

# Diseño de una aplicación web de predicción de cáncer de mama mediante machine learning



Universitat  
Oberta  
de Catalunya



UNIVERSITAT DE  
BARCELONA

**Alberto Aparicio Rabadán**

MU Bioinf. y Bioest.

Área de trabajo final:  
Bioinformática Estadística y  
Aprendizaje Automático

**Nombre Tutor/a de TF**

Romina Astrid Rebrij

**Profesor/a responsable de  
la asignatura**

Carles Ventura Royo

**Fecha Entrega: 20/06/2023**



Esta obra está sujeta a una licencia de  
Reconocimiento-NoComercial-  
SinObraDerivada [3.0 España de Creative  
Commons](https://creativecommons.org/licenses/by-nc-nd/3.0/es/)

## FICHA DEL TRABAJO FINAL

<b>Título del trabajo:</b>	Diseño de aplicación web mediante machine learning para predicción de cáncer de mama
<b>Nombre del autor:</b>	Alberto Aparicio Rabadán
<b>Nombre del consultor/a:</b>	Romina Astrid Rebrij
<b>Nombre del PRA:</b>	Carles Ventura Royo
<b>Fecha de entrega (mm/aaaa):</b>	20/06/2023
<b>Titulación o programa:</b>	Máster en Bioinformática y Bioestadística
<b>Área del Trabajo Final:</b>	Bioinformática Estadística y Aprendizaje Automático
<b>Idioma del trabajo:</b>	Castellano
<b>Palabras clave</b>	Machine Learning, redes neuronales, cáncer de mama, Streamlit

### Resumen del Trabajo

Este estudio se enfoca en desarrollar un modelo predictivo basado en una red neuronal convolucional e integrar dicho modelo en una aplicación web para diagnosticar cáncer de mama. En primer lugar, se explora mediante una revisión bibliográfica, la actualidad del cáncer de mama y la utilidad de ciertas técnicas de machine learning como las redes neuronales en la investigación de esta enfermedad. Después, mediante el uso de la librería *Scikit-learn*, se desarrolla un modelo predictivo basado en una red neuronal artificial haciendo uso de la base de datos Breast Cancer Wisconsin (Diagnostic) de UCI. Las muestras recogidas en dicha base de datos se obtuvieron mediante biopsia por aspiración en pacientes con tumores y se procesaron mediante procesamiento de imágenes dando lugar a 569 instancias y 30 variables predictoras. Para optimizar y validar la calidad del modelo se utiliza *GridSearchCV* y se dibuja una curva ROC obteniendo valores altos de F1-score y un valor de área bajo la curva cercano a la unidad. Finalmente, utilizando la plataforma Streamlit se diseña una aplicación web que integra el modelo de predicción realizado y permite a cualquier usuario introducir nuevas instancias de datos para realizar predicciones. A pesar de que este tipo de modelos predictivos podría servir de ayuda al diagnóstico temprano del cáncer de mama, se debe ser consciente de las limitaciones de la técnica y se debe utilizar de forma cautelosa.

## **Abstract**

This study focuses on developing a predictive model based on a convolutional neural network and integrating this model into a web application for breast cancer diagnosis. Firstly, a literature review is conducted to explore the current state of breast cancer and the utility of certain machine learning techniques, such as neural networks, in researching this disease. Subsequently, an artificial neural network-based predictive model is developed using the Scikit-learn library and the Breast Cancer Wisconsin (Diagnostic) dataset from UCI. The samples collected in this database were obtained through aspiration biopsy in patients with tumors and were processed through image processing, resulting in 569 instances and 30 predictive features. GridSearchCV and a ROC curve are used to optimize and validate the quality of the model, obtaining high F1-score values and an area under the curve value close to unity. Finally, using the Streamlit platform, a web application is designed integrating the performed predictive model and allowing any user to input new data instances to make predictions. Although these types of predictive models could assist in early breast cancer diagnosis, awareness of the technique's limitations is crucial, and it should be used cautiously.

# Índice

1.	Introducción.....	1
1.1.	Contexto y justificación del Trabajo.....	1
1.2.	Objetivos del Trabajo .....	2
1.3.	Impacto en sostenibilidad, ético-social y de diversidad.....	3
1.4.	Enfoque y método seguido.....	4
1.5.	Planificación del Trabajo .....	4
1.6.	Breve resumen de productos obtenidos.....	5
1.7.	Breve descripción de los otros capítulos de la memoria .....	5
2.	Estado del arte .....	7
2.1	El cáncer de mama en la actualidad.....	7
2.2	Machine Learning.....	8
2.3	Redes neuronales.....	10
2.4	Uso de machine learning en la investigación contra el cáncer.....	12
3.	Materiales y métodos .....	14
3.1	Análisis exploratorio de los datos .....	15
3.2	Preprocesamiento de los datos .....	15
3.3	Optimización, prueba y validación del modelo de predicción .....	16
3.4	Diseño de la aplicación web mediante Streamlit .....	19
4.	Resultados .....	20
4.1	Análisis exploratorio de los datos .....	20
4.2	Creación, optimización, prueba y validación del modelo de predicción....	21
5.	Conclusiones y trabajos futuros.....	26
6.	Glosario .....	28
7.	Bibliografía .....	29
8.	Anexo 1.....	33

# Lista de figuras

Figura 1. Diagrama de Gantt que describe la planificación de las diferentes tareas a realizar para el desarrollo completo del modelo de predicción y la aplicación web.....	5
Figura 2. Código utilizado para la generación del informe exploratorio de los datos. ....	15
Figura 3. División de los datos en X (variables predictoras) e Y (variable a predecir).....	16
Figura 4. División de los datos en dos conjuntos: entrenamiento (80% de los datos) y prueba (20%).....	16
Figura 5. Modelo de predicción generado con la función MLPClassifier de la librería Scikit-learn y validación cruzada mediante GridSearchCV .....	17
Figura 6. Entrenamiento del modelo predictivo. ....	18
Figura 7. Uso de la función best_params_ para observar los parámetros óptimos que ha escogido el modelo. ....	18
Figura 8. Código utilizado para calcular la probabilidad de obtener un resultado benigno o maligno.....	18
Figura 9. Código utilizado para la obtención de una curva ROC-AUC con el objetivo de valorar la capacidad del modelo para clasificar las muestras.....	19
Figura 10. Diseño de los distintos botones del menú lateral correspondientes a las variables predictoras. ....	19
Figura 11. Diseño del botón de inicio de la predicción que al ser accionado devuelve la probabilidad expresada en porcentaje de que una muestra sea benigna.....	19
Figura 12. Extracto del informe HTML obtenido con la función pandas-profiling. ....	20
Figura 13. Informe de la función pandas-profiling de la variable diagnosis. Se observan 357 instancias con diagnóstico benigno y 212 maligno.....	20
Figura 14. Extracto del informe realizado con la función pandas-profiling que muestra información sobre la variable "radius_mean" (media del radio nuclear). ....	21
Figura 15. Resultados posibles de la predicción (Benigno o Maligno) y probabilidad de obtener cada uno de ellos con los datos de prueba.....	21
Figura 16. Informe de calidad del modelo predictivo. ....	22
Figura 17. Gráfico de la curva ROC-AUC. Se observa un valor de área bajo la curva de 0,98. ....	23
Figura 18. Pantalla principal de la aplicación web. En la parte izquierda de la misma puede observarse el menú lateral con los diferentes botones para introducir los parámetros de la muestra. Por otro lado, en la parte central, se encuentra ubicado el botón para realizar la predicción.....	24
Figura 19. Pantalla final correspondiente al resultado de la predicción, en ella se puede observar el resultado expresado en porcentaje con dos decimales. ....	25

# 1. Introducció

## 1.1. Contexto y justificación del Trabajo

El cáncer de mama es una de las enfermedades más mortales y prevalentes entre las mujeres a nivel mundial. Se estima que 2.3 millones de mujeres fueron diagnosticadas de esta enfermedad en 2020 (Sung et al., 2021). Es por ello que el diagnóstico precoz y la detección temprana son fundamentales para aumentar las tasas de supervivencia y reducir su morbilidad y mortalidad. Sin embargo, los métodos de diagnóstico actuales, aunque eficaces, pueden ser invasivos, costosos y no siempre están disponibles en regiones con recursos limitados (Drukteinis et al., 2013).

Las técnicas de ML han demostrado ser realmente útiles en el análisis y predicción de enfermedades, incluyendo el cáncer de mama (Ravi et al., 2017). Los algoritmos basados en estas técnicas, ayudan a identificar patrones complejos en los datos y proporcionar diagnósticos precisos basados en características difíciles de detectar por otras vías (Esteva et al., 2019).

De entre todas las técnicas de ML, las Redes Neuronales Artificiales (ANNs) han demostrado ser especialmente eficaces a la hora de clasificar y predecir enfermedades y ya se han hecho estudios como el de Spanhol y sus colaboradores en 2016 en el que se prueba esta técnica con el cáncer de mama.

El objetivo de este trabajo de fin de máster es desarrollar un modelo predictivo de cáncer de mama utilizando ANNs e integrar dicho modelo en una aplicación web que permita ingresar nuevas instancias de datos para realizar predicciones en tiempo real. La base de datos que se va utilizar es Breast Cancer Wisconsin (Diagnostic) de UCI que constituye una de las bases más completas y utilizadas en la investigación de este tipo de cáncer (Dua y Graff, 2019). Dentro de esta base de datos, el conjunto de datos que se utilizarán en este trabajo proviene del estudio de Mangasarian et al., (1994) y corresponde a 569 imágenes procesadas de muestras de glándulas mamarias obtenidas mediante aspiración con aguja fina (FNA) (Mangasarian, Street y Wolberg, 1994).

En conclusión, este trabajo podría contribuir al diagnóstico temprano de la enfermedad, tiene un aplicación práctica y podría ser útil tanto para pacientes como para profesionales de la salud, además el uso de técnicas innovadoras de machine learning y el diseño de la aplicación web pueden contribuir al avance de la investigación en el campo del cáncer de mama. Es importante señalar que esta aplicación no pretende reemplazar el diagnóstico que realiza el médico sino complementarlo y ayudar en la toma de decisiones clínicas.

## 1.2. Objetivos del Trabajo

### **Objetivos generales:**

1. Crear un modelo predictivo mediante machine learning para predecir cáncer de mama utilizando la base de datos “Breast Cancer Wisconsin (Diagnostic)” de UCI.
2. Desarrollar una aplicación web que permita a los usuarios ingresar nuevas instancias de datos para realizar predicciones utilizando el modelo predictivo desarrollado.

### **Objetivos específicos:**

- 1.a. Realizar un análisis exploratorio y un preprocesamiento de los datos de pacientes a utilizar para entender su estructura y adecuarlos al modelo.
- 1.b. Diseñar y entrenar una red neuronal convolucional utilizando un conjunto de datos de la base de datos “Breast Cancer Wisconsin (Diagnostic)” de UCI y evaluar su rendimiento.
- 2.a. Integrar el modelo predictivo creado en una aplicación web utilizando el lenguaje de programación Python y crear una interfaz de usuario intuitiva con Streamlit que permita a los usuarios ingresar nuevas instancias de datos para realizar predicciones.
- 2.b. Poner la aplicación a disposición de los usuarios en la plataforma Streamlit.



### 1.3. Impacto en sostenibilidad, ético-social y de diversidad

Este Trabajo de Fin de Máster (TFM) cumple con el objetivo número 3 "Salud y Bienestar", de los Objetivos de Desarrollo Sostenible (ODS) contribuyendo a mejorar la detección del cáncer de mama y, por lo tanto, a mejorar la salud de las mujeres a nivel mundial. No obstante, veamos las distintas dimensiones por separado:

**Sostenibilidad y Huella Ecológica:** En cuanto a la sostenibilidad medioambiental, el desarrollo y explotación de un modelo de machine learning y su aplicación web no tiene un impacto directo en el agotamiento de materias primas, residuos o contaminación. Sin embargo, hay que tener en cuenta el consumo energético de los servidores que albergan la aplicación y realizan los cálculos del modelo. Para mitigar este impacto, se puede optar por proveedores de servicios de computación en la nube que utilicen energías renovables y tengan políticas de sostenibilidad (Google Cloud, 2021).

**Aspectos Ético-Sociales:** En la dimensión ético-social, en este trabajo se utilizan datos relacionados con la salud pero en ningún caso datos personales por lo que se garantiza el cumplimiento de las leyes de protección de datos, establecido por el GDPR en Europa (Reglamento General de Protección de Datos, 2016). Además, este modelo de predicción no se contempla como un sustituto del diagnóstico médico, sino como una herramienta de apoyo.

**Diversidad y Derechos Humanos:** Respecto a la diversidad, este TFM está diseñado para ser utilizado por todos, independientemente de su género, raza, religión, orientación sexual, etc., lo que contribuye al ODS 10, "Reducción de las Desigualdades". Además, al estar disponible en una plataforma web, puede ser accesible para personas con discapacidades, siguiendo las directrices de accesibilidad de la Web (WCAG, 2021).

En conclusión, este TFM tiene un impacto positivo en términos de sostenibilidad, aspectos ético-sociales y diversidad, aunque es importante tener en cuenta los posibles impactos negativos relacionados con el consumo energético y la privacidad de los datos.

#### 1.4. Enfoque y método seguido

Este trabajo sigue un enfoque basado en el aprendizaje automático supervisado ya que la base de datos mencionada anteriormente, contiene etiquetas de clasificación que nos permiten entrenar un modelo supervisado. Constaría de las siguientes partes:

- Análisis exploratorio de los datos: para entender su estructura (número de instancias y variables predictoras)
- Preprocesamiento de los datos: división de los datos en dos grupos, uno de prueba y otro de entrenamiento para la red neuronal.
- Diseño, entrenamiento y validación del modelo predictivo: desarrollar una red neuronal artificial adecuada para el trabajo y entrenarla utilizando la base de datos mencionada anteriormente.
- Implementación del modelo predictivo desarrollado en un aplicación web: diseñar la aplicación utilizando Python y Streamlit e integrar el modelo predictivo de modo que permita a los usuarios ingresar nuevas instancias de datos para obtener predicciones.

Para el modelo predictivo, se hace uso de una red neuronal convolucional y para su diseño y validación se utilizan las librerías: Pandas, Numpy y Scikit-learn.

Por otro lado, para desarrollar la aplicación web se utiliza Streamlit ya que es un marco de trabajo de uso abierto diseñado para facilitar el proceso de creación de aplicaciones basadas en modelos de aprendizaje automático. Esta plataforma nos permite hacer pública la aplicación favoreciendo su accesibilidad y difusión. Además, cuenta con una gran comunidad de desarrolladores por lo que existe mucha información sobre su uso.

#### 1.5. Planificación del Trabajo

La planificación del trabajo se muestra mediante un diagrama de Gantt en la figura 1.

# Diseño de una aplicación web de predicción de cáncer de mama mediante *machine learning*



**Figura 1. Diagrama de Gantt que describe la planificación de las diferentes tareas a realizar para el desarrollo completo del modelo de predicción y la aplicación web.**

## 1.6. Breve resumen de productos obtenidos

Los productos obtenidos de este trabajo son:

- Un modelo predictivo de cáncer de mama generado a partir de una red neuronal convolucional basado en un conjunto de datos de la base de datos de cáncer de mama “Breast Cancer Wisconsin”.
- Una aplicación web con el modelo predictivo integrado que permite ingresar nuevas instancias de datos para predicción de cáncer de mama.

## 1.7. Breve descripción de los otros capítulos de la memoria

- Estado del arte: consta de una revisión bibliográfica sobre los siguientes temas: el cáncer de mama en la actualidad; origen y tipos de técnicas dentro del campo de machine learning; origen, tipos y utilidad de las redes neuronales y finalmente aplicación de machine learning en investigación contra el cáncer de mama.
- Materiales y métodos: en este capítulo se detallan los distintos pasos llevados a cabo para el diseño del modelo predictivo y la aplicación web.
- Resultados: este apartado muestra los productos resultantes del procedimiento descrito en el capítulo anterior.

- Bibliografía: muestra la lista de referencias, citadas en el texto, ordenada alfabéticamente según las normas APA.
- Anexo 1: en este apartado encontramos el enlace al código fuente del modelo predictivo y la aplicación web desarrollada, así como el enlace de acceso a la aplicación.

## 2. Estado del arte

### 2.1 El cáncer de mama en la actualidad

Como se ha mencionado anteriormente, el cáncer de mama es una de las enfermedades más comunes entre las mujeres a nivel mundial y se estima que cerca de 685.000 mujeres mueren a causa de esta enfermedad, representando el 15% de todas las muertes por cáncer entre las mujeres (OMS, 2021). Aunque es más común en las mujeres, los hombres también pueden desarrollarlo. La prevalencia de esta afección va en aumento, hasta el punto de que se estima que el número de casos de cáncer de mama llegará a ser de 3,2 millones de casos anuales a partir de 2050 (Harbeck y Gnant, 2017).

Concretamente en España, el cáncer de mama es el tipo de cáncer más frecuente entre las mujeres. Según el Observatorio del Cáncer de la Asociación Española Contra el Cáncer (AECC), en 2020 se diagnosticaron más de 32.953 nuevos casos (AECC, 2021). A pesar de ser una enfermedad muy extendida, la tasa de supervivencia ha mejorado significativamente en las últimas décadas. El índice de supervivencia a los 5 años en España es del 90%, una de las tasas más altas de Europa, gracias a las mejoras en la detección temprana y los tratamientos (AECC, 2021).

El cáncer de mama es una enfermedad heterogénea con varios subtipos, cada uno de los cuales tiene diferentes implicaciones terapéuticas y pronósticos (Perou et al., 2000). El tratamiento del cáncer de mama puede implicar una combinación de cirugía, radioterapia, quimioterapia, terapia hormonal y terapia dirigida, y el plan de tratamiento se personaliza para cada paciente en función de las características del tumor y de la paciente (National Cancer Institute, 2019).

Las campañas de concienciación y los programas de detección temprana, como las mamografías, han tenido un impacto significativo en el diagnóstico del cáncer de mama. Sin embargo, existe la necesidad de desarrollar herramientas más precisas y menos invasivas para mejorar aún más las tasas de detección y reducir las biopsias innecesarias (Drukteinis et al., 2013).

## 2.2 Machine Learning

El Machine Learning (ML), o aprendizaje automático, se basa en el desarrollo de algoritmos y modelos estadísticos que permiten a los ordenadores aprender a realizar tareas sin ser explícitamente programados para ello (Samuel, 1959). Su principal objetivo es permitir a las máquinas mejorar su rendimiento o adaptarse a nuevas circunstancias de manera autónoma a medida que se exponen a más datos.

El término "Machine Learning" fue acuñado por Arthur Samuel en 1959, durante su trabajo en IBM, donde desarrolló un programa de ordenador para jugar a las damas que mejoraba su rendimiento con el tiempo (Samuel, 1959). Sin embargo, el concepto de aprendizaje automático tiene sus raíces en la obra de Alan Turing, quien en los años 40 ya especulaba con la idea de que las máquinas podrían aprender (Turing, 1950). Desde entonces, el campo del ML ha experimentado un crecimiento significativo y ha evolucionado pasando por distintas fases. Durante los años 60 y 70, el enfoque principal eran los métodos simbólicos basados en reglas y la construcción de sistemas expertos. La década de los 80 estuvo dominada por los sistemas basados en el conocimiento, mientras que los años 90 supusieron el auge de las redes neuronales y el aprendizaje profundo, gracias a los avances en los algoritmos y el aumento de la capacidad de cálculo (Goodfellow et al., 2016).

En la actualidad, el ML se ha convertido en una herramienta esencial en muchas industrias y campos de la ciencia, desde la visión por computadora y el procesamiento del lenguaje natural hasta la bioinformática y la medicina. El aprendizaje profundo, una subcategoría del ML que se inspira en la estructura del cerebro humano para crear redes neuronales artificiales, ha demostrado un rendimiento excepcional en una amplia gama de tareas y ha impulsado gran parte del reciente éxito de la IA (LeCun et al., 2015).

En cuanto a las técnicas, el ML se puede dividir en tres categorías principales: aprendizaje supervisado, no supervisado y por refuerzo. El aprendizaje supervisado implica entrenar un modelo con datos de entrada etiquetados,

mientras que el aprendizaje no supervisado busca patrones en datos no etiquetados. El aprendizaje por refuerzo es un enfoque en el que un agente aprende a tomar decisiones basándose en recompensas y castigos (Sutton y Barto, 2018).

El ML presenta muchas ventajas, como su capacidad para manejar grandes cantidades de datos, su adaptabilidad a nuevas entradas y su habilidad para descubrir patrones y relaciones complejas. Sin embargo, también tiene desventajas. Entre ellas se incluyen su dependencia de la calidad y cantidad de los datos de entrenamiento, el riesgo de sobreajuste y subajuste, y los desafíos en la interpretación de los modelos, especialmente en el caso de los modelos de aprendizaje profundo, a menudo denominados "cajas negras" debido a su falta de transparencia (Doshi-Velez y Kim, 2017).

El futuro del Machine Learning promete ser emocionante, ya que sigue siendo un campo de rápido desarrollo con muchas oportunidades para la innovación. Se estima que las técnicas de aprendizaje automático van a volverse cada vez más sofisticadas, capaces de manejar una mayor cantidad de datos, más complejos y proporcionar modelos más precisos y eficientes. En particular, el aprendizaje profundo, que ya ha logrado avances significativos en áreas como el procesamiento del lenguaje natural, continuará evolucionando y encontrando nuevas aplicaciones (Goodfellow et al., 2016).

También se prevé que se preste cada vez más atención a cuestiones como la explicabilidad y la justicia en el ML. La explicabilidad se refiere a la capacidad de entender y explicar las decisiones tomadas por los modelos de ML. Esto es especialmente importante en contextos sensibles, como la medicina o la toma de decisiones legales, donde es necesario poder explicar y justificar las decisiones tomadas por las máquinas. Por otro lado, la justicia en el ML se refiere al desarrollo de modelos que no refuercen ni perpetúen las desigualdades existentes, un desafío cada vez más reconocido en la comunidad de IA (Barocas et al., 2019).

Otra tendencia emergente en el ML es la del aprendizaje federado (también llamado aprendizaje colaborativo), que permite el entrenamiento de modelos de

ML en múltiples dispositivos descentralizados, mientras se mantienen los datos en el dispositivo original. Este enfoque podría ayudar a superar algunos de los desafíos relacionados con la privacidad y la seguridad de los datos en el ML (Konečný et al., 2016).

En resumen, el Machine Learning, desde sus humildes inicios hasta su estado actual de desarrollo avanzado, ha recorrido un largo camino y continúa evolucionando rápidamente, con su potencial para transformar industrias y revolucionar nuestra comprensión del mundo.

## 2.3 Redes neuronales

Las redes neuronales artificiales (ANNs, por sus siglas en inglés) son modelos de Machine Learning que se inspiran en las redes neuronales biológicas presentes en el cerebro humano. Estas redes están diseñadas para imitar la forma en que las neuronas biológicas interactúan y procesan la información, con el objetivo de aprender a realizar tareas complejas a partir de un conjunto de datos (LeCun et al., 2015).

Las ANNs se componen de unidades de procesamiento interconectadas, llamadas nodos o "neuronas", organizadas en diferentes capas: una capa de entrada, una o más capas ocultas y una capa de salida. Cada neurona recibe una serie de entradas, las pondera en función de sus respectivos pesos, aplica una función de activación y luego transmite la salida a las neuronas de la siguiente capa. El aprendizaje en las redes neuronales se logra ajustando los pesos de las conexiones a través de un proceso llamado retropropagación, que utiliza un algoritmo de descenso del gradiente para minimizar la diferencia entre la salida de la red y la salida deseada (Goodfellow et al., 2016).

Existen diferentes tipos de redes neuronales, cada uno de los cuales tiene sus propias características y aplicaciones específicas. Las redes neuronales feedforward son las más simples y comunes, en las que la información se mueve en una sola dirección, desde la capa de entrada hasta la capa de salida, sin bucles. Las redes neuronales convolucionales (CNNs) son un tipo especial de redes feedforward que son especialmente efectivas en tareas de



procesamiento de imágenes, ya que están diseñadas para procesar datos con una topología de cuadrícula (Krizhevsky et al., 2012).

Por otro lado, las redes neuronales recurrentes (RNNs) son un tipo de ANN que incorporan bucles, permitiendo que la información persista de un paso a otro en el tiempo, lo que las hace adecuadas para tareas que implican datos secuenciales, como el procesamiento del lenguaje natural (Hochreiter y Schmidhuber, 1997). Dentro de las RNNs, las redes de memoria a corto y largo plazo (LSTM) son una variante popular que supera algunos de los desafíos de las RNNs tradicionales, como el problema del desvanecimiento del gradiente (Hochreiter y Schmidhuber, 1997).

Las redes neuronales han demostrado ser herramientas muy efectivas para tareas como el reconocimiento de voz y la traducción automática. Gracias a sus capacidades de aprendizaje profundo, pueden aprender representaciones complejas y jerárquicas de los datos, permitiéndoles manejar datos de alta dimensionalidad y descubrir patrones y relaciones intrincadas que pueden ser inaccesibles para los métodos de aprendizaje automático tradicionales (LeCun et al., 2015).

En la actualidad, las redes neuronales son una de las áreas más activas de la investigación en IA. Su capacidad para aprender a partir de grandes cantidades de datos y manejar tareas complejas las ha hecho indispensables en muchos campos y aplicaciones, desde la medicina y la biología hasta la industria y el entretenimiento. Por ejemplo, en medicina, las redes neuronales se utilizan para detectar enfermedades en imágenes médicas con una precisión comparable o incluso superior a la de los especialistas humanos (Esteva et al., 2019). En la industria, se utilizan para optimizar procesos, predecir fallas en equipos y mejorar la eficiencia energética (Kusiak, 2018). En el entretenimiento, son la base de los algoritmos de recomendación que se utilizan en plataformas de streaming como Netflix y Spotify (Gomez-Urbe y Hunt, 2016). Además, las redes neuronales constituyen el componente central de los sistemas de conducción autónoma, donde se utilizan para interpretar los datos de los

sensores del vehículo, detectar obstáculos y tomar decisiones durante la conducción (Bojarski et al., 2016).

A pesar de su éxito, las redes neuronales también presentan desafíos. Entre ellos, se incluyen su demanda de grandes cantidades de datos y capacidad de cálculo, su tendencia al sobreajuste y la dificultad para interpretar los resultados, el "problema de la caja negra" (Doshi-Velez y Kim, 2017). A pesar de estos desafíos, el futuro de las redes neuronales es prometedor y conforme la investigación avanza y las técnicas mejoran, podemos ver redes neuronales cada vez más potentes y capaces de aprender de manera más eficiente y efectiva.

## 2.4 Uso de machine learning en la investigación contra el cáncer

Los avances recientes en estas técnicas de aprendizaje profundo están permitiendo a los científicos e investigadores médicos descubrir nuevas formas de diagnosticar, tratar y monitorizar esta afección compleja y prevalente.

Un área en la que el ML y las redes neuronales han demostrado ser particularmente útiles es la detección temprana y el diagnóstico del cáncer de mama. Un estudio de 2020 llevado a cabo por McKinney et al., (2020) utilizó una red neuronal convolucional (CNN) para interpretar las imágenes de mamografía, demostrando que el sistema superaba la precisión de los radiólogos humanos. El modelo CNN fue capaz de reducir la tasa de falsos positivos y falsos negativos, lo que potencialmente podría llevar a un diagnóstico más preciso y a un tratamiento más temprano y eficaz del cáncer de mama (McKinney et al., 2020). En una investigación similar, Ardila et al., (2019) usaron una red neuronal para detectar cáncer de pulmón en tomografías computarizadas de tórax, demostrando la utilidad de este tipo de técnicas. Además, las redes neuronales también se están utilizando para predecir el pronóstico del cáncer de mama. En un estudio realizado por Chaudhary et al., (2018), se utilizó una red neuronal para predecir la supervivencia en pacientes con cáncer de hígado. El modelo pudo identificar patrones en los datos clínicos y genéticos que podrían indicar un mejor o peor pronóstico, lo que puede ayudar a los médicos a personalizar los tratamientos.

Otro uso prometedor del ML en la investigación del cáncer de mama es el desarrollo de nuevos tratamientos. Algunos estudios están explorando cómo el ML puede usarse para identificar nuevas dianas terapéuticas o para predecir la respuesta a los tratamientos existentes. Por ejemplo, el trabajo de Kourou et al., (2015) utilizó técnicas de ML para predecir la eficacia de diferentes tratamientos de quimioterapia en función de los perfiles genéticos de los pacientes.

Además de estas aplicaciones directas, el ML también está ayudando a los investigadores a entender mejor la biología del cáncer de mama. Un estudio de Yuan et al., (2020) utilizó redes neuronales para analizar grandes conjuntos de datos de expresión génica, lo que permitió descubrir nuevos genes y vías metabólicas implicadas en el cáncer de mama.

A pesar de que estas aplicaciones pueden ser realmente útiles, hay que tener en cuenta los términos de explicabilidad y justicia que se han mencionado anteriormente, especialmente en el campo de la salud, donde los errores pueden tener graves consecuencias y donde los médicos y los pacientes exigen una comprensión clara de cómo y por qué se toman las decisiones clínicas (Holzinger et al., 2017). Por tanto, el desarrollo de técnicas de aprendizaje automático más interpretables es actualmente un área activa de investigación.

Por otro lado, trasladar los modelos de ML del laboratorio a la clínica también es un desafío. Aunque los modelos pueden funcionar bien en los conjuntos de datos de entrenamiento, pueden no generalizar bien a nuevas poblaciones de pacientes. Por lo tanto, es importante realizar estudios de validación en diferentes contextos y poblaciones antes de adoptar estos modelos en la práctica clínica (Wiens et al., 2019).

### 3. Materiales y métodos

Como se mencionó anteriormente, los datos que se utilizan en este trabajo provienen de un estudio en el que se analizaban mediante procesamiento de imágenes los núcleos de las células de glándulas mamarias obtenidas por biopsia por aspiración (Mangasarian, Street y Wolberg, 1994). En dicho estudio, utilizaron un árbol de decisión denominado "Multisurface Method Tree", que es un algoritmo de clasificación supervisado basado en el método de múltiples superficies (MSM). Con esta técnica dividieron 569 muestras en benignas y malignas utilizando 30 variables relacionadas con el núcleo de las células mamarias y concluyeron que estudiar factores como el perímetro nuclear puede ayudar en el diagnóstico del cáncer de mama sustituyendo a técnicas que más invasivas como la biopsia del ganglio linfático centinela. Calcularon las diferentes características del núcleo celular de la siguiente forma (Wolberg, Street y Mangasarian, 1999):

- a) El radio se calculó con el promedio la longitud de los segmentos de línea radiales desde el centro de la masa nuclear hasta cada uno de los puntos del borde nuclear.
- b) El perímetro se midió como la distancia alrededor del borde nuclear.
- c) El área se midió contando el número de píxeles en el interior del borde nuclear y sumando la mitad de los píxeles en el perímetro.
- d) El perímetro y el área se combinaron para obtener una medida de la compactación de los núcleos celulares utilizando la siguiente fórmula:  $\text{perímetro}^2 / \text{área}$ .
- e) La suavidad se cuantificó midiendo la diferencia entre la longitud de cada radio y la longitud media de los radios adyacentes.
- f) La concavidad se determinó midiendo el tamaño de las hendiduras existentes en el borde nuclear.
- g) Los puntos cóncavos contaban el número de puntos en el borde nuclear que se encuentran en una hendidura.
- h) La simetría se midió encontrando la diferencia relativa en longitud entre segmentos de línea perpendiculares a ambos lados del eje mayor.
- i) La dimensión fractal se aproximó utilizando la "aproximación de la línea de costa" que mide la irregularidad del borde nuclear.

- j) La textura se midió encontrando la varianza de las intensidades en la escala de grises de los píxeles.

Se calculó el valor medio, el valor más alto (resultante de la media de los tres valores más grandes de todos los núcleos de la imagen analizada) y la desviación estándar para cada característica dando un total de 30 variables.

En este trabajo, procederemos a crear un modelo de predicción a partir de los datos de las 569 imágenes utilizando una red neuronal convolucional y se integrará dicho modelo en una aplicación web con el objetivo de que otros investigadores puedan ingresar nuevas instancias de datos para realizar predicciones. Como se ha mencionado anteriormente, las redes convolucionales son excepcionalmente buenas en la identificación de patrones en los datos cuando se trabaja con volúmenes altos de datos como es este caso.

La versión de Python utilizada en este trabajo es la 3.10.11 y se desarrolla en jupyter notebook y Visual Studio Code. En este escrito sólo se muestra el código más relevante, pero se proporciona un enlace a GitHub en el que se puede visualizar el código fuente (Anexo 1).

### 3.1 Análisis exploratorio de los datos

En primer lugar, se realiza un análisis exploratorio de los datos. Para ello, se genera un “DataFrame” cargando los datos y como observamos en la figura 2, mediante la librería “Pandas profiling” obtenemos un informe HTML de los datos en el que podemos visualizar información relevante del conjunto de datos como el número de instancias, variables y sus tipos y la presencia o ausencia de datos perdidos.

```
from pandas_profiling import ProfileReport
prof = ProfileReport(df)
prof.to_file(output_file='output.html')
```

Figura 2. Código utilizado para la generación del informe exploratorio de los datos.

### 3.2 Preprocesamiento de los datos

Como hemos observado en el análisis exploratorio, todas las variables predictoras son numéricas por lo que no es necesario realizar cambios en el tipo de las mismas para trabajar con ellas. Sin embargo, es preciso separar la

variable categórica y eliminar la variable numérica correspondiente al ID del paciente. Para ello dividimos los datos en X (variables predictoras) e Y (variable “diagnosis”) como se observa en la figura 3.

```
X = df.drop("diagnosis", axis=1)
y = df["diagnosis"]
```

**Figura 3. División de los datos en X (variables predictoras) e Y (variable a predecir)**

Se podría aplicar algún método de “feature selection” o selección de variables para determinar aquellas variables que son más relevantes para la predicción, pero la red neuronal que vamos a generar es capaz de aprender y seleccionar dichas variables.

Finalmente, para poder crear la red neuronal, primero es necesario dividir los datos en dos grupos (Figura 4): datos de entrenamiento y datos de prueba. Según lo explicado por Özesmi, Tan y Özesmi (2006) en su trabajo sobre los problemas metodológicos a la hora de construir, entrenar y probar redes neuronales artificiales, si se trabaja con un volumen bajo de muestras del mismo orden que el número de variables predictoras, existe el riesgo de realizar un sobreentrenamiento (overfitting) lo que afecta a la exactitud de la predicción. Sin embargo, en el caso de este trabajo se cuenta con una alta cantidad de instancias (569) y 30 variables predictoras, por lo que ese riesgo no es tan importante.

```
X_train, X_test, y_train, y_test= train_test_split(X, y, test_size=0.2, random_state=1)
```

**Figura 4. División de los datos en dos conjuntos: entrenamiento (80% de los datos) y prueba (20%).**

### 3.3 Optimización, prueba y validación del modelo de predicción

A continuación, se crea el modelo de predicción (Figura 5) utilizando la función “MLPClassifier” de la librería “Scikit-learn” (Scikit-learn developers, 2023). Los parámetros que conforman la función de la figura X son los siguientes:

- **Hidden\_layer\_sizes:** Se realizaron diferentes pruebas cambiando el número de capas ocultas y neuronas, pero no se observaron apenas diferencias en la capacidad de predicción. Los números hacen referencia al número de neuronas en cada capa.

- Activation: Hace referencia a la función de activación de las capas, se utiliza la función de la tangente hiperbólica (tanh) y la función de activación ReLu (Rectified Linear Units) ya que se ha visto en estudios como el de LeCun et al., (2015) que ayuda a minimizar los errores en la predicción.
- Solver: este parámetro está involucrado en la optimización del modelo proporcionando un gradiente descendiente estocástico que reduce la función de pérdida de la red.
- Alpha: se optó por los valores predeterminados de 0,0001 y 0,5.
- Learning rate: Se optó por los parámetros más usados con un gradiente descendiente estocástico (Scikit-learn developers, 2023), “constant” y “adaptive”.

Además, se realizó una validación cruzada utilizando GridSearchCV, que es una función que optimiza los hiperparámetros de la red neuronal (Figura 5).

```
mlp_gs = MLPClassifier(max_iter=2000, random_state=1)
parameter_space = {
    'hidden_layer_sizes': [(10,30,10),(20,),(20,20,20),(10,10,10,10)],
    'activation': ['tanh', 'relu'],
    'solver': ['sgd', 'adam'],
    'alpha': [0.0001, 0.05],
    'learning_rate': ['constant','adaptive'],
}
clf = GridSearchCV(mlp_gs, parameter_space, n_jobs=-1, cv=5)
```

Figura 5. Modelo de predicción generado con la función MLPClassifier de la librería Scikit-learn y validación cruzada mediante GridSearchCV

Una vez optimizado y validado el modelo podemos pasar a entrenar el modelo con la función de entrenamiento (Figura 6).

```
clf.fit(X_train, y_train)
✓ 29.0s

GridSearchCV
GridSearchCV(cv=5, estimator=MLPClassifier(max_iter=2000, random_state=1),
n_jobs=-1,
param_grid={'activation': ['tanh', 'relu'],
'alpha': [0.0001, 0.05],
'hidden_layer_sizes': [(10, 30, 10), (20,),
(20, 20, 20),
(10, 10, 10, 10)],
'learning_rate': ['constant', 'adaptive'],
'solver': ['sgd', 'adam']})
estimator: MLPClassifier
MLPClassifier(max_iter=2000, random_state=1)
MLPClassifier(max_iter=2000, random_state=1)
```

Figura 6. Entrenamiento del modelo predictivo.

Una vez entrenado, podemos visualizar qué parámetros han resultado ser los más óptimos tras la validación cruzada (Figura 7).

```
clf.best_params_
✓ 0.0s
{'activation': 'relu',
'alpha': 0.0001,
'hidden_layer_sizes': (20,),
'learning_rate': 'constant',
'solver': 'adam'}
```

Figura 7. Uso de la función `best_params_` para observar los parámetros óptimos que ha escogido el modelo.

Después, podemos obtener información como por ejemplo la probabilidad de obtener un resultado benigno y maligno con los datos de prueba (Figura 8).

```
clf.predict_proba(X_test[:1])
y_pred = clf.predict(X_test)
```

Figura 8. Código utilizado para calcular la probabilidad de obtener un resultado benigno o maligno

Para finalizar con la validación del método se optó por analizar un informe de clasificación (que contiene parámetros evaluadores de la calidad del modelo como el f1-score) y realizar una curva ROC – AUC (Receiver Operating Characteristic y Area Under the Curve) que es un método para evaluar la eficiencia de un algoritmo de clasificación. Consiste en un gráfico que explica,



en este caso, con qué precisión puede distinguir el modelo entre pacientes con cáncer o sin cáncer.

```
RocCurveDisplay.from_estimator(clf, X_test, y_test, pos_label="B")
```

Figura 9. Código utilizado para la obtención de una curva ROC-AUC con el objetivo de valorar la capacidad del modelo para clasificar las muestras.

Finalmente, creamos un archivo “.joblib” con el fin de guardar en un archivo el modelo de predicción creado para después poder integrarlo fácilmente en la aplicación web que vamos a diseñar.

### 3.4 Diseño de la aplicación web mediante Streamlit

Streamlit es un “framework” de uso abierto para la creación de aplicaciones web creado para facilitar el diseño de aplicaciones basadas en modelos de machine learning. Para llevar a cabo el diseño, se importa el archivo .joblib que hemos generado anteriormente y se utiliza la librería streamlit para diseñar los botones correspondientes a las variables predictoras (Figura 10), el botón de inicio de la predicción y el resultado expresado en porcentaje de que la muestra del paciente sea catalogada como “benigna” (Figura 11).

```
radius_mean = st.sidebar.number_input("radius_mean", min_value=0.0, value=0.0, step=0.1)
texture_mean = st.sidebar.number_input("texture_mean", min_value=0.0, value=0.0, step=0.1)
perimeter_mean = st.sidebar.number_input("perimeter_mean", min_value=0.0, value=0.0, step=0.1)
area_mean = st.sidebar.number_input("area_mean", min_value=0.0, value=0.0, step=0.1)
```

Figura 10. Diseño de los distintos botones del menú lateral correspondientes a las variables predictoras.

```
if st.button("Haz la predicción"):
    st.markdown(
        f""La probabilidad de que la muestra sea Benigna es:
        <br/>
        <font size="6"> {model.predict_proba([input_values])[0][0]*100:.2f}% </font>""",
        unsafe_allow_html=True,
    )
```

Figura 11. Diseño del botón de inicio de la predicción que al ser accionado devuelve la probabilidad expresada en porcentaje de que una muestra sea benigna.

## 4. Resultados

A continuación, se procede a mostrar y comentar los resultados obtenidos.

### 4.1 Análisis exploratorio de los datos

En primer lugar, pasamos a analizar el informe de la librería “pandas-profiling” (figura 12):

Dataset statistics		Variable types	
Number of variables	31	Categorical	1
Number of observations	569	Numeric	30
Missing cells	0		
Missing cells (%)	0.0%		
Duplicate rows	0		

Figura 12. Extracto del informe HTML obtenido con la función pandas-profiling.

El informe nos muestra que hay 569 instancias, 30 variables numéricas y 1 categórica y nos confirma que no existe presencia de valores perdidos ni filas duplicadas.

Como podemos observar en la figura 13, en la variable categórica “diagnosis” tenemos dos posibles valores (B y M) que hacen referencia a si el diagnóstico del análisis de la muestra es benigno o maligno.

diagnosis  
Categorical

Distinct	2	B	357
Distinct (%)	0.4%	M	212

Figura 13. Informe de la función pandas-profiling de la variable diagnosis. Se observan 357 instancias con diagnóstico benigno y 212 maligno.

Utilizando esta librería también se puede obtener un informe de cada variable en el que podemos encontrar información interesante como los valores máximo y mínimo, la media o los valores perdidos. En la figura 14 podemos observar un ejemplo de estos informes.

radius\_mean

Real number ( $\mathbb{R}$ )

Distinct	456	Minimum	6.981
Distinct (%)	80.1%	Maximum	28.11
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	14.127292	Memory size	8.9 KiB

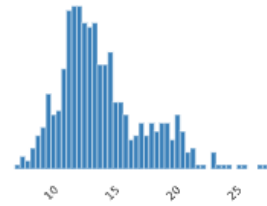


Figura 14. Extracto del informe realizado con la función pandas-profiling que muestra información sobre la variable “radius\_mean” (media del radio nuclear).

## 4.2 Creación, optimización, prueba y validación del modelo de predicción

La función GridSearchCV, va a buscar de forma exhaustiva entre los parámetros propuestos aquellas que producen un mejor resultado en la predicción optimizando el modelo (Adnan et al., 2022).

Como se observó en la figura 7 en el apartado de materiales y métodos, tras aplicar la función de entrenamiento con la validación cruzada, se puede observar que los parámetros óptimos para este modelo han sido:

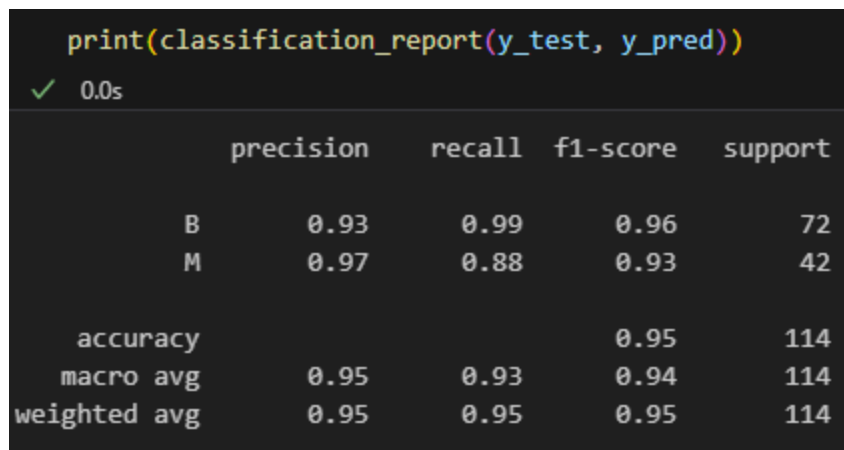
- En la función de activación: la función ReLu que es la más utilizada en redes convolucionales (Lecun et al., 2015).
- Alpha de 0,0001 que es el predeterminado por la librería.
- Un ratio de aprendizaje constante y gradiente estocástico propio de la librería.

En cuanto a la probabilidad de obtener un resultado maligno o benigno con los datos de prueba, como podemos observar en la figura 15, existe una probabilidad de aproximadamente el 70% de que la muestra sea benigna y un 30% de que sea maligna.

```
clf.classes_ 0.0s  clf.predict_proba(X_test[:1])
✓ 0.0s  ✓ 0.0s
array(['B', 'M'], dtype='<U1') array([[0.69277593, 0.30722407]])
```

Figura 15. Resultados posibles de la predicción (Benigno o Maligno) y probabilidad de obtener cada uno de ellos con los datos de prueba.

La calidad del modelo de clasificación se valora mediante un informe de clasificación que calcula una serie de parámetros de evaluación como la precisión, el recall y el f1-score (Figura 16). La precisión se podría definir como la proporción de predicciones correctas entre el total de predicciones, por otro lado, el recall sería la proporción de verdaderos positivos que el modelo ha predicho correctamente y el f1-score sería la media entre los dos parámetros anteriores, lo que facilita la evaluación del rendimiento combinándolos en un solo parámetro.



```
print(classification_report(y_test, y_pred))
```

✓ 0.0s

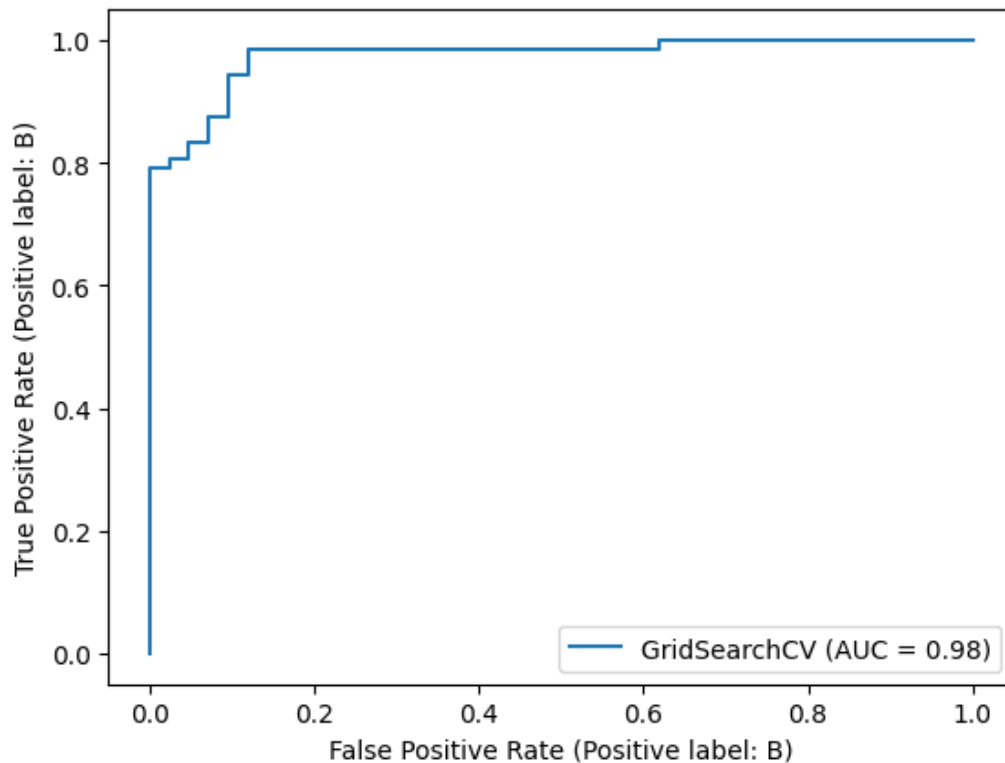
	precision	recall	f1-score	support
B	0.93	0.99	0.96	72
M	0.97	0.88	0.93	42
accuracy			0.95	114
macro avg	0.95	0.93	0.94	114
weighted avg	0.95	0.95	0.95	114

**Figura 16. Informe de calidad del modelo predictivo.**

Observando la figura 16, el modelo predictivo creado presenta un valor de f1-score alto de 0,96 y 0,93 para predecir casos benignos y malignos respectivamente. Estos valores indican que el modelo posee una precisión y exhaustividad (recall) alta a la hora de clasificar las muestras.

Finalmente, en la figura 17 podemos observar el gráfico correspondiente a la curva ROC. Si el modelo es capaz de realizar esta distinción con facilidad, el valor de AUC será cercano a 1, por otro lado, si el modelo no es capaz de discriminar entre pacientes con cáncer o sin cáncer, dicho valor será aproximadamente 0,5. Este método de validación se ha visto que es muy útil para medir el rendimiento de un modelo de predicción generado a partir de redes neuronales (Woods y Bowyer, 1997) y ha sido utilizado en otros estudios con redes neuronales como el de Lorente et al., (2013) y en estudios de cáncer de mama como el de Antropova et al., (2016) donde obtuvieron un área bajo la curva de 0,85 demostrando una buena calidad de predicción.

En este caso, como podemos observar en la figura 17, hemos obtenido un valor de área bajo la curva (AUC) muy cercano a la unidad (0,98) por lo que el modelo es capaz de discernir con precisión entre una muestra maligna y una benigna:



**Figura 17. Gráfico de la curva ROC-AUC. Se observa un valor de área bajo la curva de 0,98.**

### 4.3 Aplicación web

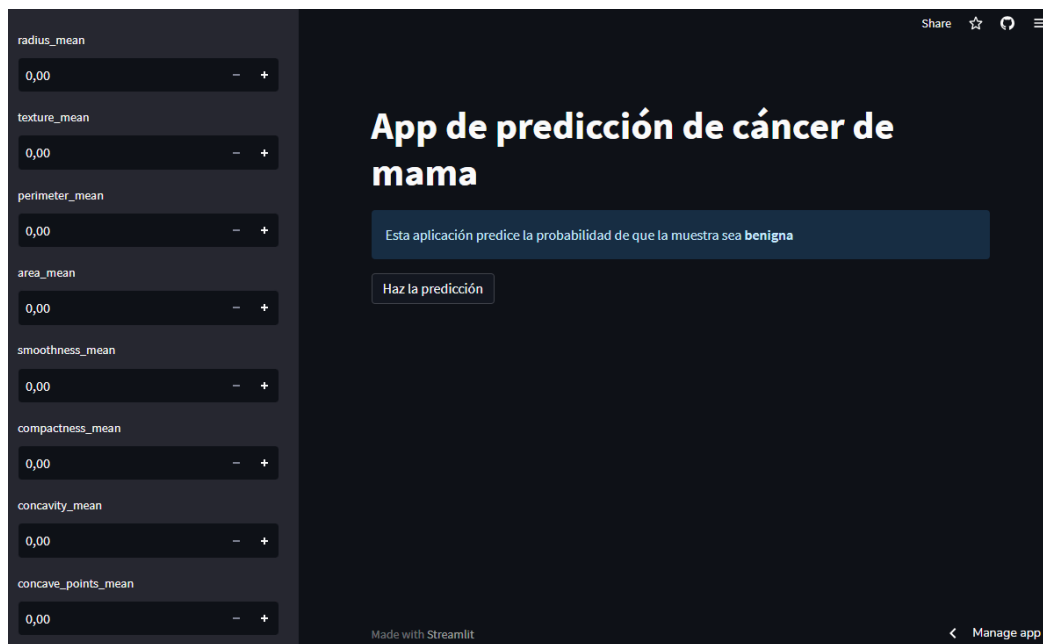
Mediante la librería Streamlit, se ha desarrollado una aplicación web sencilla que permite ingresar nuevas instancias de datos para obtener predicciones sobre la probabilidad de que un paciente padezca o no cáncer de mama.

Se ha elegido esta aplicación porque existen estudios como el de Gopiseti et al., en los que se ha utilizado esta herramienta para crear una aplicación para predecir enfermedades utilizando machine learning. Otro ejemplo de su uso es el estudio de Chalicham y Vineetha (2023), en el que se utilizó una aplicación desarrollada con Streamlit para reconocimiento facial utilizando métodos de aprendizaje automático.

Esta plataforma nos ofrece la posibilidad de crear y compartir aplicaciones web de manera sencilla lo que incrementa enormemente la accesibilidad ya que permite abrir la aplicación en teléfonos móviles y ordenadores de todo el

mundo (Moharekar, Pol, Ombase y Moharekar, 2022). De esta forma, es posible hacer público el uso del modelo de predicción de cáncer de mama que hemos desarrollado.

Para ello, el investigador debe introducir en el menú lateral de la aplicación (Figura 18), los diferentes parámetros nucleares obtenidos mediante procesamiento de imágenes.



**Figura 18.** Pantalla principal de la aplicación web. En la parte izquierda de la misma puede observarse el menú lateral con los diferentes botones para introducir los parámetros de la muestra. Por otro lado, en la parte central, se encuentra ubicado el botón para realizar la predicción.

Una vez introducidos los parámetros de la muestra, el usuario debe hacer click en “Haz la predicción” para que tenga lugar la predicción. Como podemos observar en la figura 19 el resultado es la probabilidad, expresada en porcentaje, de que esa muestra sea benigna.

# App de predicción de cáncer de mama

Esta aplicación predice la probabilidad de que la muestra sea **benigna**

Haz la predicción

La probabilidad de que la muestra sea Benigna es:

**68.33%**

Figura 19. Pantalla final correspondiente al resultado de la predicción, en ella se puede observar el resultado expresado en porcentaje con dos decimales.

Para probar la aplicación, se puede acceder a ella través del enlace ubicado en el anexo 1 de este trabajo.

## 5. Conclusiones y trabajos futuros

En este estudio, se ha llevado a cabo una revisión bibliográfica que proporciona una visión global de la situación actual en el ámbito de la investigación del cáncer de mama, con un enfoque particular en la utilidad de las redes neuronales en este contexto. Se ha explorado el origen y los diferentes tipos de redes neuronales y se ha puesto de manifiesto su utilidad en el estudio y la detección del cáncer de mama. Además, se ha diseñado un modelo predictivo basado en una red neuronal convolucional usando la función `MLPClassifier` de la librería `Scikit-learn`. Para hacerlo, se ha utilizado un conjunto de datos procedente de la base de datos Breast Cancer Wisconsin (Diagnostic) de UCI, que contiene 569 instancias y 30 variables relacionadas con la morfología nuclear de las células tumorales. Para valorar la calidad del modelo creado, se han aplicado técnicas de optimización, como `GridSearchCV`, y se ha evaluado el rendimiento con una curva ROC, logrando un AUC de 0,98, y un valor de F1-scores de 0,96 y 0,93 para resultados benignos y malignos, respectivamente.

Se ha desarrollado una aplicación web interactiva y fácil de usar en `Streamlit` que permite a los usuarios introducir nuevas instancias de datos de las 30 variables predictoras para realizar predicciones. Este desarrollo tiene el objetivo de mostrar que el modelo predictivo y la aplicación podrían servir como una herramienta de apoyo a investigadores y médicos a la hora de diagnosticar el cáncer de mama, aumentando la eficiencia y la precisión del proceso.

No obstante, es importante mencionar que existen ciertas limitaciones en las técnicas utilizadas. Por ejemplo, aunque las redes neuronales son altamente efectivas en la identificación de patrones y tendencias, su interpretación es compleja y la precisión del modelo depende en gran medida de la calidad y la cantidad de los datos disponibles para el entrenamiento. Cualquier sobreajuste o error en estos datos puede conducir a predicciones incorrectas o sesgadas.

En conclusión, este proyecto muestra la utilidad y el potencial de las redes neuronales y las tecnologías de machine learning en la investigación y el diagnóstico del cáncer de mama.



Para continuar con esta línea de trabajo, se podrían realizar estudios de correlación entre variables y técnicas de selección de variables para reducir el número de variables predictoras y ver realmente qué variables son las más importantes a la hora de predecir el diagnóstico de cáncer de mama. También, se podrían explorar otras técnicas de aprendizaje automático como las redes neuronales recurrentes (RNN) con el objetivo de intentar mejorar aún más la precisión y utilizar conjuntos de datos más grandes para valorar la robustez del modelo.

En cuanto a la aplicación web, en este trabajo se ha elaborado una interfaz sencilla, pero se podría mejorar incluyendo más funciones y haciendo posible que el usuario personalice su interfaz.

## 6. Glosario

**Red Neuronal Convolutacional (CNN):** Tipo de red neuronal artificial diseñada para procesar datos con una estructura de cuadrícula, como una imagen. Las CNN pueden identificar patrones con variación de escala y orientación y tienen una capacidad robusta para detectar características espaciales.

**Machine Learning:** Rama de la inteligencia artificial que se centra en el desarrollo y aplicación de algoritmos que permiten a las máquinas aprender de los datos y hacer predicciones o decisiones sin ser programadas explícitamente.

**Modelo Predictivo:** Modelo que se utiliza para predecir los valores futuros basándose en los datos históricos.

**FNA (Fine Needle Aspirates):** método para realizar biopsias en el que se utiliza una aguja fina para extraer tejidos o fluidos del paciente.

**Scikit-learn:** Biblioteca para aprendizaje automático en Python. Incluye algoritmos de clasificación, regresión, agrupación y reducción de dimensionalidad.

**GridSearchCV:** Método para realizar la búsqueda exhaustiva de hiperparámetros en un modelo de aprendizaje automático para encontrar la mejor combinación de parámetros que mejoren la precisión del modelo.

**Curva ROC (Receiver Operating Characteristic):** Gráfico que muestra la capacidad de diagnóstico de un sistema de clasificación binario a medida que varía su umbral de discriminación.

**F1-Score:** Medida de precisión para modelos de clasificación binaria. Es la media armónica de la precisión y la sensibilidad.

**Streamlit:** Framework de código abierto en Python para crear aplicaciones de machine learning y ciencia de datos.

**Procesamiento de Imágenes:** Conjunto de técnicas que se aplican a las imágenes para resolver problemas como la mejora de la calidad de la imagen, la extracción de información útil, entre otros.

## 7. Bibliografía

- Adnan M, Alarood AAS, Uddin MI, ur Rehman I. 2022. Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. PeerJ Computer Science 8:e803 <https://doi.org/10.7717/peerj-cs.803>).
- AECC (2021). Observatorio del Cáncer - AECC. <https://observatorio.aecc.es/>
- Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., y Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6), 954-961.
- Barocas, S., Hardt, M., y Narayanan, A. (2019). Fairness and machine learning. *fairmlbook.org*. <http://www.fairmlbook.org>.
- Chalicham, L.P., Vineetha, K. (2023). Streamlit Web Application for Finding Similar Face Using Deep Learning. In: Senjyu, T., So-In, C., Joshi, A. (eds) *Smart Trends in Computing and Communications. SmartCom 2023. Lecture Notes in Networks and Systems*, vol 650. Springer, Singapore. [https://doi.org/10.1007/978-981-99-0838-7\\_50](https://doi.org/10.1007/978-981-99-0838-7_50)
- Chaudhary, K., Poirion, O. B., Lu, L., y Garmire, L. X. (2018). Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clinical cancer research*, 24(6), 1248-1259.
- Doshi-Velez, F., y Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Drukteinis, J. S., Mooney, B. P., Flowers, C. I., y Gatenby, R. A. (2013). Beyond mammography: new frontiers in breast cancer screening. *American Journal of Medicine*, 126(6), 472-479.
- Dua, D., y Graff, C. (2019). UCI Machine Learning Repository. University of California, School of Information and Computer Science. <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>

- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... y Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine*, 25(1), 24-29.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., y Thrun, S. (2019). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.
- Goodfellow, I., Bengio, Y., y Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>.
- Google Cloud (2021). Environmental report: How we're building a carbon-free future. <https://cloud.google.com/sustainability>.
- Harbeck, N., y Gnant, M. (2017). Breast cancer. *Lancet*, 389(10074), 1134–1150.
- Hochreiter, S., y Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
- Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., y Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.
- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., y Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13, 8-17.
- Krizhevsky, A., Sutskever, I., y Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).
- Kusiak, A. (2018). Smart manufacturing. *International Journal of Production Research*, 56(1-2), 508-517
- K. Woods and K. W. Bowyer, "Generating ROC curves for artificial neural networks," in *IEEE Transactions on Medical Imaging*, vol. 16, no. 3, pp. 329-337, June 1997, doi: 10.1109/42.585767.
- LeCun, Y., Bengio, Y., y Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
- L. D. Gopiseti, S. K. L. Kummera, S. R. Pattamsetti, S. Kuna, N. Parsi and H. P. Kodali, "Multiple Disease Prediction System using Machine Learning and Streamlit," *2023 5th International Conference on Smart*

*Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2023, pp. 923-931, doi: 10.1109/ICSSIT55814.2023.10060903.

- Lorente, D., Aleixos, N., Gómez-Sanchis, J. et al. Selection of Optimal Wavelength Features for Decay Detection in Citrus Fruit Using the ROC Curve and Neural Networks. *Food Bioprocess Technol* 6, 530–541 (2013). <https://doi.org/10.1007/s11947-011-0737-x>
- Mangasarian, O. L., Street, W. N., y Wolberg, W. H. (1994). Breast cancer diagnosis and prognosis via linear programming. In *AAAI Spring Symposium on Artificial Intelligence in Medicine*, Stanford, CA.
- McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... y Etemadi, M. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89-94.
- Moharekar, T. T., Pol, U. R., Ombase, R., & Moharekar, T. J. (2022). Detection and classification of plant leaf diseases using convolution neural networks and Streamlit. *International Research Journal of Modernization in Engineering, Technology and Science*, 4(7), 4305. <https://www.irjmets.com>
- OMS (2021). Breast Cancer: Prevention and control. <https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/>
- Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., ... y Borresen-Dale, A. L. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797), 747-752.
- Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., y Yang, G. Z. (2017). Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1), 4-21.
- Reglamento General de Protección de Datos (2016). <https://www.boe.es/doue/2016/119/L00001-00088.pdf>.
- Samuel, A. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), 210-229.
- Scikit-learn developers. (2023). MLPClassifier. Scikit-learn: Machine Learning in Python. <https://scikit->

learn.org/stable/modules/generated/sklearn.neural\_network.MLPClassifier.html.

- Spanhol, F. A., Oliveira, L. S., Petitjean, C., y Heutte, L. (2016). Breast cancer histopathological image classification using Convolutional Neural Networks. 2016 International Joint Conference on Neural Networks (IJCNN), 2560-2567.
- Stacy L. Özesmi, Can O. Tan, Uygur Özesmi. Methodological issues in building, training, and testing artificial neural networks in ecological applications, Ecological Modelling, Volume 195, Issues 1–2, 2006, Pages 83-93, ISSN 0304-3800, <https://doi.org/10.1016/j.ecolmodel.2005.11.012>. (<https://www.sciencedirect.com/science/article/pii/S0304380005005806>)
- Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., y 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.
- Sutton, R. S., y Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.
- Turing, A. (1950). Computing machinery and intelligence. Mind, 59(236), 433-460.
- Web Content Accessibility Guidelines (WCAG) (2021). <https://www.w3.org/WAI/standards-guidelines/wcag/>.
- Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... y Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. Nature Medicine, 25(9), 1337-1340.
- Yuan, Y., Van Allen, E. M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., ... y Kim, P. (2020). Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nature biotechnology, 32(7), 644.

## 8. Anexo 1

- Enlace de acceso a la aplicación: <https://bert-tfm.streamlit.app/>
- Enlace al código fuente del modelo predictivo y de la aplicación: <https://github.com/bert25/tfm-app>