et al., 2014). Esta heterogeneidad representa un desafío significativo en la práctica clínica, especialmente en lo cual se refiere a la estratificación pronóstica y a la definición de estrategias terapéuticas personalizadas.

Con el avance de las tecnologías de secuenciación de próxima generación, se ha hecho posible para identificar variantes somáticas relevantes en la carcinogénesis mamaria, especialmente en genes implicado en las vías de proliferación celular, reparación del ADN y señalización intracelular (CÁNCER) (GENOME ATLAS NETWORK, 2012). Genes como PIK3CA, TP53 y ERBB2 han sido ampliamente... descritos como determinantes de la progresión tumoral y la respuesta terapéutica, y considerados objetivos centrales en la oncología de precisión (ANDRÉ et al., 2019; SILWAL-PANDIT et al., 2017).

Sin embargo, la creciente disponibilidad de datos genómicos de alta dimensión hace necesario... Las limitaciones de los métodos estadísticos tradicionales dificultan la identificación de patrones complejos. asociado con resultados clínicos. En este escenario, las técnicas de aprendizaje automático han surgido como herramientas prometedoras que permiten el análisis integrado de múltiples variables y la construcción de modelos predictivos más robustos (KOURI et al., 2020; ESTEVA et al., 2019).

Estudios recientes demuestran que los algoritmos de aprendizaje automático son capaces de Integrar datos genómicos y clínicos para predecir el pronóstico, la respuesta terapéutica y el riesgo de progresión tumoral, con un rendimiento superior en comparación con los modelos convencionales en diferentes contextos. oncológicas (TOPOL, 2019; LIBBRECHT; NOBLE, 2015). Además, enfoques interpretables, Cómo los métodos basados en valores basados en SHAP han permitido una mayor transparencia en la comprensión a partir de la contribución individual de las variables, ampliando la aplicabilidad clínica de estos modelos. (LUNDBERG; LEE, 2017).

Dado este contexto, el desarrollo de modelos predictivos basados en variantes El uso de técnicas somáticas y de aprendizaje automático representa una estrategia prometedora para mejorar... Estratificación pronóstica en el cáncer de mama luminal B. Por lo tanto, el presente estudio propone la Construcción y evaluación de modelos de aprendizaje automático capaces de identificar patrones. La genómica está asociada a la agresividad tumoral, lo que contribuye al avance de la medicina de precisión.



Año VI, vol. 1 2026 | Envío: 15/03/2026 | Aceptado: 17/03/2026 | Publicación: 19/03/2026

## 2. Materiales y métodos

### 2.1 Diseño del estudio

Este es un estudio observacional con un enfoque analítico, en el que un Modelo predictivo basado en técnicas de aprendizaje automático para evaluar resultados clínicos. basado en firmas genómicas. La estrategia metodológica se estructuró de acuerdo con las directrices para estudios que involucran modelos predictivos en salud (COLLINS et al., 2015).

### 2.2 Proceso de aprendizaje automático

Se implementó un proceso de aprendizaje automático, que consta de etapas previas. Procesamiento, selección de variables, entrenamiento del modelo y evaluación del rendimiento. El preprocesamiento incluyó la estandarización de variables y el manejo de datos faltantes mediante Se utilizaron métodos de imputación apropiados, basados en la proximidad estadística. La selección de variables fue... Esto se logró mediante un análisis de importancia predictiva, que combina la regularización L1 y el análisis de importancia. permutacional, priorizando atributos con relevancia biológica.

### 2.3 Modelos predictivos

Se evaluaron tres algoritmos supervisados ampliamente utilizados en bioinformática. Potenciación de gradiente extremo (XGBoost), máquina de vectores de soporte (SVM) y redes neuronales artificiales (ARN). La optimización de hiperparámetros se realizó mediante búsqueda bayesiana con validación cruzada. La validación cruzada anidada tiene como objetivo reducir el riesgo de sobreajuste y aumentar la capacidad de generalización. de los modelos (BERGSTRA et al., 2011).

### 2.4 Evaluación del desempeño

El rendimiento de los modelos se evaluó de forma independiente y en conjunto utilizando las métricas. Área bajo la curva ROC (AUC), sensibilidad, especificidad, valor predictivo positivo (VPP), valor Valor predictivo negativo (VPN) y puntuación F1. Teniendo en cuenta el posible desequilibrio entre las clases, se utilizaron métricas adicionales. basado en precisión-exhaustividad.

### 2.5 Interpretabilidad de los modelos

La interpretabilidad de los modelos se evaluó utilizando la metodología SHAP (Shapley). Explicaciones aditivas), lo que permite cuantificar la contribución individual de las variables a la predicciones (LUNDBERG; LEE, 2017). Este enfoque permitió identificar los principales Determinantes genómicos asociados con los resultados analizados.

### 2.6 Análisis estadístico y software

Los análisis se realizaron en un entorno computacional utilizando el lenguaje R (versión).

Año VI, vol. 1 2026 | Envío: 15/03/2026 | Aceptado: 17/03/2026 | Publicación: 19/03/2026  
4.2.1) y Python (versión 3.9), utilizando bibliotecas específicas para modelado estadístico y aprendizaje automático.

Los resultados se consideraron estadísticamente significativos cuando  $p < 0,05$ , con  
Aplique la corrección para comparaciones múltiples cuando corresponda.

### 3. Resultados y discusión

Aplicación de modelos de aprendizaje automático para el análisis de firmas genómicas.  
Demostró una alta capacidad predictiva para identificar patrones asociados con la agresividad.  
Tumor en cáncer de mama subtipo luminal B. Entre los algoritmos evaluados, el modelo basado en  
En Extreme Gradient Boosting (XGBoost) se observó el mejor rendimiento general, con AUC-ROC  
0,88, seguido de Redes Neuronales Artificiales (0,87) y Máquina de Vectores de Soporte (0,85), según  
presentado en la Tabla 1.

Este rendimiento superior de XGBoost puede atribuirse a su capacidad para modelar  
interacciones no lineales y captura de relaciones complejas entre variables de alta dimensión.  
una característica intrínseca de los datos genómicos. Además, los algoritmos basados en conjuntos tienden a...  
exhibiendo mayor robustez frente al ruido y la variabilidad biológica, aspectos que son frecuentemente  
observado en estudios sobre cáncer (TOPOL, 2019).

Tabla 1 – Rendimiento de los modelos de aprendizaje automático.

Modelo	AUC-ROC	Sensibilidad	Especificidad	VPP	VPN	Puntuación F1
XGBoost	0,88	0,82	0,83	0,81	0,84	0,82
ARN	0,87	0,81	0,82	0,80	0,83	0,81
SVM	0,85	0,79	0,80	0,78	0,81	0,79

Fuente: Autores (2026).

Mantener valores equilibrados de sensibilidad y especificidad entre los modelos.  
sugiere una buena capacidad discriminadora y estabilidad predictiva, incluso en escenarios de potencial  
desequilibrio de clases. Este hallazgo es particularmente relevante en oncología, donde los eventos  
Los eventos clínicos adversos suelen distribuirse de forma desigual.

El análisis de la importancia de las variables mostró que genes como PIK3CA, TP53 y  
ERBB2 fueron los principales determinantes de las predicciones, como se muestra en la Tabla 2.



Año VI, vol. 1 2026 | Envío: 15/03/2026 | Aceptado: 17/03/2026 | Publicación: 19/03/2026

Tabla 2 – Importancia de las variables (análisis SHAP).

Importancia relativa variable (SHAP)	Función biológica principal
PIK3CA 0,32	Vía PI3K/AKT/mTOR – proliferación celular
TP53 0,28	Control del ciclo celular y la apoptosis
ERBB2 0,25	Señalización del crecimiento tumoral
Ki-67 0,20	Índice proliferativo
TFRC 0,15	Metabolismo del hierro y crecimiento tumoral

Fuente: Autores (2026).

La relevancia de estas variables refuerza la coherencia biológica del modelo, ya que estas

Se reconoce ampliamente que los genes desempeñan un papel fundamental en la carcinogénesis y la resistencia al cáncer de mama. terapéutico (CANCER GENOME ATLAS NETWORK, 2012; ANDRÉ et al., 2019).

Además, la contribución del índice proliferativo (Ki-67) resalta la importancia de Integración entre datos genómicos y parámetros fenotípicos, ampliando la capacidad predictiva de modelos. Este hallazgo coincide con la evidencia que apunta a Ki-67 como marcador pronóstico. relevante en el subtipo Luminal B (PRAT et al., 2015).

El análisis integrado también permitió identificar patrones genómicos asociados con diferentes niveles de riesgo, destacando que la interacción entre variantes es crucial para el comportamiento. tumoral (Tabla 3).

Tabla 3 – Patrones genómicos asociados con el riesgo previsto.

Patrón genético	Interpretación biológica	Asociación con de agresividad
Se ha aislado una mutación en el gen PIK3CA que activa una vía proliferativa.		Moderado
ERBB2 modificado	Señalización de crecimiento aumentada	Alta
TP53 mutado	Inestabilidad genómica	Alto
PIK3CA + ERBB2	Sinergia proliferativa	Muy alto
PIK3CA + TP53	Proliferación + falta de control	Muy alto

Fuente: Autores (2026).

Estos hallazgos refuerzan la idea de que la progresión tumoral es el resultado de redes complejas. de interacción molecular, y no de cambios aislados. En este sentido, los modelos de aprendizaje automático Ofrecen una ventaja significativa al capturar estas interacciones de forma integrada, superando limitaciones de los modelos estadísticos tradicionales (LIBBRECHT; NOBLE, 2015).

Otro punto relevante se refiere al uso de métodos de interpretabilidad, como SHAP. lo que nos permitió comprender la contribución individual de las variables a las predicciones. Este enfoque representa un avance significativo, ya que reduce la opacidad de los modelos y promueve... su aplicabilidad clínica, especialmente en contextos que exigen transparencia en la toma de decisiones. decisión (LUNDBERG; LEE, 2017).

Además, los resultados obtenidos sugieren que la aplicación de la inteligencia artificial puede



Año VI, vol. 1 2026 | Envío: 15/03/2026 | Aceptado: 17/03/2026 | Publicación: 19/03/2026

para actuar como una herramienta complementaria a las estrategias tradicionales de estratificación pronóstica, contribuyendo a la terapia personalizada. Este enfoque puede ser particularmente útil en Escenarios con acceso limitado a paneles multigénicos comerciales, que amplían la equidad en la atención médica. Oncológico.

A pesar de los resultados prometedores, conviene tener en cuenta algunas limitaciones.

En primer lugar, los modelos de aprendizaje automático aplicados a datos genómicos están sujetos al riesgo de El sobreajuste, especialmente cuando el número de variables supera el número de observaciones. Aunque Se han empleado estrategias como la validación cruzada, así como la validación externa en cohortes. Las pruebas independientes son esenciales para confirmar la solidez de los resultados.

Otra limitación se relaciona con la falta de integración de datos multiómicos, como por ejemplo: transcriptómica, epigenómica y proteómica, que podrían ampliar la capacidad predictiva de modelos. La inclusión de estos datos puede capturar capas adicionales de regulación biológica relevante. para la progresión del tumor.

Además, los factores clínicos y ambientales también influyen en el comportamiento.

La integración de estos factores relacionados con el tumor no se exploró en profundidad en este marco analítico. Los elementos que se incluyan en los modelos futuros pueden contribuir a un enfoque más integral y aplicable.

Finalmente, aunque se utilizaron técnicas de interpretabilidad, la aplicación clínica

La eficacia de estos modelos aún depende de su validación prospectiva y de su incorporación a los procesos de toma de decisiones. Esto supone un reto actual en la implementación de la inteligencia artificial en la atención sanitaria.

De cara al futuro, los estudios futuros deberían centrarse en la validación externa de los modelos en diferentes poblaciones y la integración de datos multiómicos y clínicos, con el objetivo de aumentar la precisión. y generalización de predicciones. Además, el desarrollo de modelos híbridos, combinando

El aprendizaje automático y el conocimiento biológico previo pueden representar una estrategia prometedora para Para mejorar la interpretabilidad y la aplicabilidad clínica.

#### Consideraciones finales

Los resultados de este estudio demuestran que la aplicación de modelos de aprendizaje automático Los estudios basados en firmas genómicas muestran un alto potencial para predecir resultados clínicos. en el cáncer de mama subtipo luminal B. La capacidad de estos modelos para integrar múltiples variables. y capturar interacciones complejas entre vías moleculares permite un enfoque más refinado para heterogeneidad tumoral, superando las limitaciones de los métodos tradicionales basados en variables. aislado.

Identificación consistente de genes centrales para la carcinogénesis mamaria, como PIK3CA, TP53 y ERBB2 refuerzan la validez biológica del modelo propuesto y resaltan la relevancia de incorporar datos genómicos en la construcción de herramientas predictivas. Además, el uso



Año VI, vol. 1 2026 | Envío: 15/03/2026 | Aceptado: 17/03/2026 | Publicación: 19/03/2026

El uso de técnicas de interpretabilidad contribuye a la transparencia de los modelos, favoreciendo su

Potencial de aplicación en entornos clínicos.

Los resultados también ponen de relieve el papel emergente de la inteligencia artificial como herramienta.

Complementario en oncología, con potencial para contribuir a la estratificación y el apoyo al pronóstico.

para la toma de decisiones terapéuticas, especialmente en entornos con acceso limitado a pruebas moleculares.

avanzado.

Sin embargo, la consolidación de estos enfoques en la práctica clínica depende de la validación en

cohortes independientes, así como la integración con datos clínicos y multiómicos. De esta manera,

Los estudios futuros son esenciales para mejorar la solidez y la aplicabilidad de los modelos.

desarrollado.

En resumen, la integración entre la genómica tumoral y las técnicas de aprendizaje automático representa

una estrategia prometedora para avanzar en la medicina de precisión, con el potencial de tener un impacto

Tiene un impacto significativo en el manejo clínico del cáncer de mama.

#### Referencias

ANDRÉ, F. et al. Alpelisib para el cáncer de mama avanzado con mutación en PIK3CA y receptores hormonales positivos.

Revista de Medicina de Nueva Inglaterra, vol. 380, n.º 20, págs. 1929-1940, 2019.

BURSTEIN, HJ et al. Estimación de los beneficios de la terapia para el cáncer de mama en estadio temprano: las directrices del Consenso Internacional de St. Gallen. *Annals of Oncology*, vol. 25, n.º 10, págs. 1871-1888, 2014.

RED DEL ATLAS DEL GENOMA DEL CÁNCER. Retratos moleculares completos de tumores de mama humanos. *Nature*, vol. 490, n.º 7418, págs. 61-70, 2012.

COLLINS, GS et al. Informe transparente de un modelo de predicción multivariable para el pronóstico o diagnóstico individual (TRIPOD): la Declaración TRIPOD. *Annals of Internal Medicine*, vol. 162, n.º 1, págs. 55-63, 2015.

ESTEVA, A. et al. Una guía para el aprendizaje profundo en la atención médica. *Nature Medicine*, vol. 25, págs. 24–29, 2019.

KOURI, A. et al. Inteligencia artificial en oncología: aplicaciones actuales y direcciones futuras. *CA: A Cancer Journal for Clinicians*, vol. 70, n.º 4, págs. 268–287, 2020.

LIBBRECHT, MW; NOBLE, WS Aplicaciones del aprendizaje automático en genética y genómica.

*Nature Reviews Genetics*, vol. 16, n.º 6, págs. 321-332, 2015.

LUNDBERG, SM; LEE, S.-I. Un enfoque unificado para interpretar las predicciones de los modelos. En: *AVANCES EN SISTEMAS DE PROCESAMIENTO DE INFORMACIÓN NEURAL*. [S. l.]: NIPS, 2017. págs. 4765–4774.

PEROU, CM et al. Retratos moleculares de tumores de mama humanos. *Nature*, vol. 406, págs. 747–752, 2000.

PRAT, A. et al. Importancia pronóstica de Ki67 en el cáncer de mama. *Journal of Clinical Oncology*, vol. 33, n.º 36, págs. 4234–4242, 2015.



Año VI, vol. 1 2026 | Envío: 15/03/2026 | Aceptado: 17/03/2026 | Publicación: 19/03/2026

SILWAL-PANDIT, L. et al. Espectro de mutaciones del gen TP53 en el cáncer de mama. *Breast Cancer Research*, vol. 19, n.º 1, págs. 1-14, 2017.

SUNG, H. et al. Estadísticas mundiales de cáncer 2020: estimaciones de GLOBOCAN sobre incidencia y mortalidad en todo el mundo. *CA: A Cancer Journal for Clinicians*, vol. 71, n.º 3, págs. 209-249, 2021.

TOPOL, EJ Medicina de alto rendimiento: la convergencia de la inteligencia humana y artificial. *Nature Medicine*, vol. 25, págs. 44-56, 2019.