Clasificación de tumores de mama benignos y malignos a partir de características morfológicas mediante algoritmos de minería de datos

Martinez Briones Angel Jesus

***Resumen*—El cáncer de mama es el cáncer más frecuente entre mujeres a nivel global. La disparidad en la infraestructura sanitaria provoca que la tasa de mortalidad en países de renta baja supere el 50 %, frente a un 17 % en naciones de renta alta, debido principalmente a diagnósticos tardíos. El presente proyecto propone la aplicación y evaluación de dos modelos de clasificación basados en minería de datos como herramienta de diagnóstico para contextos de recursos limitados. Utilizando el conjunto de datos Wisconsin Breast Cancer Dataset, compuesto por 589 observaciones y 33 variables morfológicas obtenidas mediante aspirado de aguja fina (FNA), se realizó un preprocesamiento en lenguaje R, incluyendo el tratamiento de datos. Se implementó un modelo de regresión logística para la clasificación binaria de tumores (benignos y malignos) y un modelo SVM. El modelo ganador fue el SVM no obstante ambos modelos fueron satisfactorios, con este trabajo se busca demostrar la viabilidad de modelos computacionales como una alternativa accesible y precisa para la detección temprana de patologías mamarias.**

I. INTRODUCCIÓN

El cáncer de mama es una de las principales enfermedades oncológicas a nivel mundial. De acuerdo con la Organización Mundial de la Salud (OMS, 2024), el cáncer de mama es el cáncer más común entre las mujeres en 157 de los 187 países examinados en el 2022. Se estima que en ese mismo año fallecieron 670,000 personas por cáncer de mama en todo el mundo. Esta enfermedad está presente en todos los países del mundo y afecta predominantemente a mujeres, independientemente de su edad o estrato socioeconómico.

Las estrategias impulsadas por la OMS para enfrentar esta enfermedad se centran principalmente en la educación para la detección temprana y el acceso oportuno al tratamiento. Si bien no todos los tumores que se desarrollan en la mama son malignos, la identificación temprana de aquellos con potencial cancerígeno es fundamental, ya que se evita la propagación de estas células por el sistema linfático y aumenta significativamente la supervivencia de la persona afectada.

Bajo este contexto el uso de modelos de minería de datos se ha convertido en una herramienta prometedora que pueden ayudar a los profesionales de la salud a hacer un diagnóstico temprano, estas técnicas resultan relevantes por la desigualdad en la infraestructura de los sistemas de salud entre países. Mientras que, en los países de renta alta, el 83 % de las mujeres diagnosticadas con cáncer de mama sobreviven mientras que, en los países de renta baja, más del 50 % mueren debido a diagnósticos tardíos y acceso limitado a tratamientos adecuados. (Noticias ONU, 2025).

La pregunta de investigación que guía este estudio es: ¿qué tan efectivos son los modelos de regresión logística y máquinas de soporte

vectorial para clasificar tumores mamarios benignos y malignos a partir de características morfológicas?

II. MATERIALES

La base de datos utilizada cuenta con 589 observaciones y 33 variables, de las cuales 32 son numéricas y solo 1 es categórica, correspondiente al diagnóstico del tumor mamario. Esta variable categórica representa la clase objetivo, la cual indica si el tumor es benigno o maligno, y constituye la variable de respuesta empleada en los modelos de clasificación.

Las variables numéricas corresponden a características morfológicas, entre las que se incluyen medidas como radio, perímetro, área, textura, suavidad, simetría y número de puntos, entre otras. Estas características describen propiedades relevantes para la diferenciación entre tumores benignos y malignos.

Este conjunto de datos se basa en el Wisconsin Breast Cancer Dataset (Diagnostic), creado originalmente por el Dr. William H. Wolberg y colaboradores en la Universidad de Wisconsin-Madison. Las características se calcularon a partir de imágenes digitalizadas de aspirado de aguja fina (FNA) de masas mamarias, lo que permite una descripción cuantitativa del tejido analizado.

III. MÉTODO

El tratamiento de la base de datos se llevó a cabo en lenguaje de programación R. En primer lugar, el archivo datos.csv fue importado y almacenado en un data frame para su análisis. Posteriormente, se realizó una etapa de preprocesamiento de los datos, la cual incluyó la identificación y el tratamiento de valores faltantes y valores atípicos (outliers).

Los valores faltantes fueron abordados mediante técnicas de eliminación o imputación, según la naturaleza de la variable y la proporción de datos ausentes. La detección de valores atípicos se realizó utilizando el método del rango intercuartílico (IQR), con el objetivo de reducir la influencia de observaciones extremas que pudieran afectar el desempeño del modelo, se eliminaron 33 pacientes que presentaban más de 5 outliers en sus registros, Esta decisión se fundamentó en que dichas observaciones representaban casos clínicos de severidad extrema o anomalías de medición que podrían sesgar la capacidad de generalización de los algoritmos. Para los valores faltantes menores, se procedió con una imputación basada en la media, respetando los parámetros de del método IQR.

Dado que la variable de respuesta corresponde a una clasificación binaria (tumor benigno o maligno) y que las variables predictoras son de naturaleza numérica continua, se empleó un modelo de regresión logística y un modelo SVM como métodos. Para evaluar el desempeño del modelo, el conjunto de datos fue dividido en subconjuntos de entrenamiento 70 % y prueba 30 %, y se emplearon métodos de evaluación como la curva ROC, AUC y matriz de confusión.

IV. RESULTADOS

Ambos modelos arrojaron resultados positivos esta es la comparación de ambos modelos. El Cuadro 1 y 2 muestra la matriz de confusión de ambos modelos, en el cuadro 3 se comparan las métricas de desempeño, la figura 1 y 2 representa la curva ROC respectivamente

Cuadro I: Modelo SVM (Kernel Lineal)

Predicción / Referencia	B	M
Predicción (B)	99	5
Predicción (M)	0	57

Accuracy	96.8 %
AUC	0.992
Sensibilidad	91.9 %
Especificidad	100 %
Balanced Accuracy	95.97 %

0 falsos positivos = estrés, pruebas innecesarias
5 falsos negativos = riesgo clínico

Cuadro II: Modelo de Regresión Logística

Predicción / Referencia	B	M
Predicción (B)	91	6
Predicción (M)	9	56

Accuracy	91.3 %
AUC	0.956
Sensibilidad	90.3 %
Especificidad	91.9 %
Balanced Accuracy	91.1 %

6 falsos positivos = estrés, pruebas innecesarias
9 falsos negativos = riesgo clínico

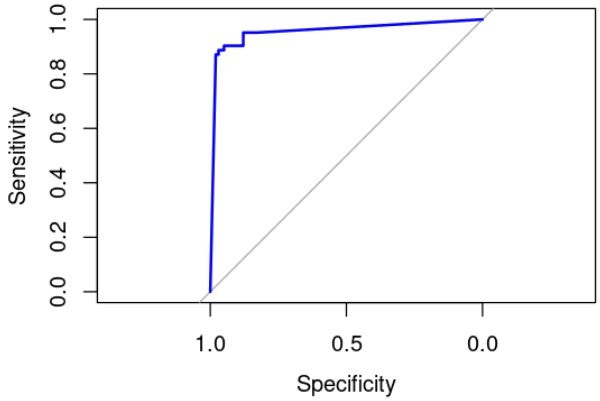

Figura 1: Curva ROC del Modelo de regresión logistica.

Cuadro III: Comparación de métricas

Métrica	Regresión Logística	SVM (Lineal)	Ganador
Accuracy	91.3 %	96.8 %	SVM
AUC	0.956	0.992	SVM
Sensibilidad (M)	90.3 %	91.9 %	SVM
Especificidad (B)	91.9 %	100 %	SVM
Balanced Accuracy	91.1 %	96.0 %	SVM

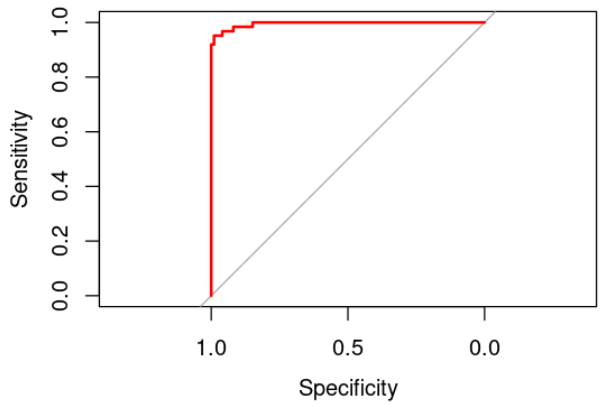

Figura 2: Curva ROC del Modelo SVM.

V. DISCUSIÓN

Los resultados obtenidos demuestran que ambos modelos poseen un desempeño sobresaliente para la clasificación diagnóstica. No obstante, el modelo SVM con kernel lineal se posiciona como el ganador con una Exactitud del 96.8 % y un AUC de 0.992. Estos hallazgos son congruentes con lo visto en clase, donde el SVM suele superar a la Regresión Logística en tareas de clasificación binaria gracias a su capacidad para maximizar la separación entre clases.

Una implicación de este estudio es la Especificidad del 100 % alcanzada por el SVM. En un contexto clínico esto significa que el modelo no generó falsos positivos garantizando que ningún sea clasificado erróneamente como maligno. Esto posiciona a la herramienta como un filtro inicial y como un apoyo para la toma de decisiones médicas basadas en la morfología de tumor.

No obstante, desde una perspectiva de riesgo clínico, el modelo SVM presentó 5 falsos negativos, lo que resulta en una sensibilidad del 91.9 %. Aunque es superior a los 9 falsos negativos de la Regresión Logística , estos casos representan pacientes con tumores malignos que el modelo clasificó erróneamente como benignos. En el contexto de países con infraestructura sanitaria limitada descrita por la OMS, un falso negativo es el escenario de mayor riesgo, ya que posterga un tratamiento. Por tanto, aunque el SVM es el modelo ganador, su implementación debe ser considerada como una herramienta de apoyo y no como un sustituto definitivo del juicio clínico.

Una limitación del presente estudio es el uso de un único conjunto de datos, lo cual puede afectar la generalización de los resultados. Asimismo, la eliminación de observaciones extremas podría excluir casos clínicos relevantes.

Como trabajo futuro, se propone la integración de bases de datos multihospitalarias y la exploración de más algoritmos o redes neuronales para intentar reducir falsos negativos y hacer más robusto el modelo.

VI. CONCLUSIONES

Se logró evaluar con éxito dos modelos de clasificación para la detección temprana del cáncer de mama, identificando al SVM (Lineal) como la herramienta con mayor capacidad predictiva, garantiza una discriminación casi perfecta entre tumores benignos y malignos.

La implementación de estos modelos responde a la necesidad planteada por la OMS sobre la detección temprana. Al alcanzar una especificidad del 100 % con el modelo SVM, se asegura que no existan falsos positivos, optimizando el uso de recursos médicos y reduciendo intervenciones innecesarias.

El uso de características morfológicas procesadas mediante minería de datos se valida como una alternativa de apoyo diagnóstico viable para reducir la brecha de mortalidad en países de renta baja. Esto permite un diagnóstico inicial preciso que no depende de infraestructura oncológica de alto costo.

VII. REFERENCIAS

1. World Health Organization. (2024, 13 de marzo). Cáncer de mama: Datos y cifras. Organización Mundial de la Salud. https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer

2. International Agency for Research on Cancer. (2025, 24 de febrero). Los casos de cáncer de mama aumentarán casi un 40 % de aquí a 2050. ONU Ginebra. https://news.un.org/es/story/2025/02/1536736

3. Neurocipher. (2020). Breast Cancer Dataset [Conjunto de datos]. Kaggle. https://www.kaggle.com/datasets/neurocipher/breast-cancer-dataset