



Universidad
Internacional
de Andalucía

TÍTULO

DESARROLLO DE UNA HERRAMIENTA DE INTELIGENCIA ARTIFICIAL PARA DETECTAR RECAÍDAS EN PACIENTES CON CÁNCER DE MAMA

AUTORA

Susana Martín Clavo

Tutor	Esta edición electrónica ha sido realizada en 2025
Institución	Dr. Tomás Urda Valcárcel
Curso	Universidad Internacional de Andalucía
©	<i>Diploma de Especialización en Digitalización e Innovación en Salud</i>
©	<i>(2022-23)</i>
Fecha documento	Susana Martín Clavo
	De esta edición: Universidad Internacional de Andalucía
	2023



Universidad
Internacional
de Andalucía



**Atribución-NoComercial-SinDerivadas
4.0 Internacional (CC BY-NC-ND 4.0)**

Para más información:

<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.es>

<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>

DESARROLLO DE UNA HERRAMIENTA DE INTELIGENCIA ARTIFICIAL PARA DETECTAR RECAÍDAS EN PACIENTES CON CÁNCER DE MAMA

Susana Martín Clavo

Curso Académico: 2022-2023

INDICE

RESUMEN.....	5
ABSTRACT.....	7
INTRODUCCIÓN.....	9
OBJETIVOS.....	19
MATERIAL Y MÉTODOS.....	20
RESULTADOS.....	25
CONCLUSIONES.....	28
BIBLIOGRAFÍA.....	30

INDICE DE TABLAS

1.- Características histológicas de los subtipos de cáncer de mama.....	10
2.- Variables demográficas.....	22
3.- Variables clínicas, tumorales y moleculares.....	23

INDICE DE FIGURAS

Figura 1.- Anticipando los datos en la era de la Medicina Personalizada de Precisión.....	13
Figura 2.- Incertidumbres en la Práctica Clínica.....	14
Figura 3.- Metodología de trabajo.....	21
Figura 4.- Ejemplo de Redes Neuronales.....	26

RESUMEN

INTRODUCCIÓN

La predicción de recidiva en pacientes que han sido operados de cáncer de mama juega un papel muy importante en tareas médicas como el diagnóstico y la planificación del tratamiento que hay que realizarle al mismo. Las redes neuronales artificiales se muestran como una herramienta potente para el análisis de conjuntos de datos donde hay relaciones no lineales entre los datos a estudio y la información a ser pronosticada.

OBJETIVO

OBJETIVO GENERAL

Analizar e identificar variables clínicas sobre pacientes con cáncer de mama con el fin de obtener patrones estadísticos y desarrollar un modelo predictivo basado en inteligencia artificial que detecte posibles recaídas.

OBJETIVOS ESPECÍFICOS

Definir la técnica o algoritmo requerido que automatice esa generación de modelos para el análisis de la información con respecto al seguimiento, control y posible recaída del cáncer de mama.

MATERIAL Y MÉTODOS

Estudio descriptivo con el objeto de identificar y caracterizar diferentes variables (clínicas, fisiopatológicas, tumorales, moleculares) de pacientes que han sufrido cáncer de mama para finalmente obtener una herramienta de inteligencia Artificial (IA) o algoritmo que pueda identificar factores de riesgo de recaída con una precisión óptima, para brindar un tratamiento adecuado de forma precoz acorde a las nuevas evidencias.

RESULTADOS

Para nuestra investigación, el modelo teórico final propuesto sería el de **Redes Neuronales Artificiales** (RNA). La RNA es un método de aprendizaje supervisado empleado para dar solución a problemas de clasificación y regresión en aprendizaje automático. Un aspecto crucial para poder realizar aprendizaje y diagnóstico en la red neuronal es seleccionar dos conjuntos independientes de datos procedentes de la base de datos de los pacientes, que serán usados respectivamente para el entrenamiento de la red y para validar la eficacia de la predicción.

CONCLUSIONES

Gracias a la digitalización de grandes cantidades de datos, el desarrollo de algoritmos de Aprendizaje Automático y la mejora de la capacidad de cálculo de los ordenadores de las últimas décadas, la IA puede abrir excelentes oportunidades para la automatización de tareas, la aplicación de la medicina de precisión o el avance en la investigación a través de la detección de patrones complejos en bases de datos médicas.

Palabras clave: Cáncer de mama, recaída, Algoritmo, Red Neuronal Artificial

ABSTRAC

INTRODUCTION

The prediction of recurrence in patients who have undergone surgery for breast cancer plays a very important role in medical tasks such as diagnosis and treatment planning that must be carried out. Artificial neural networks are shown to be a powerful tool for the analysis of data sets where there are non-linear relationships between the data under study and the information to be forecast.

AIM

GENERAL OBJECTIVE

Analyze and identify clinical variables on patients with breast cancer in order to obtain statistical patterns and develop a predictive model based on artificial intelligence that detects possible relapses.

SPECIFIC OBJECTIVES

Define the technique or algorithm required to automate the generation of models for the analysis of information regarding the follow-up, control, and possible relapse of breast cancer.

MATERIALS AND METHODS

Descriptive study in order to identify and characterize different variables (clinical, pathophysiological, tumor, molecular) of patients who have suffered from breast cancer to finally obtain an Artificial Intelligence (AI) tool or algorithm that can identify risk factors for relapse with optimal precision, to provide adequate treatment early according to new evidence.

RESULTS

For our research, the final theoretical model proposed would be that of Artificial Neural Networks (RNA). ANN is a supervised learning method used to solve classification and regression problems in machine learning. A crucial aspect to be able to perform learning and diagnostics in the neural network is to select two independent data sets from the patient database, which will be respectively used for training the neural network and to validate the efficacy of the neural network prediction.

CONCLUSIONS

Thanks to the digitization of large amounts of data, the development of Machine Learning algorithms and the improvement in the computing power of computers in recent decades, AI can open up excellent opportunities for task automation, the application of medicine of precision or advancement in research through the detection of complex patterns in medical databases.

Keywords: Breast cancer, relapse, Algorithm, Artificial Neural Network

1. INTRODUCCIÓN

El cáncer de mama es el cáncer más frecuente en la mujer española. Se calcula que 1 de cada 8 mujeres españolas tendrá un cáncer de mama en algún momento de su vida. Aunque la mortalidad por cáncer de mama ha descendido en los últimos años gracias a los programas de cribado y a la mejora de los tratamientos el cáncer de mama sigue siendo la primera causa de muerte por cáncer en España en las mujeres. El estadio en el que se ha diagnosticado el cáncer influye en la supervivencia. La supervivencia en el estadio I es de más del 98% y en cambio en los estadios IV la supervivencia desciende al 24%.

Se estima que el 5-10% de los cánceres de mama son hereditarios. Las 20-25% de las mutaciones responsables de los cánceres de mama hereditarios ocurren en los genes BRCA 1 y 2. La probabilidad de que una mujer portadora de mutación en los genes BRCA 1 o 2 presente cáncer de mama se sitúa en torno al 57% a los 70 años(1). Existen otros genes como PALB2, p53, CDH1, ATM, CHEK2 o BRIP1 que también aumentan la probabilidad de padecer un cáncer de mama a lo largo de la vida.

A pesar de que el pronóstico a largo plazo es favorable, cada año sufrirán una recaída aproximadamente el 10 % de las pacientes con cáncer de mama; la identificación de las personas con riesgo de recaída elevado puede mejorar el tratamiento(2).

El riesgo de recurrencia del cáncer de mama en etapa temprana a los 10 años o más de haber recibido el diagnóstico está vinculado al tamaño del cáncer, el número de ganglios linfáticos con cáncer, y si el cáncer es positivo para receptores de estrógeno, según un estudio danés(3) . En general, según el estudio, hubo recurrencia hasta 32 años después del diagnóstico. Esta es la primera vez que se muestra en una investigación que el riesgo de recurrencia del cáncer de mama dura más de 30 años.

La clasificación de estos tumores se realiza en función de sus características inmunohistoquímicas en cuatro subtipos moleculares, estando los mismos relacionados directamente con el pronóstico de la enfermedad(4) Tabla 1.

Tabla 1. Características histológicas de los subtipos de cáncer de mama.

SUBTIPO	RECEPTORES HORMONALES	HER 2 NEU	KI67 O ÍNDICE DE PROLIFERACIÓN	PRONÓSTICO
LUMINAL A	+	-	-	BUENO
LUMINAL B	+	-	+	INTERMEDIO
HER2 POSITIVO	-	+	+	MALO
BASAL LIKE O TRIPLE NEGATIVO	-	-	+	MUY MALO

FUENTE: Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J et al. Breast cancer intrinsic subtype classification, clinical use and future trends. American journal of cancer research. 2015;5(10):2929–2943.

El diagnóstico definitivo de cáncer de mama lo establece el especialista en anatomía patológica al observar células malignas bajo el microscopio a partir de la muestra de la biopsia. El examen patológico clasificará el cáncer como invasivo o no invasivo. Los cánceres no invasivos también se denominan in situ y son aquellos que no se han extendido fuera del conducto (ductales in situ) o del lobulillo (lobulillares in situ). El patólogo analiza varios aspectos de la pieza que recibe y así define el tumor y establece algunos factores importantes para determinar el pronóstico y la respuesta a determinados tratamientos(5):

- **Tamaño tumoral:** cuanto mayor sea el tumor más riesgo tiene de recidiva.
- **Tipo histológico:** depende de las células de las que derive el tumor. El carcinoma ductal es el tipo más frecuente (80%) y deriva de los conductos por donde fluye la leche, el siguiente en frecuencia es el carcinoma lobulillar que deriva de las células del lóbulo (donde se produce la leche).
- **Grado histológico:** describe el grado de diferenciación (maduración) de las células del tumor. Las que son más diferenciadas (grado I), son más maduras y menos agresivas en oposición a las menos diferenciadas (grado III).
- **Afectación ganglionar:** el número de ganglios afectados es el factor pronóstico más importante de forma que cuanto mayor es el número de ganglios afectos, mayor es el riesgo de recaída. Por ello, cuando se opera un cáncer de mama es fundamental el análisis

de los ganglios linfáticos de la axila puesto que se considera el primer sitio de extensión del tumor.

Una opción para evaluar los ganglios es la técnica del ganglio centinela que permite conservar la mayoría de los ganglios axilares para evitar complicaciones secundarias debidas a la extirpación de toda la cadena ganglionar.

- **Receptores hormonales:** el patólogo analiza si las células del tumor presentan receptores para dos tipos de hormonas, los estrógenos y la progesterona. Esto tiene su interés por que su expresión confiere mejor pronóstico y porque indican que las pacientes que los tienen van a responder a la terapia hormonal.
- **HER-2 (receptor 2 del factor de crecimiento epidérmico humano):** es una proteína que participa en el crecimiento de las células. Está presente en células normales y en la mayoría de los tumores, pero en un 15-20% de los tumores de mama se encuentra en concentraciones elevadas y esto confiere al tumor mayor agresividad. Estos tumores con sobreexpresión de HER-2 son con mucha frecuencia sensibles al tratamiento con tratamientos anti-HER 2.
- **Ki 67:** Ki 67 es una proteína que se halla en el núcleo de las células cuando se dividen. Ki 67 determina el índice de proliferación. Los tumores con índices de proliferación altos (>14%) tiene peor pronóstico(6).
- **Plataformas moleculares o perfiles de expresión génica:** En la biopsia puede cuantificarse la expresión de diversos genes expresados por el tumor. Estas plataformas pueden ayudar a predecir el riesgo de recaída y el beneficio de la quimioterapia adyuvante en los tumores luminales. En la actualidad existen varias plataformas comercializadas.

El patólogo también informara del estado de los márgenes tras la cirugía para saber si el tumor se ha extirpado por completo. Si los márgenes son positivos quiere decir que no se ha extirpado todo el tumor en su totalidad y es posible que tenga que realizarse una nueva cirugía.

Por otro lado, **el campo de la oncología** hoy en día ha experimentado dos de las mayores evoluciones tecnológicas: la “ómica” molecular (genómica, proteómica, epigenómica) y el “**big data**”, que analiza grandes bases de datos combinando biomarcadores y tendencias poblacionales para generar conocimiento en tiempo real. Hace un par de décadas el cáncer se diagnosticaba mediante una combinación de imágenes de rayos X y pruebas de histopatología. En contraste, las pruebas moleculares ahora pueden informar sobre los cambios en cientos de genes y proteínas para diagnosticar y determinar el pronóstico y tratamiento del cáncer en un individuo. El tratamiento del cáncer ha evolucionado de manera similar, donde agentes citotóxicos se han

reemplazado o modificado para dar lugar a planes de tratamiento y terapias agnósticas de histología o tumor de origen y, a menudo, dirigidos a biomarcadores presentes en el tumor, las vías de señalización involucradas en el crecimiento del cáncer, o modulando el sistema inmunológico del paciente para atacar la enfermedad (por ejemplo, los inhibidores ‘checkpoint’).

En este sentido, los **datos** se han convertido en un torrente que fluye en todos los ámbitos de la sociedad. Posiblemente las aplicaciones en salud sean unas de las más importantes que actualmente tiene el Big Data y sea aún un campo por explorar. La aplicación de la Big Data a la Salud obtiene un beneficio intangible que es la mejora de la salud de los ciudadanos que no se obtiene con la aplicación de estos procesos a otros campos del conocimiento(7). La aplicación de la Inteligencia Artificial (IA) al ámbito de la salud es posiblemente el de mayor impacto social, por las implicaciones que tiene para el bienestar de las personas. Muchos han depositado las esperanzas de la Medicina de Precisión en los Sistemas de Ayuda a la Decisión Médica.

Campos como el Machine Learning (ML) y el Procesamiento del Lenguaje Natural convierten los datos en la información que necesitan para conseguir esa IA que extraiga conclusiones, haga predicciones o se comunique con nosotros(8).

Cuanto más datos tengamos, más fiable y representativa será la muestra del proceso que queremos automatizar, y mejor va a funcionar el software que se genere.

El análisis de los datos en salud, desde historias clínicas a datos ómicos, con herramientas de IA y el conocimiento derivado de los mismos previsiblemente jugarán un papel fundamental en la configuración de la medicina del futuro, ya que dan lugar a nuevas oportunidades y aplicaciones en todas las áreas, desde la investigación biomédica y traslacional hasta la práctica clínica y salud pública. Figura 1

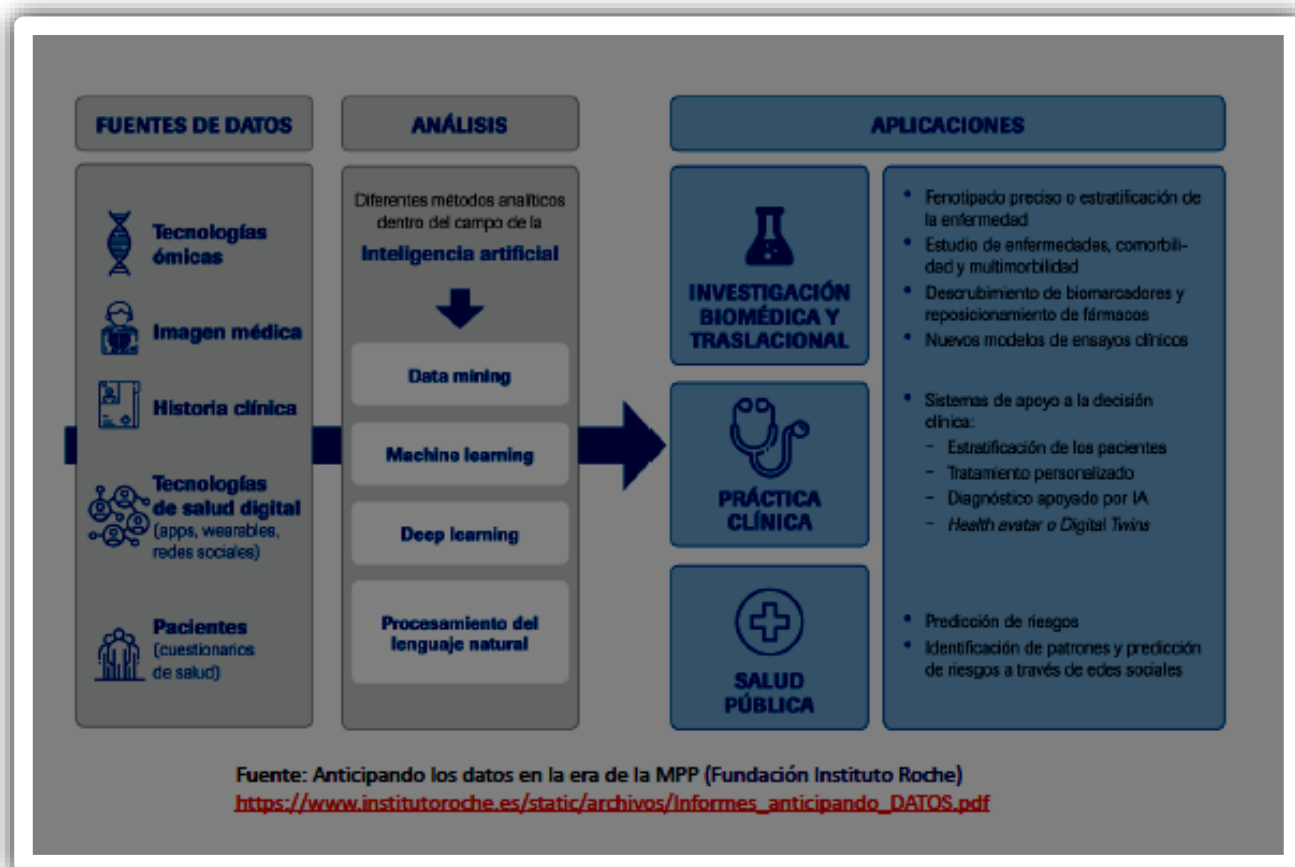


Figura 1.- Anticipando los datos en la era de la Medicina Personalizada de Precisión.

Como sabemos, la prevención y el tratamiento eficaz son las mejores herramientas para ejercer un correcto control del cáncer y gracias a tecnologías como el Big Data, podemos conocer cómo afecta la enfermedad a cada persona y prevenir un mayor desarrollo de ésta. Mediante la utilización del Big Data, se pueden recoger los datos, analizarlos y obtener una visión de cómo afecta la enfermedad a cada persona. Con todo ello, se puede entender mejor cómo puede progresar la enfermedad, los detonantes y personalizar los tratamientos para que sean eficaces y mejoran la calidad de vida de los pacientes(9).

Es importante recordar que el Big Data sin el análisis de los expertos, no es eficaz. Se necesita, por un lado, la mejor tecnología y datos de calidad y por el otro, un equipo multidisciplinar de expertos que sepan utilizar estas herramientas y entiendan el funcionamiento de los procesos más complejos, que les permitan analizar y extraer conclusiones de calidad.

En este sentido, la tecnología, y en particular la IA, se ha propuesto a menudo como una **solución a la incertidumbre que rodea muchos aspectos de la práctica sanitaria (Figura 2):**

- Cómo clasificar las condiciones de los pacientes (incertidumbre diagnóstica)
- Por qué y cómo los pacientes desarrollan enfermedades (incertidumbre pato-fisiológica)
- Qué tratamiento será más apropiado para ellos (incertidumbre terapéutica)
- Si se recuperarán con o sin un tratamiento específico (incertidumbre pronóstica).



Figura 2.- Incertidumbres en la Práctica Clínica

A día de hoy es difícil encontrar una definición universal de lo que se conoce como IA. A menudo se refiere al campo de las ciencias de la computación que trata de imitar los procesos cognitivos humanos, la capacidad de aprendizaje y el almacenamiento de conocimiento. Otras definiciones lo estudian en un sentido más amplio, a través de la comprensión y construcción de entidades inteligentes, generalmente entendidas como software informático. Centrándonos en el tipo de tareas, se podría considerar la IA como sistemas automáticos capaces de, por ejemplo, traducir un documento, reconocer a una persona por sus rasgos faciales o conducir un automóvil. La IA no se limita a imitar tareas humanas, sino que en algunos casos es capaz de superar al mejor experto en la materia tomando decisiones con menores tasas de error que una persona o identificando patrones imperceptibles para el ojo humano. Por lo tanto, la IA permite analizar información con una aproximación diferente de la tradicional. Ya no nos limitamos a describir la información de la que disponemos para dar respuesta a lo sucedido. La IA permite responder otras preguntas: ¿qué sucedió? (diagnosticar), ¿qué pasará? (predecir) y ¿qué debería hacer? (prescribir).

Por lo tanto, la aplicación de la IA y el aprendizaje automatizado como tecnologías fundamentales y habilitadoras para ayudar al oncólogo tratante, son prometedoras para mejorar la atención del

cáncer. A través del desarrollo continuo de sistemas escalables de aprendizaje profundo, los médicos pueden implementar de manera más libre y eficiente la medicina basada en resultados.

Dentro del campo de la IA, los algoritmos pueden analizar grandes cantidades de datos médicos, incluidos los registros de pacientes, diagnósticos y resultados de pruebas de laboratorio, para detectar patrones y hacer predicciones sobre la probabilidad de recaídas. Estos algoritmos también pueden identificar paciente con un mayor riesgo de sufrir recaídas y brindar recomendaciones de tratamiento personalizadas(10).

Un significativo número de trabajos de investigación(11) están demostrando un impresionante rendimiento dentro de estas áreas por parte de sistemas construidos mediante el aprendizaje automático (Machine Learning). Son tecnologías de enorme potencialidad y frecuente uso en la investigación biomédica básica, en la investigación traslacional y en la práctica clínica y asistencial.

Por ejemplo, un equipo de investigadores de la Universidad de California en Los Ángeles desarrolló un algoritmo de aprendizaje profundo que puede predecir la recaída en pacientes con cáncer de mama. El algoritmo analizó imágenes de Tomografía por emisión de positrones (PET) y Resonancia Magnética de más de 400 mujeres con cáncer de mama. Los resultados mostraron que el algoritmo podía predecir con precisión qué pacientes estaban en mayor riesgo de recaída(12).

Otro equipo de investigadores de la Universidad de Stanford desarrolló una herramienta de IA llamada MammoScreen® que puede detectar tempranamente el cáncer de mama en mamografías(13). La herramienta utiliza un algoritmo de aprendizaje profundo para analizar mamografías y detectar anomalías que podrían indicar la presencia de cáncer de mama. Los resultados del estudio indicaron que la herramienta superó a los radiólogos humanos en la detección de cáncer de mama en las mamografías.

Asimismo, para crear IA hay dos fundamentos básicos de los que se valen: los **algoritmos** y los **datos para configurarlos**. Sintetizando se puede decir que los algoritmos son instrucciones paso a paso que ayudan a un ordenador a completar un cálculo. De esta manera, la IA utiliza algoritmos para crear máquinas que aprendan de su propia experiencia, se reconfiguran ante nuevos escenarios y

desarrollan las tareas de manera similar a cómo lo haríamos nosotros. Para poder comprender el funcionamiento de ML es necesario conocer las siguientes variantes que lo componen:

Los principales **tipos de algoritmos** de este campo son:

- **Aprendizaje por refuerzo:**

Este tipo de aprendizaje consiste en la interacción constante basada en la “prueba y error”, que una máquina puede realizar en tiempo récord en determinadas condiciones o en un entorno dado (en un juego, por ejemplo) y con un objetivo específico que se denomina “recompensa”. De esta manera, se pueden obtener resultados, patrones, correlaciones, caminos y conclusiones basadas en experiencias previas generadas por la propia máquina.

Un ejemplo de este modelo de aprendizaje es la IA ajedrecista AlphaZero de DeepMind. Donde enseñas a la máquina a jugar una partida introduciendo las reglas en su sistema, luego le das una recompensa cada vez que consigue su objetivo (por ejemplo, ganar una partida).

Los principales algoritmos utilizados en el aprendizaje por refuerzo son la programación dinámica (*dynamic programming*), Q-Learning y SARSA (*State-action-reward-state-action*).

- **Aprendizaje supervisado:**

El aprendizaje supervisado emplea modelos predictivos que utilizan datos de entrenamiento. Dado un conjunto de datos, se pretende que el sistema sea capaz de lograr una salida. Con este sistema, el modelo es ajustado (entrenado) hasta conseguir los resultados deseados(14). Un ejemplo de este aprendizaje es el de los coches autónomos. Los principales algoritmos en el Supervised machine learning son los árboles de decisión, clasificaciones Naive Bayes, la regresión ordinaria por mínimos de cuadrados, la regresión logística y el Support Vector Machines (SVM).

- **Aprendizaje no supervisado**

Los algoritmos de este aprendizaje son similares al del aprendizaje supervisado, sin embargo, estos se ajustan solo en función de los datos de entrada. Es decir, el algoritmo realiza un auto entrenamiento sin indicaciones externas. Estos algoritmos están pensados para realizar un procesamiento más complejo.

Los algoritmos más utilizados en el aprendizaje no supervisado son los algoritmos de agrupamiento o *clustering*, el análisis de componentes principales (PCA), el *Singular Value Decomposition* (SVD) y el análisis de componentes independientes.

- **Regresión Lineal**

Se utiliza para estimar los valores reales (costo de las viviendas, el número de llamadas, ventas totales, etc.) basados en variables continuas. La idea es tratar de establecer la relación entre las variables independientes y dependientes por medio de ajustar una mejor línea recta con respecto a los puntos.

- **Regresión logística**

Los modelos lineales, también pueden ser utilizados para clasificaciones; es decir, que primero ajustamos el modelo lineal a la probabilidad de que una cierta clase o categoría ocurra y, a luego, utilizamos una función para crear un umbral en el cual especificamos el resultado de una de estas clases o categorías. La función que utiliza este modelo es denominada regresión logística.

- **Árboles de decisión**

Los Árboles de Decisión son diagramas con construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema. Los Árboles de Decisión están compuestos por nodos interiores, nodos terminales y ramas que emanan de los nodos interiores.

- **Algoritmo de Random**

La idea central detrás del algoritmo de Random Forest es construir una gran cantidad de árboles muy poco profundos, y luego se toma la clase que cada árbol eligió.

- **Vectores de soporte (Support Vector Machines (SVM))**

La idea detrás de SVM es encontrar un plano que separe los grupos dentro de los datos de la mejor forma posible. Aquí, la separación significa que la elección del plano maximiza el margen entre los puntos más cercanos en el plano; estos puntos se denominan vectores de soporte.

Teniendo en cuenta todo lo anteriormente expuesto, la problemática de la recaída en el cáncer de mama y que conocemos algo más sobre el aprendizaje automático, sería interesante desarrollar un algoritmo intuitivo para identificar a las pacientes con potencial de recaída en el momento del diagnóstico en el cáncer de mama, que puede influir en su tratamiento posterior, lo cual influirá positivamente en la paciente y en comparación con aquellas sin riesgo de recaída.

No obstante, existen cuestiones a la hora de aplicar la IA en el mundo de la medicina no estrictamente relacionada con la calidad de los algoritmos. Estas tienen que ver con los problemas legales del tratamiento automático de datos personales o cómo aplicar el Aprendizaje Automático en el día a día de la práctica clínica. La protección de los datos de carácter personal es un derecho fundamental recogido en la Constitución Española; sin embargo, se reconoce la posibilidad de utilización de dichos datos bajo ciertas condiciones si con ello se da un servicio a la sociedad. Si tanto la organización que recoge los datos como la que los trata tienen residencia en España, desde 2018 deben ajustarse a la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales(15). Según esta, se permite el uso de datos personales con fines estadísticos, siempre y cuando los resultados sean de carácter agregado. Además, también es necesario adoptar medidas para anonimizar y proteger dichos datos y se reconocen una serie de derechos de los interesados como el de ser informado, el de acceso o el de rectificación. Por todo ello, dado que la utilización de servicios de computación en la nube es muy habitual en IA por su flexibilidad y mayor potencia computacional, los proveedores más importantes han adaptado sus políticas de procesamiento de datos para suministrar un servicio completamente seguro. Además de suscribirse al Código de Conducta CISPE (Cloud Infrastructure Services Providers in Europe) para garantizar los estándares de protección de datos apropiados asociados con la legislación. Asimismo, en situaciones en las que se emplee un modelo de IA como herramienta diagnóstica o de toma de decisiones, es obligado informar al paciente y obtener su consentimiento

2.- OBJETIVOS

OBJETIVO GENERAL

- **Analizar e identificar variables clínicas** sobre pacientes con cáncer de mama con el fin de obtener patrones estadísticos y desarrollar un modelo predictivo basado en inteligencia artificial (IA) que detecte posibles recaídas.

OBJETIVOS ESPECÍFICOS

- **Definir la técnica o algoritmo** requerido que automatice esa generación de modelos para el análisis de la información con respecto al seguimiento, control y posible recaída del cáncer de mama.

3.- MATERIAL Y MÉTODOS

Estudio descriptivo con el objeto de identificar y caracterizar diferentes variables (clínicas, fisiopatológicas, tumorales, moleculares) de pacientes que han sufrido cáncer de mama para finalmente obtener una herramienta de inteligencia Artificial (IA) o algoritmo que pueda identificar factores de riesgo de recaída con una precisión óptima, para brindar un tratamiento adecuado de forma precoz acorde a las nuevas evidencias.

Para llevar a cabo los objetivos marcados en el epígrafe anterior se procedería de la siguiente manera (Figura 3):

1.- Recolección de datos.

2.- Verificación de datos.

3.- Clasificación de datos.

4.- Limpieza de datos:

4.1.- Para el correcto entrenamiento de los algoritmos y obtener un resultado acertado respecto a la previsión de recaída del cáncer de mama es necesario tener alta calidad de los datos, para ello es necesario la corrección de los errores de los mismos para posterior procesamiento con los algoritmos. Para la correcta limpieza de los datos se realizarán los siguientes pasos:

4.1.1.- Importación de datos.

4.1.2.-Formateo de datos.

4.1.3.- Comprobación de datos nulos.

4.1.4.- Eliminación de datos innecesario.

5.- Procesamiento de los datos (Minería de datos).

6.- Finalmente evaluación y validación del modelo.

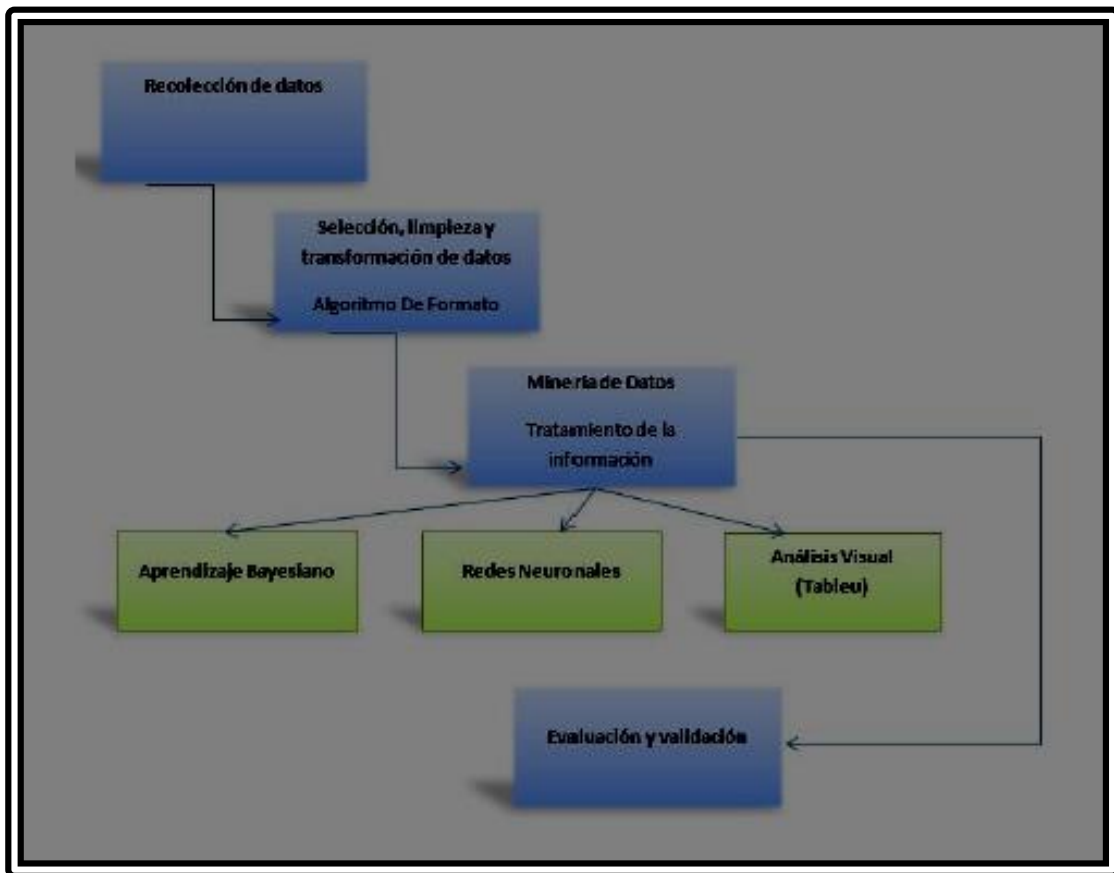


Figura 3.- Metodología de trabajo

El diseño de esta metodología partiría de la creación y definición de la propia base de datos para el desarrollo de la investigación.

Se utilizaría el proceso general del KDD (Knowledge Discovery in Databases)(16) para los datos recabados donde se contemplaría: entendimiento del problema, selección de datos y preprocesamiento y limpieza y transformación de datos. Bajo la plataforma WEKA(17), desarrollada con trabajos previos, se realizaría el análisis de datos de las variables evaluadas.

La selección y preprocesamiento de datos se obtendría mediante la identificación de variables y la limpieza de datos consistiría en seleccionar de manera coordinada los datos más relevantes del proceso y en el posterior análisis de los expedientes clínicos.

Lo primero que se necesitaría, sería la base de datos con la que se pretende trabajar, que en nuestro caso se emplearía una base de datos del Servicio Médico de Oncología del Hospital Materno-Infantil de Málaga. Una vez descargados y descomprimidos los datos comprobaríamos la información que contienen.

En la tabla 2 y 3 se muestra la descripción de cada uno de los atributos seleccionados (variables demográficas, clínicas, tumorales y moleculares) para el análisis y modelo de minería de datos(18). En la construcción y elaboración de la base de datos se tendrían en cuenta 47 variables independientes y una variable dependiente denominada recaída cáncer de mama.

Tabla 2.- Variables demográficas

Variables
Datos demográficos del paciente
Estado Civil
Información Demográfica
Raza/Etnia
Historia Médico Personal
Edad en el momento del diagnóstico
Estado Menopáusico
Edad en el momento de la menarquia
Estado de tabaquismo
Historial de Infertilidad
Consumo de alcohol
Muerte (relacionado con el cáncer de mama o no)
Historia de otro tipo de cáncer
Historia de otras enfermedades crónicas
Lactancia
Índice de Masa Corporal
Índice de comorbilidad de Charlson
Antecedentes familiares
Cáncer de mama
Otros tipos de cáncer

Tabla 3.- Variables clínicas, tumorales y moleculares.

/Estadio Anatómico	Variables
	1. Estadificación Clínica
	Imagen de diagnóstico por Resonancia Magnética Biopsia
	2. Estadificación Patológica TNM
	Tumor primario Ganglio linfático regional Metástasis
	3. Estadificación posterior a la Terapia
	3.1 Información clínica Radiación Hormonoterapia Quimioterapia Tipo de Cirugía Terapia NAC Terapia anti-HER2
	3.2 Información Patológica Respuesta a la terapia Neoadyuvante Respuesta completa patológica
	4. Re estadificación en caso de recidiva del tumor
	Resultado (recurrencia o no) Tiempo de recurrencia
Estadio pronóstico	Variables
	Grado del tumor Receptor Hormonal Invasión tumoral HER2 Tipo de Tumor Ki67 Expresión oncogénica Prueba en paneles multigénicos
	Otros marcadores moleculares
	CD44 ABCC4 ABCC11 n- Cadherina Receptor del Factor de Crecimiento Epidérmico (EGFR) Citokeratina 5/6 (CK 5/6)

Definida la base de datos anterior, y seleccionadas las variables clínicas que se han considerado más oportunas para preparar el modelo predictivo, el problema que se nos presenta es el siguiente: dado un paciente determinado que presenta unos factores de pronóstico concretos ¿sufrirá recaída de la enfermedad durante el seguimiento? Para responder a esta pregunta y predecir la recurrencia

del cáncer necesitamos modelos concretos de Machine Learning (ML) y clasificar los datos de los pacientes en aquellos con o sin recaída de cáncer.

Partiendo de los datos brutos (raw data) se pasaría a una información con datos estructurados (preprocesado) y se construiría una base de datos inicial. A partir de esta base de datos realizaríamos un análisis descriptivo y exploratorio, con el objetivo de identificar y seleccionar las variables más significativas.

Tomando como referencia la base de datos elaborada en el paso anterior, se estructuraría el encabezado y detalle del archivo ARFF. Este archivo sería corrido desde la plataforma de aprendizaje automático y minería de datos WEKA, con el fin de predecir el comportamiento de la variable dependiente, denominada recaída cáncer de mama para posterior predicción realizada por medio del algoritmo.

Para la poder llevar a cabo nuestro proyecto, debe dirigirnos a todos aquellos/as profesionales, que consideramos necesario para el proyecto. A través del área de Salud correspondiente, esta iniciativa se llevaría a cabo en el marco de colaboración de la Subdirección de Sistemas de Información, Servicio de Informática, Control de Gestión, Servicio de Farmacia Hospitalaria, Unidad de Bioestadística, Servicio de Biblioteca y Dirección Médica con el servicio clínico de Oncología Médica, en el desarrollo de un concepto nuevo de abordaje de la información.

Nuestro ámbito de estudio será el área de Atención especializada en la Unidad de mama del Servicio de Oncología del Hospital Materno Infantil de Málaga.

4.- RESULTADOS

Como consecuencia del desarrollo de cada uno de los pasos de la metodología, el proyecto en su formulación y desarrollo nos genera problemas de carácter informático que necesitarían de un enfoque experto.

Para nuestra investigación, el modelo teórico final propuesto sería el de **Redes Neuronales Artificiales** (RNA). A continuación se describe de manera general la solución propuesta.

La RNA es un método de aprendizaje supervisado empleado para dar solución a problemas de clasificación y regresión en aprendizaje automático(19). Un aspecto crucial para poder realizar aprendizaje y diagnóstico en la red neuronal es seleccionar dos conjuntos independientes de datos procedentes de la base de datos de los pacientes, que serán usados respectivamente para el entrenamiento de la red y para validar la eficacia de la predicción(20).

La RNA aprende de los datos que se proporcionan y se inspira en nuestros propios sistemas neuronales, aunque no funcionan de la misma manera. Estos modelos reciben una gran cantidad de datos en una capa que llamamos capa de entrada. A partir de estos datos, se hacen comparaciones y el modelo identificará automáticamente las características de los datos y los etiquetará.

Se trata de una interconexión de unidades de procesamiento denominadas neuronas, modeladas a partir de la relación:

$$y = f(\mathbf{w}^T \mathbf{x})$$

Donde y corresponde la salida de la neurona, $\mathbf{x}$ a las entradas, $\mathbf{w}^T$ a los pesos de las entradas, y f a la función de activación de la neurona. A partir de estas unidades, es posible construir redes complejas (como la mostrada en la Figura 4), denominadas también perceptrones multicapa, que pueden implementar o modelar funciones complejas con capacidad de construir diferentes representaciones de los datos de entrada y salida.

Hay tres tipos de capas: Entrada capa, capas ocultas y capa de salida.

Así es como funciona una RNA: primero a cada neurona en la capa de entrada se le asigna un valor, llamado función de activación. Luego se le asignan un peso aleatorio, mientras que a las neuronas de la capa oculta se les asigna un valor de sesgo aleatorio. En la capa oculta, un algoritmo llamado función de activación asigna un nuevo peso a la neurona de la capa oculta, que se multiplica por un

valor de sesgo aleatorio en la capa de salida. Esta función de activación se multiplica por un peso aleatorio, que mejora con más interacciones a través de un proceso llamado backpropagation. A través de esto, el modelo desarrolla una predicción aleatoria sobre su salida en la instancia dada. Usando la propagación hacia atrás, este modelo ajusta sus parámetros para que la respuesta sea más precisa.

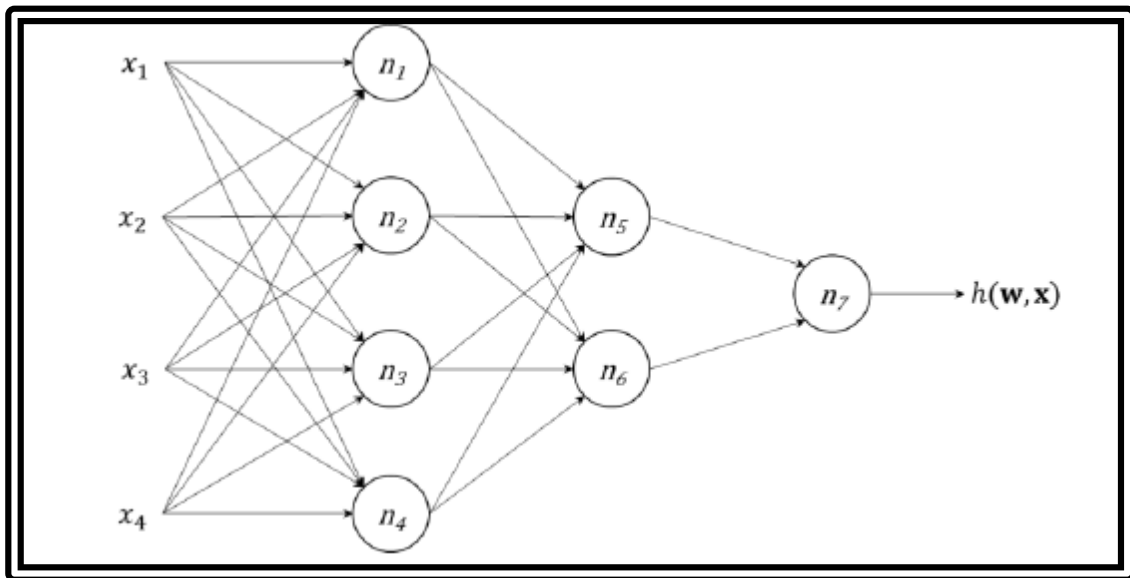


Figura 4.- Ejemplo de Redes Neuronales

La red aprende examinando los registros individuales, generando una predicción para cada registro y realizando ajustes a las ponderaciones cuando realiza una predicción incorrecta. Este proceso se repite muchas veces y la red sigue mejorando sus predicciones hasta haber alcanzado uno o varios criterios de parada.

Al principio, todas las ponderaciones son aleatorias y las respuestas que resultan de la red son, posiblemente, disparatadas. La red aprende a través del **entrenamiento**. Continuamente se presentan a la red ejemplos para los que se conoce el resultado, y las respuestas que proporciona se comparan con los resultados conocidos. La información procedente de esta comparación se pasa hacia atrás a través de la red, cambiando las ponderaciones gradualmente. A medida que progresa el entrenamiento, la red se va haciendo cada vez más precisa en la replicación de resultados conocidos. Una vez entrenada, la red se puede aplicar a casos futuros en los que se desconoce el resultado.

Una vez elegido el modelo teórico es preciso el ajuste del mismo, es por ello que este estudio debe culminar ejecutando los siguientes pasos: elaborar modelos neuronales con distinto número de neuronas; calcular sus respectivos desempeños predictivos por medio de un indicador estadístico; y establecer el modelo más adecuado a través de la comparación de los valores de este indicador para finalmente comprobar la efectividad de cada modelo como método de estimación de recaída de cáncer de mama.

Asimismo, se realizarán varias pruebas con diferentes conjuntos de prueba y entrenamiento, esto con el fin de comparar resultados con las diversas variables mostradas en las tablas 2 y 3.

5.- CONCLUSIONES

Predecir el riesgo de recurrencia del cáncer de mama es crucial para elegir los métodos de tratamiento adecuados, así como reducir la morbilidad y la mortalidad.

La edad del paciente, tamaño del tumor, número de ganglios axilares, grado de diferenciación celular, expresión de receptores hormonales, expresión o amplificación de HER2, **Ki 67**, subtipo de cáncer de mama (todos ellos obtenidos en el momento del diagnóstico), número de embarazos, edad del primer embarazo y la edad de menarquía (primera menstruación), de pacientes que presentaron un diagnóstico de Cáncer de mama pueden constituir factores de riesgo de recaída.

El planteamiento de determinar el valor en la predicción de la recaída de cada una de estas variables se antoja fundamental para abordar el tratamiento individualizado de este tipo de pacientes. Para ello consideramos fundamental la aplicación de tecnologías que nos ayuden al clínico en su diagnóstico posterior.

Aunque a menudo se identifica la IA como un concepto futurista, lo cierto es que es una tecnología que ya se emplea hoy en todo tipo de áreas, entre ellas, la oncología. Gracias a la digitalización de grandes cantidades de datos, el desarrollo de algoritmos de AA y la mejora de la capacidad de cálculo de los ordenadores de las últimas décadas, la IA puede abrir excelentes oportunidades para la automatización de tareas, la aplicación de la medicina de precisión o el avance en la investigación a través de la detección de patrones complejos en bases de datos médicas.

Sin embargo, aún queda un largo camino que recorrer para aplicar de manera generalizada estas técnicas en la práctica clínica, para lo cual será fundamental la generación de grandes bases de datos con información de calidad y la evaluación e integración de la IA en contextos clínicos realistas; por ello, conocer en qué consiste la IA y sus aplicaciones en nuestro campo se antoja esencial para la oncología presente y futura.

Hay algunos obstáculos contra un uso más amplio de este tipo de herramienta. En primer lugar, se necesitan datos de buena calidad, que deben someterse a un proceso de uniformización. Es necesario disponer de cohortes exhaustivas con acceso simplificado, que contengan datos estructurados, tanto para el entrenamiento de los algoritmos como para la validación, que es un paso crucial para poner a prueba el potencial de generalización del modelo.

Para hacer realidad la colaboración entre los médicos y la IA, un buen comienzo sería constituir equipos interdisciplinarios, en los que puedan trabajar juntas las distintas partes encargadas de las actividades relacionadas con los proyectos en una atmósfera abierta de intercambio.

En la actualidad, las redes neuronales artificiales poseen un amplio campo de investigación y desarrollo. Este es otro motivo para seguir realizando estudios sobre modelación neuronal de la recaída de cáncer de mama. En un futuro se espera que se puedan usar los modelos neuronales de recaída en tiempo real, y que su uso llegue a ser una alternativa a los métodos convencionales de estimación de la recaída de cáncer de mama.

No obstante, serán necesarias estrategias de validación que permitan decidir qué técnicas son las más apropiadas según el problema en cuestión.

6.- BIBLIOGRAFIA

1. Tomka M, Sedlakova O, Reinerova M, Veselovska Z, Stevurkova V, Bartosova Z, et al. Mutation screening of the BRCA1 gene in Slovak patients. *Neoplasma*. 2001;48(6):451-5.
2. Kreklau A, Nel I, Kasimir-Bauer S, Kimmig R, Frackenpohl AC, Aktas B. An Observational Study on Breast Cancer Survival and Lifestyle Related Risk Factors. *In Vivo*. 2021;35(2):1007-15.
3. Pedersen RN, Esen BO, Mellemkjaer L, Christiansen P, Ejlersen B, Lash TL, et al. The Incidence of Breast Cancer Recurrence 10-32 Years After Primary Diagnosis. *J Natl Cancer Inst*. 2022;114(3):391-9.
4. Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, et al. Breast cancer intrinsic subtype classification, clinical use and future trends. *Am J Cancer Res*. 2015;5(10):2929-43.
5. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-updagger. *Ann Oncol*. 2019;30(8):1194-220.
6. Ohara M, Matsuura K, Akimoto E, Noma M, Doi M, Nishizaka T, et al. Prognostic value of Ki67 and p53 in patients with estrogen receptor-positive and human epidermal growth factor receptor 2-negative breast cancer: Validation of the cut-off value of the Ki67 labeling index as a predictive factor. *Mol Clin Oncol*. 2016;4(4):648-54.
7. Luo J, Wu M, Gopukumar D, Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomed Inform Insights*. 2016;8:1-10.
8. Millard LA, Flach PA, Higgins JP. Machine learning to assist risk-of-bias assessments in systematic reviews. *Int J Epidemiol*. 2016;45(1):266-77.
9. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2(10):719-31.
10. Raza K, Hasan AN. A comprehensive evaluation of machine learning techniques for cancer class prediction based on microarray data. *Int J Bioinform Res Appl*. 2015;11(5):397-416.
11. Petrick N, Sahiner B, Chan HP, Helvie MA, Paquerault S, Hadjiiski LM. Breast cancer detection: evaluation of a mass-detection algorithm for computer-aided diagnosis -- experience in 263 patients. *Radiology*. 2002;224(1):217-24.
12. Li X, Abramson RG, Arlinghaus LR, Chakravarthy AB, Abramson V, Mayer I, et al. An algorithm for longitudinal registration of PET/CT images acquired during neoadjuvant chemotherapy in breast cancer: preliminary results. *EJNMMI Res*. 2012;2(1):62.

13. Eden KB, Ivlev I, Benschung KL, Franta G, Hersh AR, Case J, et al. Use of an Online Breast Cancer Risk Assessment and Patient Decision Aid in Primary Care Practices. *J Womens Health (Larchmt)*. 2020;29(6):763-9.
14. Al-Azzam N, Shatnawi I. Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer. *Ann Med Surg (Lond)*. 2021;62:53-64.
15. Ley de protección de datos personales y garantía de los derechos digitales (Ley orgánica 3/2018, de 5 de diciembre): Editorial ICB.
16. Alnahi H, Alshawhi S. Knowledge discovery in biomedical databases: a machine induction approach. *Comput Methods Programs Biomed*. 1993;39(3-4):343-9.
17. Smith TC, Frank E. Introducing Machine Learning Concepts with WEKA. *Methods Mol Biol*. 2016;1418:353-78.
18. Mazo C, Aura C, Rahman A, Gallagher WM, Mooney C. Application of Artificial Intelligence Techniques to Predict Risk of Recurrence of Breast Cancer: A Systematic Review. *J Pers Med*. 2022;12(9).
19. Lo JY, Baker JA, Kornguth PJ, Floyd CE, Jr. Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features. *Acad Radiol*. 1995;2(10):841-50.
20. Lisboa PJ. A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Netw*. 2002;15(1):11-39.

Autorización de difusión

“La abajo firmante, matriculada en el Diploma Universitario de Digitalización e Innovación en Salud autoriza a las entidades coordinadoras del DUDIS a difundir y utilizar con fines académicos no comerciales y mencionando expresamente a sus autores del presente Trabajo Fin de Diploma, realizado durante el curso académico 2022-2023, bajo la dirección de Tomás Urda Valcárcel, con el objeto de incrementar la difusión, uso e impacto del trabajo y garantizar su preservación y acceso a largo plazo”

